Review

Diagnostic Performance of Imaging-Based AI vs Radiologists for Preoperative Lymph Node Metastasis in Esophageal Cancer: Systematic Review and Meta-Analysis

Cong Tan[1], MMed; Lingfeng Xia[1], MMed; Xiongying Gao[1], BMed; Zhilin Zhong[1], BMed; Shuang Zhang[1], BMed; Wenjun Deng[1], BMed; Shen Gui[2], MD; Zhaojun Shang[1], BMed

[1]Department of Radiology, Wuhan Asia General Hospital, Wuhan, Hubei, China
[2]Clinical and Technical Support, Philips Healthcare, Wuhan, China

Corresponding Author:

Zhaojun Shang, BMed
Department of Radiology
Wuhan Asia General Hospital
300 Taizihu North Road, Wuhan Economic and Technological Development Zone
Wuhan, Hubei 430056
China
Phone: +86 15527309119
Email: 15527309119@163.com

Abstract

Background: Imaging-based AI technologies, particularly deep learning and radiomics, offer a novel approach to improving the accuracy of lymph node metastasis (LNM) diagnosis.

Objective: This study aimed to evaluate the diagnostic performance of imaging-based AI models compared with radiologists in detecting preoperative LNM in esophageal cancer (EC).

Methods: A systematic literature search was conducted across 5 databases (PubMed, Embase, Web of Science, Cochrane Library, and IEEE Xplore) up to August 19, 2026. Eligible studies assessed imaging-based AI for preoperative LNM detection in EC (pathology reference standard). The protocol was prospectively registered in the International Prospective Register of Systematic Reviews (PROSPERO CRD420251169109). The PROBAST+AI (Prediction Model Risk of Bias Assessment Tool with AI) tool was used for quality assessment. Bivariate random-effects models were used to calculate pooled sensitivity, specificity, diagnostic odds ratio (DOR), and area under the curve (AUC); between-study heterogeneity was described using τ^2 and I^2 (interpreted descriptively rather than categorically); homogeneity of study effects was tested with the Q statistic, and 95% prediction intervals (PIs) were derived for all pooled estimates. Head-to-head comparisons used a random-effects meta-analysis of within-study paired differences (restricted maximum likelihood with Hartung-Knapp-Sidik-Jonkman adjustment).

Results: A total of 18 studies were included. Across the 17 internal validation sets (2485 participants), AI models demonstrated a pooled sensitivity of 0.77 (95% CI 0.73-0.81; 95% PI 0.62-0.87), specificity of 0.79 (95% CI 0.73-0.84; 95% PI 0.48-0.94), a DOR of 12.88 (95% CI 9.57-17.33; 95% PI 4.5-37.0), and an AUC of 0.83 (95% CI 0.81-0.86; 95% PI 0.80-0.86). In the 10 head-to-head studies, AI showed higher sensitivity than radiologists (a paired difference of +0.27, 95% CI 0.11-0.37; 95% PI –0.26 to 0.45; P=.006) and higher discrimination (DOR ratio 4.65, 95% CI 2.53-8.56; 95% PI 0.97-22.42; P<.001), whereas specificity did not differ (+0.06, 95% CI –0.08 to 0.16; 95% PI –0.39 to 0.26; P=.30). Pooled arm-specific AUCs were 0.86 (95% CI 0.82-0.89; 95% PI 0.79-0.91) for AI and 0.67 (95% CI 0.59-0.75; 95% PI 0.44-0.84) for radiologists. In 4 external validation studies, pooled sensitivity was 0.66 (95% CI 0.46-0.81; 95% PI 0.31-0.90), specificity was 0.85 (95% CI 0.75-0.91; 95% PI 0.63-0.94), and the DOR was 10.42 (95% CI 3.50-31.01; 95% PI 0.6-142.0). These estimates remain subject to substantial heterogeneity and design-related bias.

Conclusions: Imaging-based AI models demonstrated promising pooled diagnostic performance for preoperative LNM detection in EC. By pairing AI with radiologists' performance in the same cohorts, reporting PIs alongside CIs, and separating internal from external validation, this review provides a more realistic assessment than previous average-effect estimates. In predominantly retrospective and heterogeneous cohorts, AI had higher pooled sensitivity and discrimination than radiologists, whereas specificity did not differ; the selective extraction of the model with the highest AUC and very low certainty of

evidence precludes conclusions about clinical superiority. These findings support AI as a decision-support tool, warranting prospective, multicenter external validation and prespecified human-AI comparisons rather than replacing radiologists.

Trial Registration: PROSPERO CRD420251169109; https://tinyurl.com/8efd2cay

J Med Internet Res 2026;28:e95893; doi: 10.2196/95893

Keywords: AI; esophageal cancer; lymph node metastasis; imaging; meta-analysis

Introduction

Esophageal cancer (EC) remains a leading cause of cancer morbidity and mortality worldwide, with an estimated 511,000 new cases and 445,000 deaths in 2022, and esophageal squamous cell carcinoma (ESCC) accounts for the majority of the global burden, which remains concentrated in East Asia [1,2]. The status of the regional lymph nodes is a pivotal determinant of prognosis and treatment strategy: patients with clinically node-positive disease are generally directed toward neoadjuvant chemoradiotherapy or perioperative systemic therapy before resection, whereas carefully selected patients with node-negative disease may proceed directly to endoscopic or surgical treatment [3,4]. Because treatment intensity is decided on the preoperative nodal stage, the accuracy of that stage carries direct consequences for patient management.

Accurate preoperative assessment of lymph node metastasis (LNM) in EC is crucial for treatment decision-making and prognosis [5]. The incidence of LNM increases significantly with higher tumor stages and serves as an independent adverse prognostic factor for survival [6]. Current primary assessment methods include imaging techniques such as contrast-enhanced computed tomography (CECT), contrast-enhanced magnetic resonance imaging (CEMRI), and [^{18}F] fluorodeoxyglucose positron emission tomography/computed tomography ([^{18}F] FDG PET/CT), as well as pathologic biopsy, all of which have limitations [7]. CECT and CEMRI rely on morphological criteria, exhibiting insufficient sensitivity for detecting micrometastases [7]. PET/CT is influenced by metabolic activity, limiting its ability to detect lesions with low metabolic uptake [8]. Although pathological biopsy is considered the "gold standard," its invasive nature and sampling bias hinder comprehensive preoperative evaluation [9]. In clinical practice, CT reports modest sensitivity, commonly in the range of 40% to 70%, for malignant lymph nodes because size and morphological criteria cannot reliably distinguish micrometastases from reactive hyperplasia [10-12]. Similarly, the value of PET/CT is constrained by partial-volume effects and the low metabolic uptake of small deposits.

AI has moved from methodological promise to active clinical investigation across radiology and oncology [13,14]. Deep learning (DL) architectures, in particular, convolutional neural networks, now perform detection, segmentation, and classification tasks in medical images at levels approaching those of expert readers [15]. In parallel, radiomics converts routine imaging (CT, MRI, and PET) into high-dimensional quantitative features that capture intratumoral and intranodal heterogeneity not visible to the naked eye, and radiomics signatures have been associated with staging, treatment response, and survival in EC and gastric cancer [16]. However, the translation of such models has been tempered by recurring methodological concerns, including small single-center samples, overfitting, inconsistent feature definitions, and limited reproducibility [17]. Most consequentially for diagnostic test accuracy, many models have never been evaluated beyond their development site; a systematic review of DL for radiologic diagnosis showed that external validation is frequently absent and, when performed, usually yields lower performance than internal testing [18].

Imaging-based AI technologies, particularly DL and radiomics, offer a novel approach to improving the accuracy of LNM diagnosis. AI can automatically extract high-throughput features from medical images, and multicenter studies have demonstrated its significant potential. AI-assisted computer-aided diagnostic systems achieved an accuracy of 0.744 in predicting LNM in ESCC during external validation, enhancing radiologists' diagnostic accuracy from 0.712 to 0.833 and specificity from 0.697 to 0.891 [19]. Some integrated models have reported an area under the curve (AUC) as high as 0.955 in internal validation [20]. However, key controversies remain: the diagnostic performance of different AI models shows considerable heterogeneity, their generalizability to external datasets is not yet well-established, and direct comparison of efficacy between AI and radiologists, particularly those with varying levels of expertise, remains inconclusive.

Because research in this area has expanded rapidly, several evidence syntheses have already appeared. Systematic reviews and meta-analyses have pooled the diagnostic accuracy of CT-based radiomics, DL, or machine learning (ML) models for predicting LNM in EC, with summary AUCs of approximately 0.74 to 0.85 [21-24], and a parallel synthesis exists for AI-based diagnosis of nodal metastases in abdominopelvic malignancies [25]. These works confirm that quantitative synthesis is feasible, but they leave unresolved the questions that matter most for clinical adoption. None benchmarked AI against radiologists reading the same images, although head-to-head comparisons in other imaging domains indicate that AI performance is strongly comparator- and setting-dependent [26]; none separated internal from external validation, so the generalizability of pooled estimates could not be judged; ESCC and esophageal adenocarcinoma (EAC), which differ markedly in epidemiology and biology [2], were rarely examined separately; and quality appraisal relied on tools that predate the AI-specific PROBAST+AI guidance, while heterogeneity was typically interpreted through categorized I^2 thresholds rather than prediction intervals (PIs).

The present review was therefore designed to address these gaps. Unlike those two 2024 meta-analyses, which neither compared AI with radiologists nor stratified pooled estimates by validation type nor separated pathological subtypes [21,22], our review extends the evidence by incorporating multiple modalities, directly comparing AI with radiologists, assessing generalizability via validation stratification, and evaluating performance across ESCC and EAC.

In this context, the aim of the present systematic review and diagnostic test accuracy meta-analysis was to synthesize the evidence on imaging-based AI models for preoperative detection of LNM in EC and to contrast AI performance with that of radiologists. The specific objectives were (1) to estimate the pooled sensitivity, specificity, diagnostic odds ratio (DOR), and area under the receiver operating characteristic (ROC) curve of AI models, stratified by internal vs external validation; (2) to quantify AI vs radiologist differences in studies providing head-to-head comparisons within the same cohorts; (3) to explore performance across imaging modalities and between ESCC and EAC; and (4) to evaluate risk of bias with PROBAST+AI (Prediction Model Risk of Bias Assessment Tool with AI) and the certainty of evidence with GRADE (Grading of Recommendations Assessment, Development, and Evaluation). We hypothesized that AI models would achieve higher sensitivity than radiologists but that their overall diagnostic performance would be attenuated in externally validated cohorts.

Methods

Study Design

This systematic review and meta-analysis adhered strictly to the guidelines set forth by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA; Checklist 1) [27]. The study protocol was prospectively registered with the PROSPERO (International Prospective Register of Systematic Reviews) database under the registration number CRD420251169109.

Information Sources and Search Strategy

We performed a comprehensive literature search across 5 major electronic databases (PubMed, Embase, Web of Science, Cochrane Library, and IEEE Xplore) covering the period from their inception through August 19, 2026. The search logic was constructed to encompass 3 primary domains: (1) AI technologies (including terms such as “ML” and “DL”), (2) LNM, and (3) EC. To maximize sensitivity, we used a combination of MeSH and free-text keywords appropriate for each database. We did not apply any filters regarding publication language or date. The complete search syntax used for each database is detailed in Table S1 in Multimedia Appendix 1. The search process is reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search; Checklists 2 and 3) extension for search reporting. Two investigators (CT and LX) independently evaluated the retrieved records in a 2-stage process: first by screening titles and abstracts, and subsequently by reviewing the full texts of potentially eligible papers. Additionally, we manually examined the bibliographies of all included studies to identify any further relevant research.

Eligibility Criteria

Eligibility criteria were organized using a purpose-adapted PITROS framework, which specifies participants, index test, target condition, reference standard, outcomes, and setting. Participants were patients with confirmed EC undergoing preoperative staging. The index test was imaging-based AI, including ML and DL models that used CECT, CEMRI, or [^{18}F] FDG PET/CT. The target condition was pathology-confirmed LNM. The reference standard was pathological analysis from surgical specimens or biopsies. Outcome measures included sensitivity, specificity, AUC, and DOR. Retrospective and prospective studies conducted using public databases or in hospital settings were eligible.

Studies were excluded for the following reasons: clearly irrelevant titles or abstracts; ineligible publication types such as reviews, case reports, conference abstracts, meta-analyses, and letters; non-English publications due to lack of accessibility; overlapping patient cohorts; lack of focus on preoperative lymph node staging in EC; or insufficient data to construct a 2 × 2 contingency table. The screening was performed independently by 2 reviewers (ZS and XG), with discrepancies resolved through discussion or adjudication by a third reviewer (ZZ).

Risk of Bias Assessment

The methodological quality and risk of bias of the included diagnostic accuracy studies were critically appraised using the PROBAST+AI tool, an updated instrument that supersedes the PROBAST 2019 criteria [28]. This tool evaluates studies in 2 phases—model development and model evaluation—each encompassing 7 domains related to participants, predictors, outcome assessment, and analysis. Each domain is judged as having low (L), high (H), or unclear (U) risk of bias based on responses to specific signaling questions. The complete set of signaling questions and detailed assessments are provided in Tables S2 and S3 in Multimedia Appendix 1. Two reviewers (WD and SG) independently assessed study quality, resolving disagreements through discussion to reach consensus, thereby ensuring the reliability of the systematic review. PROBAST+AI was selected because it is tailored to the development and evaluation of AI prediction models. It provides a structured assessment of risk of bias and applicability but does not independently address all reader-interpretation, blinding, threshold, and comparative test-design issues relevant to radiologist comparisons.

Certainty of Evidence

A standardized data extraction form was used to collect relevant information from the included studies. The evidence certainty for pooled sensitivity, specificity, and DOR was then assessed using the GRADE framework [29]. This involved evaluating 5 domains (risk of bias, indirectness, inconsistency, imprecision, and publication bias), and downgrading the evidence where necessary; inconsistency was downgraded when substantial heterogeneity remained

unexplained, as the identification of potential sources through exploratory analyses alone was not considered a sufficient explanation. Complete assessment details are provided in Table S4 in Multimedia Appendix 1.

Data Items and Data Extraction

Data extraction was performed independently by 2 reviewers (CT and ZS) using a standardized form. Discrepancies were resolved by consultation with a third reviewer (ZZ). The extracted data included study and patient characteristics, technical and algorithmic details, and diagnostic performance metrics for AI models and radiologists in internal and external validation sets. Extraction was manual and independently cross-checked; automated systematic review or AI-assisted data extraction tools were not used.

When explicit diagnostic contingency tables were unavailable, values were derived using 2 methods. First, true positive (TP), false positive (FP), false negative (FN), and true negative (TN) counts were calculated when sensitivity, specificity, and the number of positive cases were reported. Second, for 2 studies reporting ROC curves without sufficient tabulated metrics, coordinates were extracted using GetData Graph Digitizer, and the reported optimal sensitivity and specificity based on the Youden index were used to reconstruct contingency tables. These reconstructions are approximate, may not reflect a prespecified clinical operating threshold, and contributed to downgrading the certainty of the evidence.

Outcomes

Sensitivity and specificity were the core diagnostic accuracy outcomes for internal validation sets, external validation sets, and radiologists. DOR and AUC were complementary summary measures and were not interpreted as direct clinical endpoints. For studies reporting multiple nonoverlapping datasets, corresponding contingency tables were extracted under the published assumption of independence. When multiple AI algorithms were reported for a study, only the single model with the highest reported AUC was extracted to avoid reuse of participants. This selection may favor optimistic estimates and does not establish superiority under matched external evaluation.

Synthesis Methods and Statistical Analysis

Given the anticipated inherent heterogeneity in true diagnostic accuracy across studies, a bivariate random-effects model (a bivariate generalized linear mixed model) was used [30,31]. The analysis was stratified into 3 distinct categories: internal validation sets, external validation sets, and radiologists' assessments. When a study reported radiologists' data from more than one reader group or cohort, a single dataset per study was selected using a prespecified hierarchical rule: datasets derived from the same cohort as the AI analysis were preferred, and, among those, the dataset from the most experienced readers was used (Table S5 in Multimedia Appendix 1). To visualize aggregate performance, including sensitivity, specificity, and DOR, we constructed forest plots and summary ROC curves with 95% CIs. For the head-to-head comparison, within-study paired differences in logit-transformed sensitivity and specificity and in log-transformed DORs were synthesized with random-effects models using restricted maximum likelihood with the Hartung-Knapp-Sidik-Jonkman adjustment. Per-arm variances were summed as a conservative variance estimate that does not model the within-study covariance. Pooled logit-scale differences were back-transformed to probability differences anchored at the mean radiologists' sensitivity and specificity, and log-DOR differences were exponentiated to the ratio scale.

PIs for the pooled estimates of sensitivity, specificity, and DOR were calculated for the internal and external validation sets, where the number of studies was sufficient (using bootstrap resampling when fewer than 10 studies contributed). PIs for the AUC were calculated using Monte Carlo simulation from the estimated between-study distribution. CIs describe the uncertainty of the average effect, whereas PIs describe the expected distribution of true effects across different populations and settings [32].

Our registered protocol initially planned to perform indirect comparisons by pooling AI and radiologists' estimates separately from all eligible studies. However, during revision, we recognized that indirect comparisons across different populations, modalities, and thresholds may introduce substantial confounding. Therefore, we deviated from the protocol by restricting the primary comparative analysis to studies that provided direct head-to-head comparisons between AI and radiologists within the same patient cohorts (10/18 studies). All indirect comparisons from the original protocol are now presented as secondary or exploratory analyses, whereas the paired head-to-head comparisons are reported as the primary comparative analysis.

Between-study heterogeneity was described using the between-study variance (τ^2) and the proportion of variability attributable to heterogeneity (I^2; reported descriptively without categorization); homogeneity of study-specific effects was assessed using the Q statistic [33]. The Q statistic is a test of homogeneity for which a significant result indicates the presence of between-study heterogeneity. It does not quantify the magnitude of heterogeneity, which was judged from τ^2, I^2, and PIs. Interpretation of heterogeneity focused on PIs rather than on categorized heterogeneity indices, consistent with current guidance [34]. Given the limited number of internal validation studies, exploratory univariable meta-regression was restricted to study-level factors that could be consistently categorized, including AI method, AI model type, data-splitting method, study design, and unit of analysis. Segmentation, feature extraction, and feature selection methods were not analyzed because of substantial between-study heterogeneity and sparse categories. Diagnostic thresholds were also excluded because they were model-specific and not directly comparable across studies.

For radiologists' data, bivariate boxplots were used. Prespecified subgroup analyses were performed to assess the impact of different imaging modalities (CECT, CEMRI,

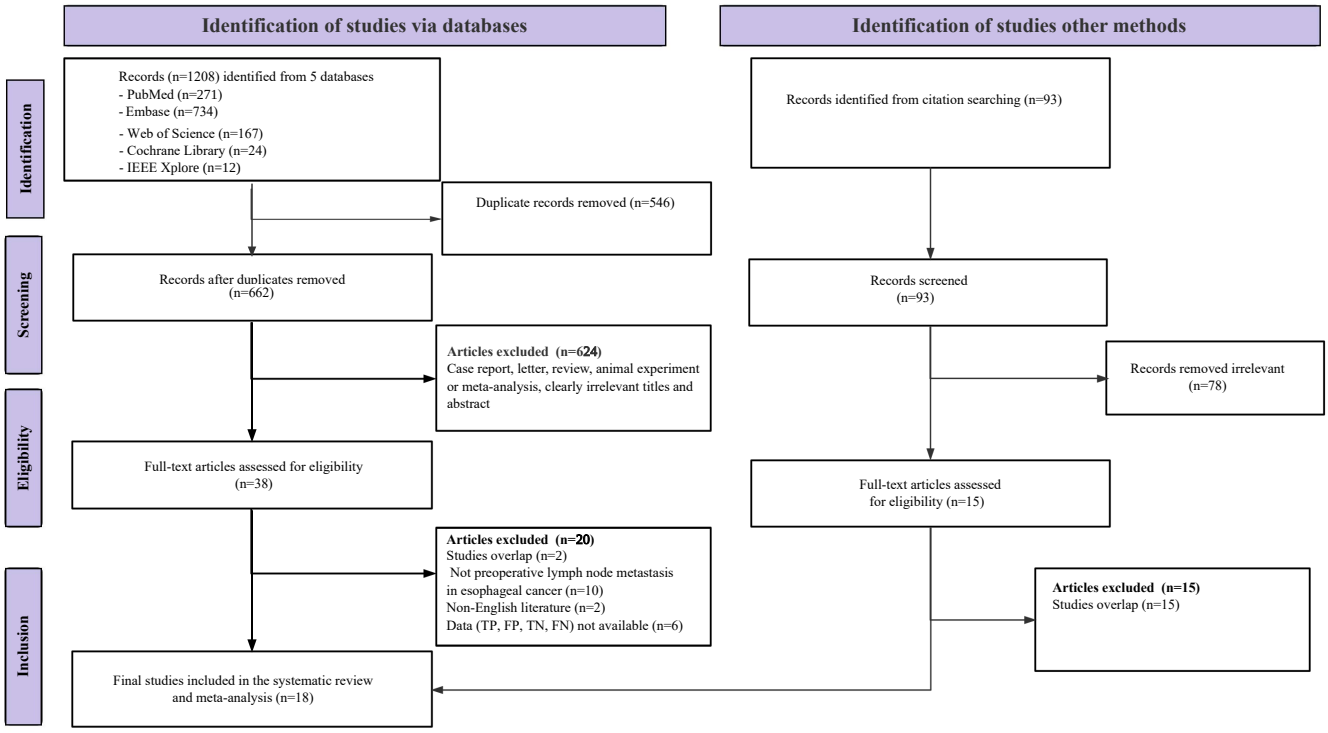
and [^{18}F] FDG PET/CT) and EC types (ESCC and EAC). A bubble plot explored the descriptive association between DOR and publication year and was not treated as a test of temporal improvement. A radar chart illustrated the distribution of ML and DL algorithms. A Fagan nomogram used an illustrative pretest probability, and Deeks' funnel plot assessed small-study effects rather than publication bias alone [35,36]. All statistical analyses were performed using R version 4.5.2 (R Foundation for Statistical Computing). Bivariate random-effects models were fitted using the lme4 package; random-effects models and PIs were calculated using metafor, and summary ROC and AUC computations were performed with mada. Figures were generated using ggplot2.

Results

Study Selection

The initial database search identified 1208 potentially relevant papers. After removing 546 duplicate records, 662 papers proceeded to the preliminary screening stage. During this initial screening, 624 publications were excluded based on clearly irrelevant titles/abstracts or ineligible publication types, resulting in 38 papers for full-text review. Following a detailed assessment, 20 studies were excluded for the following reasons: 2 studies had overlapping patient populations, 2 were non-English publications, 10 did not specifically address preoperative LNM in EC, and 6 were excluded due to insufficient or incomplete diagnostic data (TP, FP, FN, and TN). Consequently, 18 studies met the predetermined inclusion criteria and were included in the final meta-analysis [19,37-53]. The study selection process is detailed in Figure 1.

Figure 1. Flow diagram of the study selection process for the systematic review and meta-analysis of imaging-based AI and radiologist performance for preoperative lymph node metastasis in esophageal cancer. FN: false negative; FP: false positive; TN: true negative; TP: true positive.



Study Characteristics and Risk of Bias

A total of 18 eligible studies were included in the analysis. The internal validation sets, comprising 17 studies [37-53], involved a total sample size of 2485 patients. External validation was performed in 4 studies [19,49,52,53], encompassing a total of 514 patients. Among the included studies, AI models were based on CECT in 14 studies, CEMRI in 2 studies [44,50], and [^{18}F] FDG PET/CT in 2 studies [19,52]. Furthermore, diagnostic data from radiologists could be extracted from 10 studies. The publication years of the included studies spanned from 2017 to 2025. Two studies were prospective in design [44,50], while the remaining studies were retrospective. All 18 studies used pathology as the reference standard. Except for one study where patient data originated from the Netherlands and the United Kingdom [52], all other studies were based on patient data from China. The pathological type was not specified in 2 studies [41,44]. One study included patients with EAC [52], and the remaining studies focused on ESCC.

The unit of analysis was lesion-based in 2 studies [38,51] and patient-based in the others. Data for one study were sourced from a public database [52], whereas data for all other studies were collected from local hospitals.

Five studies used radiomics-only AI models, and the remaining studies used models combining radiomics with clinical features [19,42,47,48,51]. Regarding data partitioning methods, 4 studies used K-fold cross-validation [39,40,48,50], one study used a time-series split [45], and the remaining studies used a hold-out method. The AI methodology was based on DL in 4 studies [19,38,48,51] and on ML in the others. Across both internal and external validation sets, the most frequently used AI modeling algorithm was logistic regression (12/21, 57.1%; Figure 2). Study and patient characteristics, along with technical and imaging features, are summarized in Tables 1 and 2 and Tables S5 to S9 in Multimedia Appendix 1.

Figure 2. Radar chart of AI algorithm types included in the meta-analysis. DL: deep learning; DT: decision tree; FAGNN: frequency-adaptive graph neural network; LR: logistic regression; LS-SVM: least squares support vector machine; M2ANet: multiscale multiattention network; ML: machine learning; ResNet-18: residual network-18; RF: random forest; U-Net: U-shaped network; XGBoost: extreme gradient boosting.

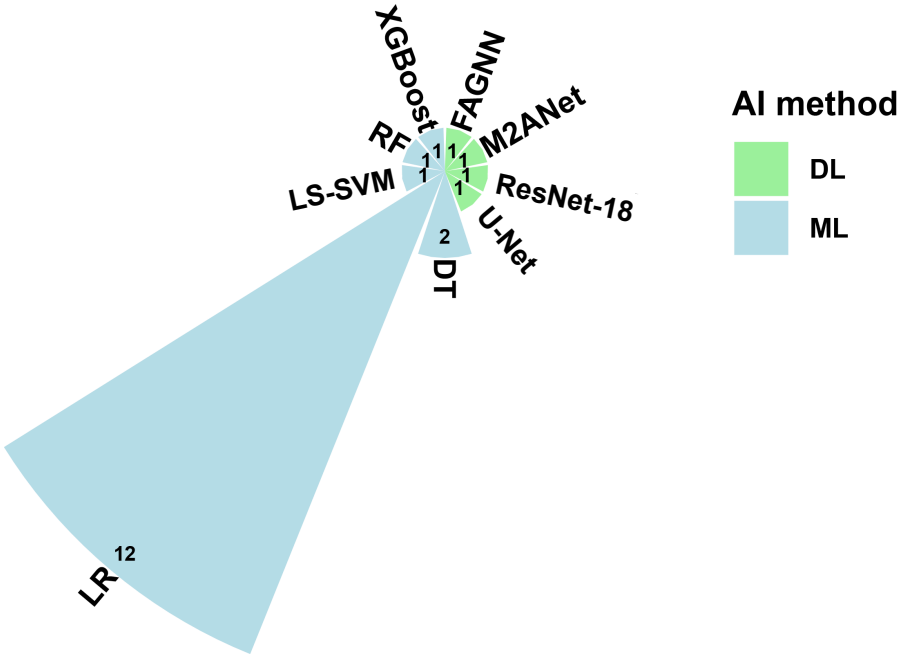


Table 1. Characteristics of the 18 included studies of imaging-based AI for preoperative lymph node metastasis in esophageal cancer (EC).

Author	Country	Study design	Type of histopathology	Type of imaging	Reference standard	Analysis unit	Number of total patients/lesions			Number of LNM[a]+patients/lesions
							Training	Internal validation	External validation	
Zhang et al [19], 2023	China	Retro	ESCC[b]	[^{18}F] FDG[c] PET/CT[d]	Pathology	PB[e]	532	—[f]	281	Training: not available, external validation: 116
Shen et al [45], 2018	China	Retro	ESCC	CECT[g]	Pathology	PB	140	57	—	Training: 42, internal validation: 19,
Li et al [41], 2021	China	Retro	EC	CECT	Pathology	PB	135	60	—	Training: 56, internal validation: 25
Ding et al [38], 2023	China	Retro	ESCC	CECT	Pathology	LB[h]	738	261	—	Training: 97, internal validation: 45
Wu H et al [48], 2025	China	Retro	ESCC	CECT	Pathology	PB	441	85	—	Training: 221, internal validation: 45
Huang et al [40], 2022	China	Retro	ESCC	CECT	Pathology	PB	350	350	—	Training: 55, internal validation: 55
Ou et al [42], 2021	China	Retro	ESCC	CECT	Pathology	PB	233	101	—	Training: 122, internal validation: 60
Zhang et al [53], 2025	China	Retro	ESCC	CECT	Pathology	PB	176	75	123	Training: 57, internal validation: 30, external validation: 42

Author	Country	Study design	Type of histopathology	Type of imaging	Reference standard	Analysis unit	Number of total patients/lesions			Number of LNM[a]+patients/lesions
							Training	Internal validation	External validation	
Xu et al [50], 2024	China	Pro	ESCC	CEMRI[i]	Pathology	PB	32	32	—	Training: 15, internal validation: 15
Peng et al [43], 2022	China	Retro	ESCC	CECT	Pathology	PB	193	81	—	Training: 99, internal validation: 41
Qu et al [44], 2019	China	Pro	EC	CEMRI	Pathology	PB	90	91	—	Training: 26, internal validation: 25
Chen et al [37], 2022	China	Retro	ESCC	CECT	Pathology	PB	216	92	—	Training: 108, internal validation: 46
Geng et al [39], 2024	China	Retro	ESCC	CECT	Pathology	PB	256	256	—	Training: 140, internal validation: 140
Tan et al [46], 2019	China	Retro	ESCC	CECT	Pathology	PB	154	76	—	Training: 81, internal validation: 37
Wang et al [47], 2017[j]	China	Retro	ESCC	CECT	Pathology	PB	—	131	—	Internal validation (combined cohort): 51
Zhang et al [52], 2021	Netherlands, United Kingdom	Retro	EAC[k]	[^{18}F] FDG PET/CT	Pathology	PB	—	130	60	Internal validation: 75, external validation: 35
Wu YP et al [49], 2025	China	Retro	ESCC	CECT	Pathology	PB	116	30	50	Training: 27, internal validation: 8, external validation: 14
Yi et al [51], 2025	China	Retro	ESCC	CECT	Pathology	LB	1327	577	—	Training: 297, internal validation: 172

[a]LNM: lymph node metastasis.
[b]ESCC: esophageal squamous cell carcinoma.
[c]FDG: fluorodeoxyglucose.
[d]PET/CT: positron emission tomography/computed tomography.
[e]PB: patient-based.
[f]Not available.
[g]CECT: contrast-enhanced computed tomography.
[h]LB: lesion-based.
[i]CEMRI: contrast-enhanced magnetic resonance imaging.
[j]For Wang et al [47], LNM counts were reported only for the combined cohort of 131 patients (66 training and 65 testing cases). This entry therefore does not represent an independent internal validation cohort.
[k]EAC: esophageal adenocarcinoma.

Table 2. Exploratory subgroup estimates of the diagnostic performance of imaging-based AI across internal and external validation contingency tables (21 tables: 17 internal and 4 external), stratified by imaging modality and histopathology[a].

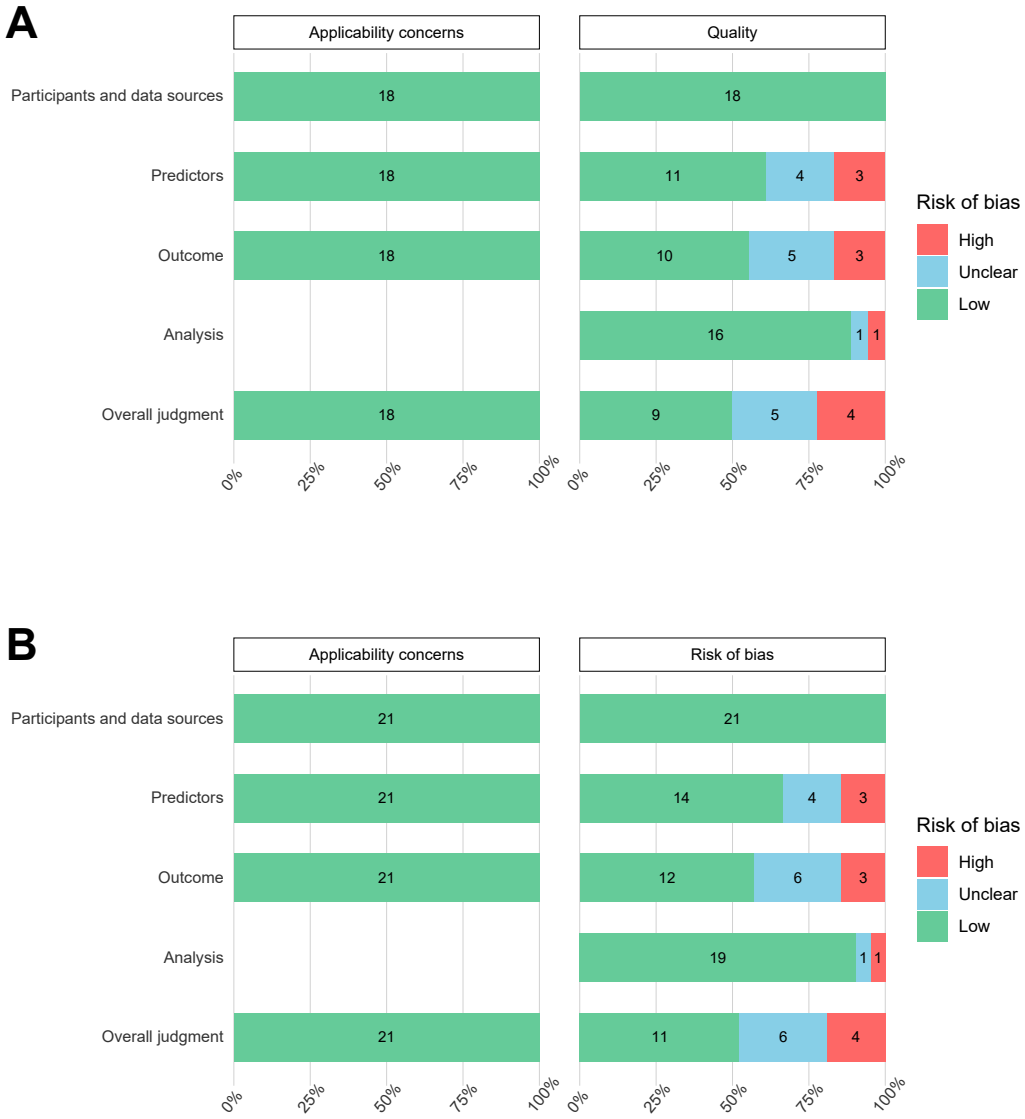
Subgroup	Number of contingency tables	Sensitivity (95% CI)	Specificity (95% CI)	AUC[b] (95% CI)	DOR[c] (95% CI)
Type of imaging					
CECT[d]	16	0.77 (0.70-0.82)	0.83 (0.76-0.88)	0.86 (0.83-0.89)	15.61 (10.81-22.55)
CEMRI[e]	2	0.80 (0.65-0.90)	0.70 (0.59-0.79)	0.74 (0.65-0.83)	9.28 (3.75-22.95)
[^{18}F] FDG[f] PET/CT[g]	3	0.70 (0.59-0.80)	0.75 (0.62-0.84)	0.78 (0.73-0.83)	6.97 (4.58-10.62)
Type of histopathology					
ESCC[h]	17	0.76 (0.70-0.82)	0.82 (0.76-0.87)	0.86 (0.83-0.89)	14.84 (10.34-21.29)
EAC[i]	2	0.75 (0.65-0.82)	0.65 (0.54-0.75)	0.85 (0.67-1.00)	5.45 (2.88-10.32)

[a]Subgroup estimates for CEMRI and EAC are each based on only 2 contingency tables and are strictly hypothesis-generating. The AUC CI for EAC was derived from a fixed-effect model and its upper limit was truncated at 1.00.
[b]AUC: area under the curve.
[c]DOR: diagnostic odds ratio.
[d]CECT: contrast-enhanced computed tomography.
[e]CEMRI: contrast-enhanced magnetic resonance imaging.
[f]FDG: fluorodeoxyglucose.
[g]PET/CT: positron emission tomography/computed tomography.
[h]ESCC: esophageal squamous cell carcinoma.
[i]EAC: esophageal adenocarcinoma.

Risk of Bias and GRADE Assessment

The risk of bias, assessed using the PROBAST+AI tool, is detailed in Figure 3 and Tables S2 and S3 in Multimedia Appendix 1. During the model development phase, 22.22% (4/18) of items in the quality domain were rated as high risk; all cohorts were judged as low risk for applicability concerns. In the model validation phase, 22.22% (4/18) of items in the risk of bias domain were rated as high risk. Similarly, all validation cohorts were assessed as low risk for applicability concerns. Among the 4 studies rated as high risk, high-risk judgments were most frequently attributable to the predictors and outcome domains (3 studies each), followed by the analysis domain (1 study). Although low-risk judgments predominated, retrospective design, selective model choice, heterogeneity, and reconstructed diagnostic tables require cautious interpretation of study quality. Based on the GRADE framework, the certainty of evidence was very low for all pooled analyses (Table S4 in Multimedia Appendix 1). The evidence was downgraded for risk of bias, inconsistency, indirectness, and imprecision; in particular, sensitivity and specificity values in 2 studies [46,47] were estimated from replotted ROC curves.

Figure 3. Quality assessment of the included studies using the PROBAST+AI tool. (A) PROBAST+AI model development. (B) PROBAST+AI model evaluation. PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for AI. Numbers in bars show the counts of assessments; zero counts are not shown. Panel A summarizes 18 model-development assessments. Panel B summarizes 21 validation datasets from 18 studies; study ratings are repeated for studies contributing both internal and external validation datasets.



Algorithm-Specific Estimates and Exploratory Temporal Pattern

Across the 18 included studies, the bubble plot (Figure 4) showed an exploratory association between DOR and publication year. Because this descriptive display did not adjust for cohort size, modality, validation design, or reporting practices, it was not interpreted as a tested temporal trend or evidence of algorithmic improvement. Figure 5 presents study-level performance estimates across algorithms and is not intended to rank their clinical performance.

Figure 4. Descriptive bubble chart of study-level diagnostic odds ratio (DOR) by publication year. Bubble size represents the DOR. This exploratory display does not test temporal improvement in algorithm performance. DT: decision tree; FAGNN: frequency-adaptive graph neural network; LR: logistic regression; LS-SVM: least squares support vector machine; M2ANet: multiscale multiattention network; ResNet-18: residual network-18; RF: random forest; U-Net: U-shaped network; XGBoost: extreme gradient boosting.

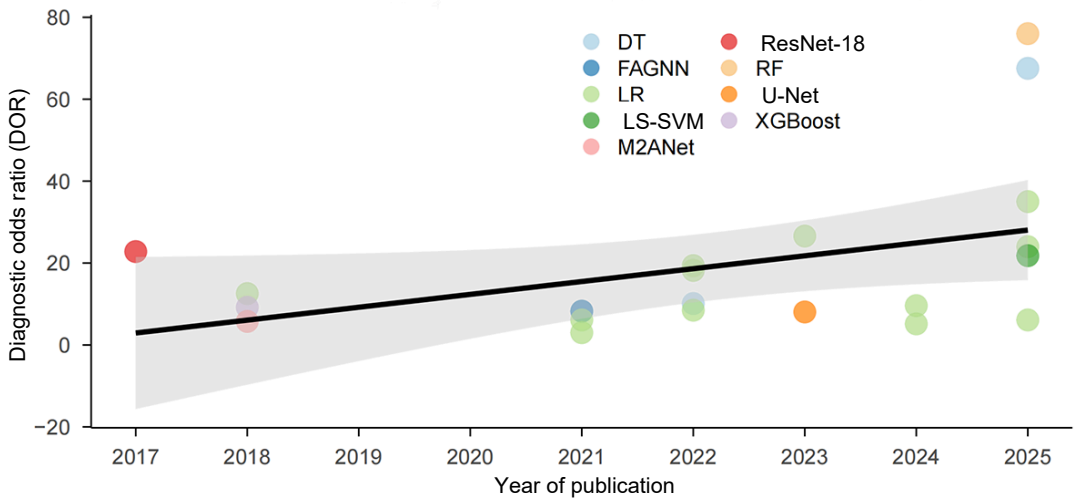
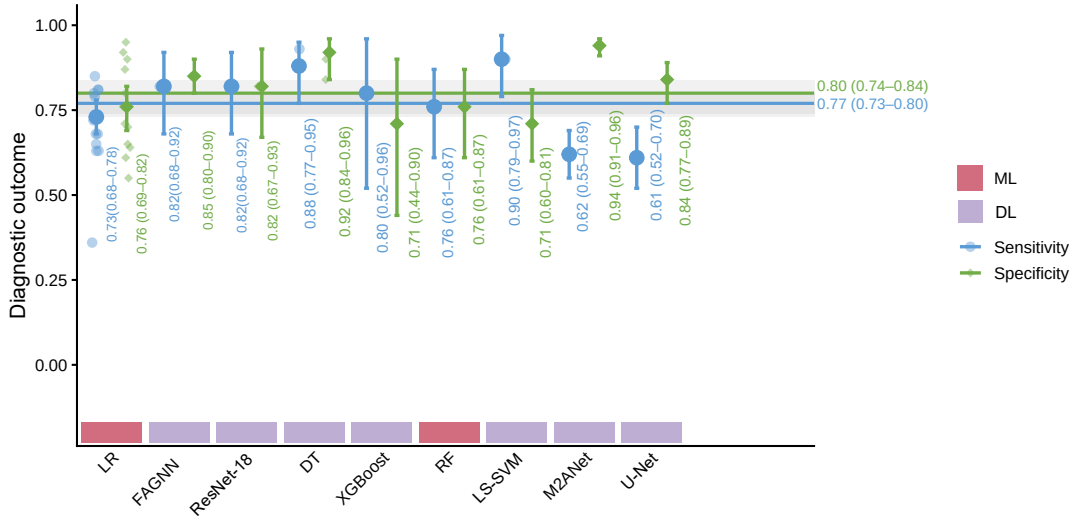


Figure 5. Study-level point-and-error-bar plot of sensitivity and specificity estimates, with 95% CIs, for AI algorithms in diagnosing preoperative lymph node metastasis in esophageal cancer. Points represent study-level estimates, and vertical error bars represent 95% CIs. Colors distinguish sensitivity and specificity, while the lower color strip distinguishes ML and DL algorithms. Horizontal lines denote pooled internal validation sensitivity and specificity. The display is descriptive and does not rank the clinical performance of individual algorithms. DL: deep learning; DT: decision tree; FAGNN: frequency-adaptive graph neural network; LR: logistic regression; LS-SVM: least squares support vector machine; M2ANet: multiscale multiattention network; ML: machine learning; ResNet-18: residual network-18; RF: random forest; U-Net: U-shaped network; XGBoost: extreme gradient boosting.

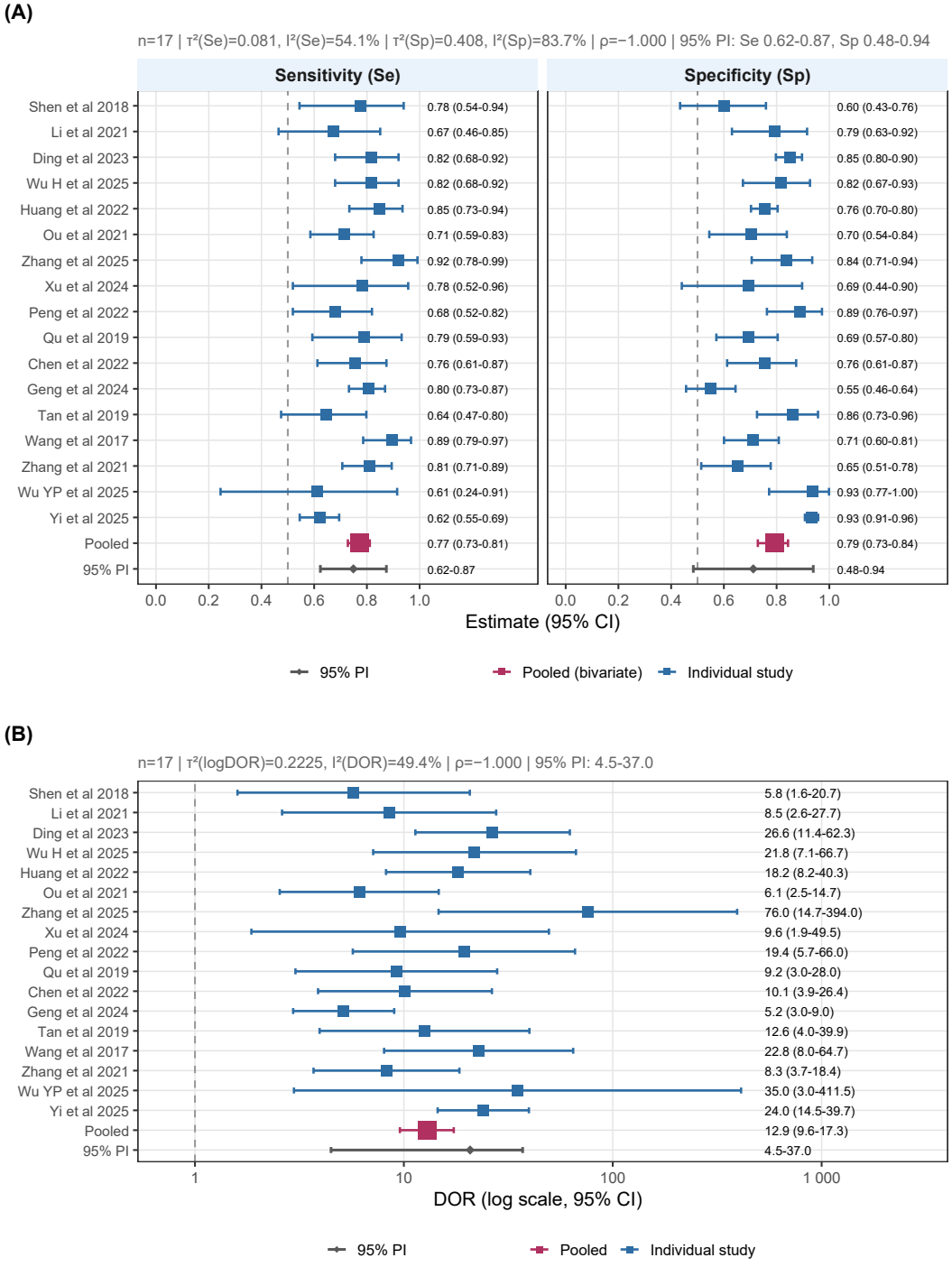


AI Estimates in Internal Validation Sets

In the internal validation set, AI models demonstrated a pooled sensitivity of 0.77 (95% CI 0.73-0.81; 95% PI 0.62-0.87; I^2=54.1%; very low certainty) and a pooled specificity of 0.79 (95% CI 0.73-0.84; 95% PI 0.48-0.94; I^2=83.7%; very low certainty). The pooled DOR was 12.88 (95% CI 9.57-17.33; 95% PI 4.5-37.0; I^2=49.4%; very low certainty), and the AUC was 0.83 (95% CI 0.81-0.86; 95% PI 0.80-0.86). τ^2 was 0.08 for sensitivity and 0.41 for specificity,

and the tests of homogeneity of study effects were significant for both outcomes (Q=39.5 and 115.1, respectively; df=16; both P<.001). These estimates are detailed in Figure 6 and Figures S1 to S3 in Multimedia Appendix 1.

Figure 6. Forest plots of pooled sensitivity and specificity (A) and DOR (B) for imaging-based AI models in internal validation sets (17 studies). Squares represent study-level estimates with 95% CIs; diamonds represent pooled bivariate random-effects estimates; shaded rows show pooled estimates and 95% PIs. The between-study τ^2 and I^2 values and PIs are shown in the subtitles [37-53]. DOR: diagnostic odds ratio; PI: prediction interval.



Head-to-Head Comparison Between AI Models and Radiologists

In the 10 head-to-head studies, paired analyses of within-study differences showed that AI models had a higher sensitivity than radiologists (pooled probability difference of +0.27; 95% CI 0.11-0.37; P=.006) and a higher overall diagnostic discrimination (DOR ratio 4.65; 95% CI 2.53-8.56; P<.001), whereas the specificity difference was not statistically significant (+0.06; 95% CI –0.08 to 0.16; P=.30; Figures 7 and 8). The 95% PIs for these paired differences were –0.26 to 0.45 (sensitivity), –0.39 to 0.26 (specificity), and 0.97 to 22.42 (DOR ratio), indicating that the

magnitude and, potentially, the direction of the sensitivity and DOR advantages may vary across settings. Pooled arm-specific estimates within these studies were AI sensitivity of 0.79 (95% CI 0.71-0.85), specificity of 0.80 (0.75-0.84), and AUC of 0.86 (0.82-0.89); radiologists' sensitivity of 0.53 (0.40-0.65), specificity of 0.73 (0.62-0.83), and AUC of 0.67 (0.59-0.75; Figure 7 and Figures S4-S6 in Multimedia Appendix 1). Between-study heterogeneity was substantial in the radiologist pool (τ^2=0.60 for sensitivity and 0.70 for specificity; Q=54.8 and 71.6, respectively; df=9; both P<.001). Eight of these studies contributed internal validation datasets, Zhang et al [19] contributed an external validation dataset, and Wang et al [47] contributed a combined training and testing cohort; the audited contingency tables and dataset selection notes are provided in Table S10 in Multimedia Appendix 1. Reader expertise and available clinical information were not standardized across studies (Table S11 in Multimedia Appendix 1).

Figure 7. Forest plots of pooled sensitivity and specificity (A) and DOR (B) for radiologists (10 studies). Squares represent study-level estimates with 95% CIs; diamonds represent pooled bivariate random-effects estimates; the shaded rows show pooled estimates and 95% PIs [19,37,38,41,45-48,50,53]. DOR: diagnostic odds ratio; PI: prediction interval; Se: sensitivity; Sp: specificity.

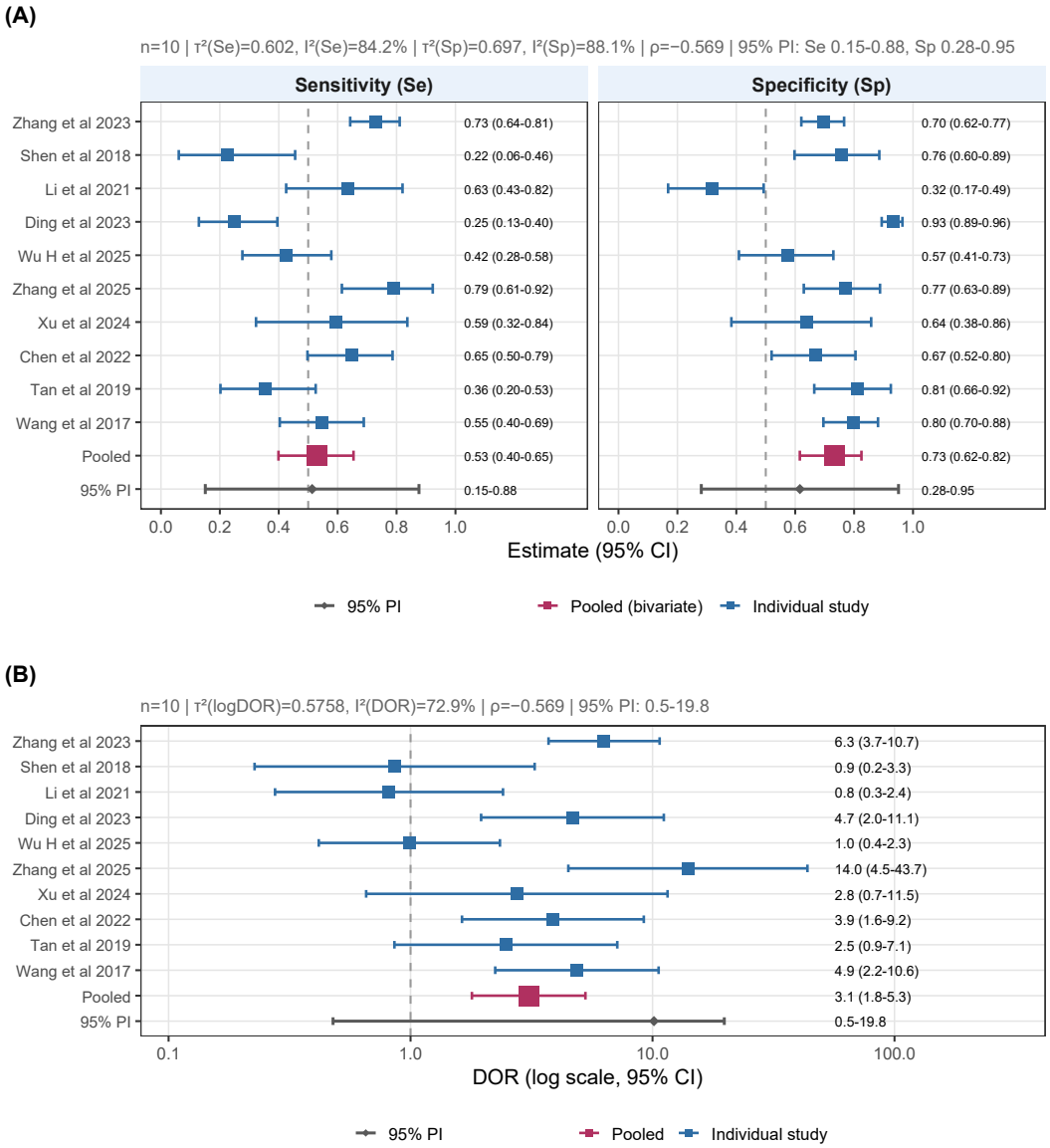
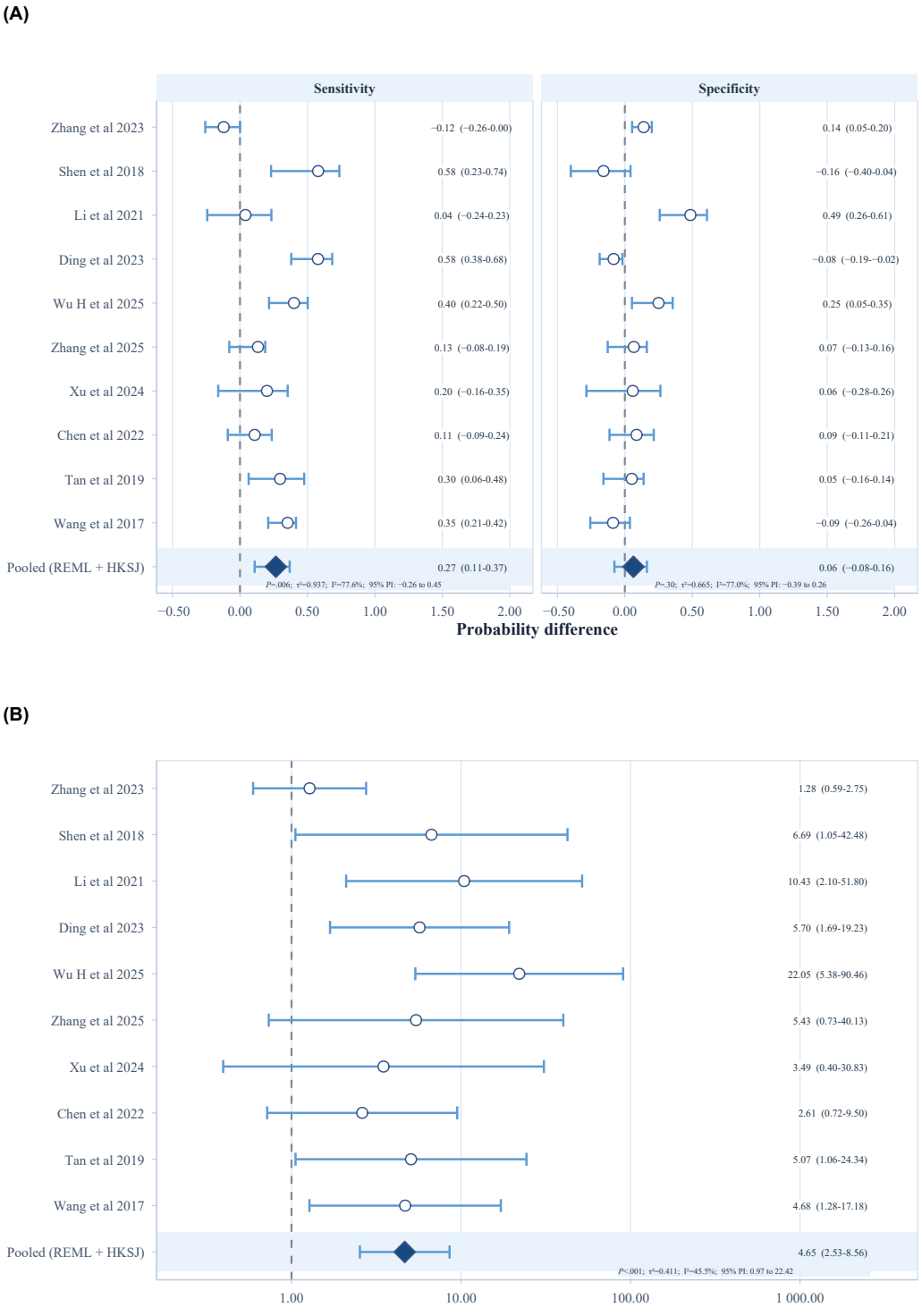


Figure 8. Paired comparison of imaging-based AI models vs radiologists in the 10 head-to-head studies. (A) Probability differences for sensitivity and specificity. Probability difference = AI − radiologist; positive values favor AI. Diamond = REML + HKSJ pooled estimate, back-transformed from the logit scale; dashed line indicates no difference. Values are shown as decimals to 2 places. Paired covariance is not included; pooled results are anchored at mean radiologist sensitivity/specificity. PIs describe the range expected in future similar settings; (B) DOR ratios (a logarithmic axis). Diamond = REML + HKSJ pooled DOR ratio; dashed line indicates no difference (ratio=1). PIs describe the range expected in future similar settings. Diamonds represent pooled estimates from random-effects models of within-study paired differences (restricted maximum likelihood with HKSJ adjustment); positive differences and ratios above 1 favor AI; pooled probability differences are anchored at the mean radiologist sensitivity/specificity and back-transformed from the logit scale. Per-study intervals use a conservative variance estimate that does not include the within-study covariance. 95% CIs and 95% PIs are shown [19,37,38,41,45-48,50,53]. Ratios above 1 favor AI in overall diagnostic discrimination. DOR: diagnostic odds ratio; HKSJ: Hartung-Knapp-Sidik-Jonkman; PI: prediction interval.



AI Estimates in External Validation Sets

In the external validation set, AI models demonstrated a pooled sensitivity of 0.66 (95% CI 0.46-0.81; 95% PI 0.31-0.90; I^2=85.9%; very low certainty) and a pooled specificity of 0.85 (95% CI 0.75-0.91; 95% PI 0.63-0.94; I^2=76.4%; very low certainty). The pooled DOR was 10.42 (95% CI 3.50-31.01; 95% PI 0.6-142; I^2=84.2%; very low certainty), and the summary AUC was 0.85 (95% CI 0.75-0.94; 95% PI 0.61-0.95). Because fewer than 10 studies contributed, PIs for the external validation pools were derived by bootstrap resampling. Between-study heterogeneity was substantial in these pools (τ^2=0.77 for sensitivity and 0.41 for specificity; Q=13.7 and 10.5, respectively; df=3; both P<.05). These estimates are detailed in Figures S7 to S9 in Multimedia Appendix 1.

Exploratory Subgroup Analyses of AI Estimates

Exploratory subgroup analyses are summarized in Table 2. Subgroup pooling combined internal and external validation contingency tables where available (21 tables in total) because the 4 external validation tables alone were insufficient for subgroup analysis, and all subgroup estimates should be read as exploratory across validation settings. CECT-based models had the highest summary estimates for specificity, AUC, and DOR, while CEMRI-based models had the highest pooled sensitivity. Estimates for CEMRI and EAC were each based on only 2 contingency tables and should therefore be interpreted as exploratory rather than as evidence of a preferred modality or histopathologic subgroup.

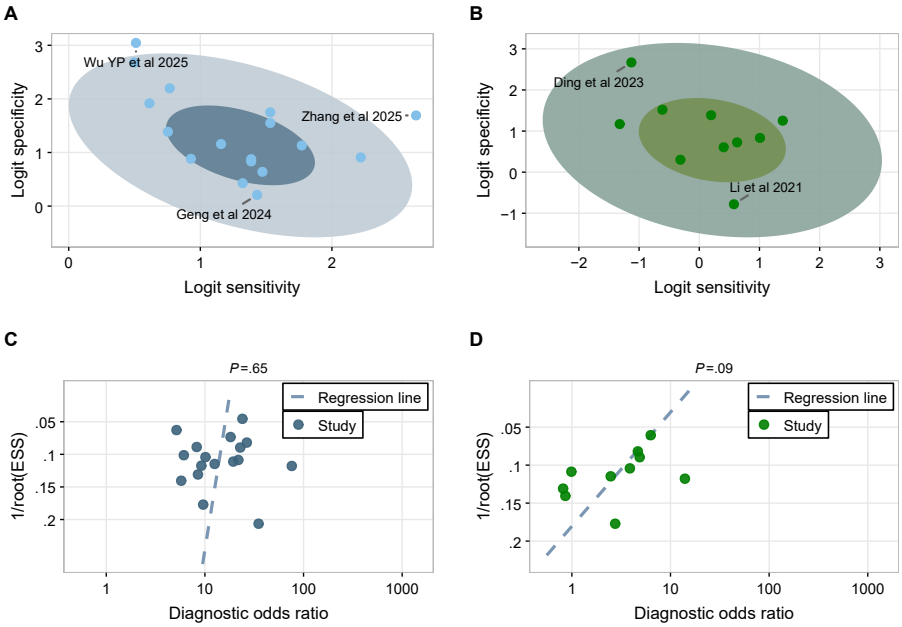
Exploratory Heterogeneity Analyses

To investigate potential sources of heterogeneity in the internal validation set, a univariable meta-regression was performed. Of the 4 prespecified study-level variables, only AI method (ML vs DL) and analysis unit (patient-based vs lesion-based analysis) were significantly associated with heterogeneity, and only for specificity (P=.04 and P=.003, respectively). AI model type and data-splitting method were not significantly associated with sensitivity or specificity. All subgroup estimates should be interpreted as exploratory correlates rather than established causal sources (Table 3). Furthermore, a bivariate boxplot analysis suggested that the studies by Geng et al [39], Zhang et al [53], and Wu YP et al [49] were potential outliers contributing to heterogeneity (Figure 9A). Sensitivity analysis after excluding these 3 outliers yielded the following pooled estimates: sensitivity of 0.77 (0.73-0.82), specificity of 0.80 (0.74-0.85), DOR of 13.33 (10.20-17.41), and AUC of 0.84 (0.82-0.86). For the analysis concerning radiologists, the bivariate boxplot analysis identified the studies by Li et al [41] and Ding et al [38] as potential sources of heterogeneity (Figure 9B). In additional sensitivity analyses excluding the 2 studies that required ROC-derived contingency tables, the internal validation estimates were similar (sensitivity 0.77 [0.72-0.81], specificity 0.79 [0.72-0.85], DOR 12.26 [9.11-16.51], and AUC 0.82 [0.80-0.85]); the corresponding radiologist estimates were sensitivity 0.55 (0.39-0.69), specificity 0.71 (0.56-0.82), DOR 2.93 (1.50-5.72), and AUC 0.67 (0.57-0.77). Similarly, excluding the 3 internal validation studies rated at high risk of bias yielded sensitivity 0.78 (0.72-0.82), specificity 0.81 (0.75-0.86), DOR 15.19 (11.61-19.88), and AUC 0.85 (0.83-0.87).

Table 3. Exploratory subgroup and univariable meta-regression analyses of AI performance in internal validation sets for preoperative lymph node metastasis in esophageal cancer.

Subgroup	Number of studies, n	Sensitivity (95% CI)	Meta-regression *P* value	Specificity (95% CI)	Meta-regression *P* value
AI method			.49		.003
Machine learning	14	0.78 (0.73-0.82)		0.76 (0.70-0.81)	
Deep learning	3	0.76 (0.62-0.86)		0.89 (0.81-0.93)	
AI model			.61		.39
Radiomic and clinical model	13	0.78 (0.74-0.82)		0.78 (0.71-0.84)	
Radiomic model	4	0.77 (0.66-0.85)		0.83 (0.70-0.91)	
Data-splitting method			.29		.14
Hold out	12	0.76 (0.70-0.81)		0.83 (0.76-0.88)	
K-fold cross-validation	4	0.83 (0.77-0.87)		0.71 (0.59-0.80)	
Study design			.70		.39
Retrospective	15	0.77 (0.72-0.81)		0.80 (0.73-0.86)	
Prospective	2	0.80 (0.65-0.90)		0.70 (0.59-0.79)	
Analysis unit			.18		.003
Patient-based analysis	15	0.78 (0.74-0.82)		0.76 (0.70-0.81)	
Lesion-based analysis	2	0.66 (0.59-0.72)		0.90 (0.87-0.92)	

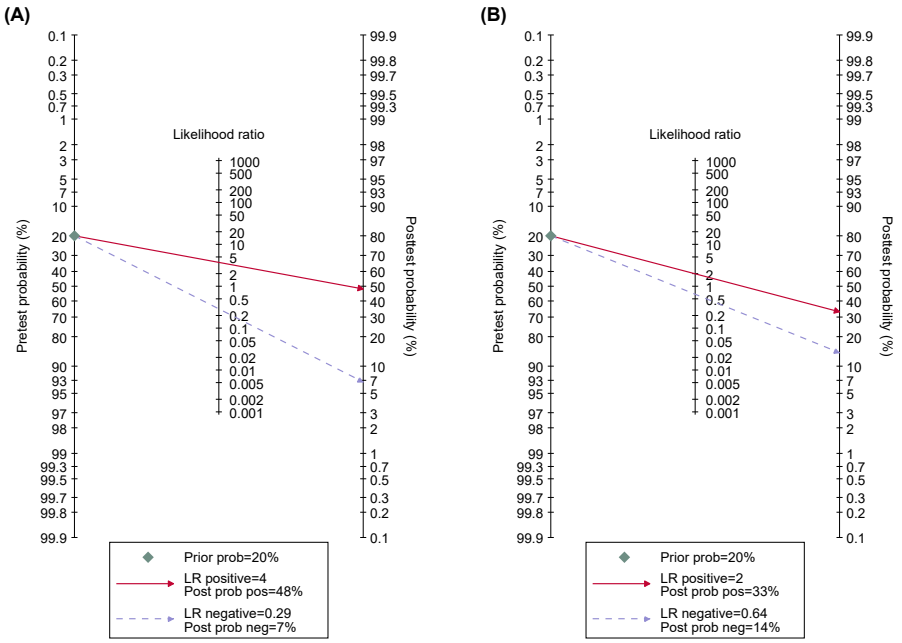
Figure 9. Bivariate boxplots and Deeks funnel plot asymmetry tests for internal validation sets and radiologist sets. (A) Bivariate boxplot for AI internal validation sets. (B) Bivariate boxplot for radiologist sets. (C) Deeks funnel plot asymmetry test for AI internal validation sets. (D) Deeks funnel plot asymmetry test for radiologist sets. The funnel plot analysis assesses small-study effects and does not by itself establish the presence or absence of such effects [38,39,41,49,53]. ESS: effective sample size.



Small-Study Effects and Illustrative Clinical Utility

Deeks funnel plot asymmetry test did not indicate significant small-study effects for the internal validation set (P=.65). For the radiologist analyses, the test was borderline (P=.09), which was considered in the GRADE assessment for publication bias (Figures 9C and 9D). These tests do not exclude small-study phenomena. The Fagan nomograms used an illustrative 20% pretest probability and yielded posttest probabilities of 42%, 48%, and 55% for AI, and 26%, 33%, and 42% for radiologists (Figures 10A and 10B and Figures S10-S13 in Multimedia Appendix 1). Because pretest LNM probability varies with clinical characteristics such as T stage, these examples should not be applied as patient-specific estimates.

Figure 10. Illustrative Fagan nomograms for AI and radiologists in diagnosing preoperative lymph node metastasis (LNM) in esophageal cancer. (A) AI (internal validation). (B) Radiologists. A 20% pretest probability is used for illustration and should not be treated as patient-specific because LNM prevalence varies with clinical context. LR: likelihood ratio; Prob: probability.



Discussion

Principal Findings

In paired analyses of the 10 head-to-head studies, AI models achieved higher pooled sensitivity and higher overall diagnostic discrimination (DOR) than radiologists, whereas specificity did not differ. The paired set comprised 8 internal validation datasets, one external validation dataset, and one combined training and testing cohort (Table S5 in Multimedia Appendix 1). These direct comparisons are vulnerable to selection, optimism, and design-related bias and do not establish intrinsic clinical superiority. AI may incorporate image features that are difficult to assess with unaided visual inspection [22]. Radiologist performance may vary with expertise, clinical information, and reading conditions, which were not standardized across the primary studies [21]. These potential advantages require evaluation in prospectively designed, matched human-AI comparisons before they can support clinical preference. In addition, only 4 studies reported external validation. Independent prospective multicenter studies with standardized acquisition, reference standards, and reader conditions are needed to assess clinical transportability [54].

Imaging modality may also affect AI performance. CECT-based AI demonstrated the highest specificity (0.83), AUC (0.86), and DOR (15.61), while CEMRI-based AI showed the highest sensitivity (0.80). This performance difference likely stems from the inherent physical and informational advantages of different imaging modalities. CECT provides high spatial resolution anatomical images with exceptional clarity, enabling AI models to more accurately exclude nonmetastatic lymph nodes based on their morphology, size, and enhancement characteristics. This significantly enhances diagnostic specificity, AUC, and DOR. In contrast, CEMRI, with its superior soft tissue contrast and multiparametric imaging capabilities (eg, diffusion-weighted imaging and T2-weighted imaging), is more sensitive to tissue edema and changes in cellular density [55]. This may allow AI models to detect subtle or early-stage metastatic lymph nodes earlier and more sensitively, thereby improving detection sensitivity [56]. Overall, the strength of CECT-based AI lies in accurate exclusion (high specificity), while that of CEMRI-based AI lies in sensitive detection (high sensitivity). This suggests that the two may hold complementary value in clinical practice [55]. Because the CEMRI and PET/CT subgroups contained few contingency tables, modality-specific estimates should be regarded as exploratory and require confirmation in prospective multicenter external validation studies.

Histopathologic subtype may also affect AI performance. In our analysis, AI models demonstrated higher overall diagnostic performance for patients with ESCC, achieving higher specificity, AUC, and DOR compared with patients with EAC. By contrast, the 2 EAC models showed numerically lower specificity, AUC, and DOR, with sensitivity estimates similar to those for ESCC. This performance discrepancy is likely attributable to the distinct biological behaviors and imaging manifestations of the 2 cancer subtypes. ESCC may follow a more predictable pattern of LNM, with imaging features of metastatic lymph nodes (such as morphology and enhancement patterns) appearing more typical and quantifiable on radiomic images. This allows AI models to more accurately differentiate nonmetastatic lymph nodes, leading to higher specificity, AUC, and DOR [55]. On the other hand, the biological characteristics of EAC may result in more variable or subtle imaging features of metastatic lymph nodes. This variability may lead models to prioritize detection over confidence, which may increase false positives and reduce specificity [6]. Only 2 studies focused specifically on EAC; therefore, subgroup differences should be regarded as exploratory and require confirmation in adequately powered, geographically diverse cohorts.

Compared with previous meta-analyses, we found 2 similar studies published in 2024 that explored AI-based detection of LNM in patients with EC. Ma et al [22] included 9 studies and reported a pooled AUC of 0.74 for radiomics models. In our study, the internal validation set achieved an AUC of 0.83. This difference may partly reflect the larger number of studies included in our analysis (18 vs 9). This broader evidence base may improve the stability of pooled estimates. Unlike their study, we additionally stratified performance by imaging modality, revealing that CECT-, CEMRI-, and PET/CT-based AI models had AUCs of 0.86, 0.74, and 0.78, respectively. This finding addresses the limitation of the study by Ma et al [22], which provided only a single pooled radiomics AUC, and highlights the clinical value of modality-specific analysis. Furthermore, while Ma et al [22] focused solely on radiomics, our study systematically compared the diagnostic performance of AI models with that of radiologists and investigated the impact of EC pathological subtypes (ESCC vs EAC) on model performance. This fills critical gaps in understanding the diagnostic value of AI relative to human experts and its performance across different EC subtypes.

Compared with the meta-analysis conducted by Jannatdoust et al [21] (which reported an AUC of 0.85 for CT-based AI models in predicting LNM in patients with EC), the CECT-based AI model in this study achieved an AUC of 0.86, indicating highly consistent results. However, our study introduces key innovations that advance the field: first, it expands the analytical scope from CT to include CEMRI- and PET/CT-based AI models, enabling a comprehensive comparison of AI performance across multiple imaging modalities; second, we distinguish between internal and external validation sets to evaluate the generalization capability of AI models—a critical dimension insufficiently addressed in previous meta-analyses. These additions may provide more comprehensive evidence and objective results for clinical practice.

Given the substantial heterogeneity among the included studies, we used a bivariate random-effects model to pool the results. This model summarizes average performance across heterogeneous studies and does not resolve differences between settings. We conducted exploratory univariable meta-regression and bivariate boxplot analyses. AI method

(ML vs DL), AI model types (radiomics and clinical models vs radiomics-only models), data-splitting method (hold-out vs K-fold cross-validation), and analysis unit (patient-based vs lesion-based) were examined as study-level correlates of heterogeneity; these associations do not establish causal sources [57]. In these analyses, only AI method and analysis unit reached statistical significance, and only for specificity: DL models showed higher pooled specificity than ML models (0.89 vs 0.76), whereas lesion-based analyses showed higher pooled specificity than patient-based analyses (0.90 vs 0.76), although both comparisons rested on only 2 to 3 studies and remain hypothesis-generating. AI model type and data-splitting method were not significantly associated with either sensitivity or specificity, so the corresponding subgroup estimates should not be interpreted as evidence of advantage [58,59]. The bivariate boxplot analysis further identified studies by Geng et al [39], Zhang et al [53], and Wu YP et al [49] as potential outliers contributing to heterogeneity. Residual heterogeneity related to unmodeled technical factors, including segmentation procedures, feature-extraction software, feature-selection strategies, and model-specific operating thresholds, cannot be excluded; therefore, these meta-regression findings should be regarded as exploratory and hypothesis-generating.

AI models may serve as decision-support tools for radiologists; their potential value lies in supporting, not replacing, clinician judgment [60]. Current evidence lacks stratified comparisons by radiologist expertise, access to clinical information, and reading conditions. The additional clinical value of AI should therefore be evaluated through prospectively designed comparisons of AI-assisted diagnosis with manual diagnosis [61]. It is essential to clarify that AI should serve to support and complement clinical decision-making rather than replace the comprehensive judgment of physicians, as cancer diagnosis and treatment involve the integration of complex, multidimensional information [62]. Future investigations into the added value of AI models combined with radiologists compared to manual diagnosis alone may be a worthwhile direction. Current AI model technology is not yet fully mature: models based on single-imaging modalities have inherent limitations, while patients in clinical practice often undergo multiple imaging examinations. Therefore, future research should focus on integrating multimodal imaging information, linking detection results with patients' clinical context, and presenting synthesized diagnostic insights through reporting [63]. In this context, multimodal DL approaches with cross-attention mechanisms have shown promise for accurate cancer diagnosis [64]. Additionally, the application of AI models faces challenges such as the scarcity of expert-annotated data, regulatory barriers, and technical bottlenecks related to data availability, model interpretability, transparency, and the generalization capability of DL models. Advances in few-shot learning, self-supervised models, and centralized platforms may support the development of an integrated AI ecosystem [65]. Despite the considerable potential of AI, its limited interpretability remains a major obstacle to clinical integration and trust. The "black-box" nature of such systems obscures the reasoning behind their decisions, and few existing models can elucidate their own decision-making logic, highlighting the urgent need for future AI systems to enhance transparency and interpretability [66,67]. Transformer-based deep networks have shown effectiveness in medical image denoising [68], and advanced feature extraction and detection techniques continue to improve segmentation accuracy in endoscopic and colorectal applications [69,70]. Accordingly, future work should determine whether AI-assisted, rather than AI vs radiologist, workflows improve clinically meaningful decisions and outcomes [71].

When interpreting the findings of this study, several limitations of the current meta-analysis must be carefully considered. First, the majority of the included studies were retrospective in design, with only Xu et al [50] and Qu et al [44] using prospective designs. The predominance of retrospective studies raises selection and spectrum-bias concerns. Second, when multiple AI models were reported using the same patient cohort, we selected a single model with the highest reported AUC to avoid repeated inclusion of the same participants and disproportionate weighting of individual studies. However, this pragmatic selection strategy may preferentially retain optimistically performing models and introduce selection bias. Accordingly, the pooled estimates should be interpreted as reflecting the selected representative models rather than providing an unbiased comparison or ranking of all candidate AI models. This strategy may therefore overestimate the performance that routinely available models would achieve in real-world deployment. Third, most cohorts were from China, and differences in model architecture, imaging acquisition, analysis unit, and pathology reference standard were not fully standardized. Finally, 2 studies required ROC-derived contingency tables based on Youden-optimal operating points. These thresholds may not reflect prespecified clinical thresholds, and variation in study-specific operating points limits the clinical interpretation of pooled sensitivity and specificity. Together, these factors, the high heterogeneity, and very low certainty of evidence mean that pooled estimates should be interpreted as preliminary and cannot establish clinical superiority or generalizability. In addition, PIs were considerably wider than CIs for most estimates, indicating that performance in a new setting may differ substantially from the pooled average.

Conclusions

Imaging-based AI models showed promising pooled diagnostic performance for preoperative LNM detection in EC. By pairing AI with radiologist performance within the same cohorts, reporting PIs alongside CIs, and distinguishing internal from external validation, this review provides a more realistic assessment of the current evidence than previous meta-analyses based on average effects alone. Across predominantly retrospective and heterogeneous cohorts, AI had higher pooled sensitivity and overall discrimination than radiologists, whereas specificity did not differ; the selective extraction of the highest AUC model and very low certainty of evidence preclude conclusions about clinical superiority. These findings support imaging-based

AI as a decision-support tool rather than a replacement for radiologists; prospective, multicenter external validation with prespecified head-to-head human-AI comparisons is needed before routine clinical adoption.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading, editing, and translation. The GenAI tool used was ChatGPT 4.0. The responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declare that no funding was received for this study.

Data Availability

The original findings of this study are encompassed within the paper. For additional inquiries, contact the corresponding authors.

Authors' Contributions

Conceptualization: CT
Data curation: LX, XG, ZZ, SZ, WD, SG
Formal analysis: CT, ZS, LX, XG, ZZ, SZ, WD, SG
Investigation: LX, XG, ZZ, SZ, WD, SG
Methodology: LX, XG, ZZ, SZ, WD, SG
Software: CT, ZS
Writing – original draft: CT, ZS
Writing – review & editing: LX, XG, ZZ, SZ, WD, SG

Conflicts of Interest

None declared.

Multimedia Appendix 1

Methodological descriptions, additional results, comprehensive search strategies, and quality assessment data.
[DOCX File (Microsoft Word File), 3724 KB-Multimedia Appendix 1]

Checklist 1

PRISMA-DTA checklist.
[DOCX File (Microsoft Word File), 21 KB-Checklist 1]

Checklist 2

PRISMA 2020 abstracts checklist.
[DOCX File (Microsoft Word File), 15 KB-Checklist 2]

Checklist 3

PRISMA-S checklist.
[DOCX File (Microsoft Word File), 16 KB-Checklist 3]

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Abnet CC, Arnold M, Wei WQ. Epidemiology of esophageal squamous cell carcinoma. Gastroenterology. Jan 2018;154(2):360-373. [doi: 10.1053/j.gastro.2017.08.023] [Medline: 28823862]
3. van Hagen P, Hulshof M, van Lanschot JJB, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med. May 31, 2012;366(22):2074-2084. [doi: 10.1056/NEJMoa1112088] [Medline: 22646630]
4. Molena D, Sewell M, Maron SB, et al. The 2025 American Association for Thoracic Surgery (AATS) expert consensus document: surgical management of esophageal and gastroesophageal junction cancer. J Thorac Cardiovasc Surg. Sep 2025;170(3):884-901. [doi: 10.1016/j.jtcvs.2025.04.009] [Medline: 40328426]
5. Le XY, Feng JB, Guo Y, Zhou YQ, Li CM. Predicting preoperative lymph node metastasis in esophageal cancer: advancement and challenges. World J Clin Oncol. Mar 24, 2025;16(3):102863. [doi: 10.5306/wjco.v16.i3.102863] [Medline: 40130042]

6. Dubecz A, Kern M, Solymosi N, Schweigert M, Stein HJ. Predictors of lymph node metastasis in surgically resected T1 esophageal cancer. Ann Thorac Surg. Jun 2015;99(6):1879-1885. [doi: 10.1016/j.athoracsur.2015.02.112] [Medline: 25929888]
7. Huang Y, Du J, Li Q, et al. Threshold size criterion to suspect malignant supraclavicular lymph node <10 mm in esophageal cancer. Insights Imaging. Mar 5, 2025;16(1):50. [doi: 10.1186/s13244-025-01929-3] [Medline: 40045142]
8. Karashima R, Watanabe M, Imamura Y, et al. Advantages of FDG-PET/CT over CT alone in the preoperative assessment of lymph node metastasis in patients with esophageal cancer. Surg Today. Apr 2015;45(4):471-477. [doi: 10.1007/s00595-014-0965-6] [Medline: 24969050]
9. Lei X, Cao Z, Wu Y, et al. Preoperative prediction of clinical and pathological stages for patients with esophageal cancer using PET/CT radiomics. Insights Imaging. Oct 15, 2023;14(1):174. [doi: 10.1186/s13244-023-01528-0] [Medline: 37840068]
10. Leonhardi J, Schnarkowski B, Mehdorn M, et al. Diagnostic accuracy and reliability of CT-based Node-RADS for esophageal cancer. In Vivo. 2025;39(1):353-359. [doi: 10.21873/invivo.13835] [Medline: 39740864]
11. Yamada H, Hosokawa M, Itoh K, et al. Diagnostic value of. Surg Today. Jul 2014;44(7):1258-1265. [doi: 10.1007/s00595-013-0725-z] [Medline: 24077997]
12. Kim JJ, Park JK, Moon SW. Usefulness of positron emission tomography-computed tomography in pre-operative evaluation of intra-thoracic esophageal cancer. Thorac Cancer. Nov 2015;6(6):687-694. [doi: 10.1111/1759-7714.12237] [Medline: 26557905]
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. Aug 2018;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
14. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin. Mar 2019;69(2):127-157. [doi: 10.3322/caac.21552] [Medline: 30720861]
15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. Dec 2017;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
16. Sah BR, Owczarczyk K, Siddique M, Cook GJR, Goh V. Radiomics in esophageal and gastric cancer. Abdom Radiol (NY). Jun 2019;44(6):2048-2058. [doi: 10.1007/s00261-018-1724-8] [Medline: 30116873]
17. Lambin P, Woodruff HC, Mali SA, et al. Radiomics Quality Score 2.0: towards radiomics readiness levels and clinical translation for personalized medicine. Nat Rev Clin Oncol. Nov 2025;22(11):831-846. [doi: 10.1038/s41571-025-01067-1] [Medline: 40903523]
18. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell. May 2022;4(3):e210064. [doi: 10.1148/ryai.210064] [Medline: 35652114]
19. Zhang ST, Wang SY, Zhang J, et al. Artificial intelligence-based computer-aided diagnosis system supports diagnosis of lymph node metastasis in esophageal squamous cell carcinoma: A multicenter study. Heliyon. Mar 2023;9(3):e14030. [doi: 10.1016/j.heliyon.2023.e14030] [Medline: 36923854]
20. Yuan P, Huang ZH, Yang YH, et al. A ^{18}F-FDG PET/CT-based deep learning-radiomics-clinical model for prediction of cervical lymph node metastasis in esophageal squamous cell carcinoma. Cancer Imaging. Nov 12, 2024;24(1):153. [doi: 10.1186/s40644-024-00799-0] [Medline: 39533388]
21. Jannatdoust P, Valizadeh P, Pahlevan-Fallahy MT, et al. Diagnostic accuracy of CT-based radiomics and deep learning for predicting lymph node metastasis in esophageal cancer. Clin Imaging. Sep 2024;113:110225. [doi: 10.1016/j.clinimag.2024.110225] [Medline: 38905878]
22. Ma D, Zhou T, Chen J, Chen J. Radiomics diagnostic performance for predicting lymph node metastasis in esophageal cancer: a systematic review and meta-analysis. BMC Med Imaging. Jun 12, 2024;24(1):144. [doi: 10.1186/s12880-024-01278-5] [Medline: 38867143]
23. Liu L, Liao H, Zhao Y, et al. CT-based radiomics for predicting lymph node metastasis in esophageal cancer: a systematic review and meta-analysis. Front Oncol. 2024;14:1267596. [doi: 10.3389/fonc.2024.1267596] [Medline: 38577325]
24. Chen J, Weng J, Lin X. Accuracy of machine learning in detecting lymph node metastasis of esophageal cancer: a systematic review and meta-analysis. Transl Cancer Res. Jun 30, 2026;15(6):501. [doi: 10.21037/tcr-2026-1-0353] [Medline: 42445376]
25. Bedrikovetski S, Dudi-Venkata NN, Maicas G, et al. Artificial intelligence for the diagnosis of lymph node metastases in patients with abdominopelvic malignancy: a systematic review and meta-analysis. Artif Intell Med. Mar 2021;113:102022. [doi: 10.1016/j.artmed.2021.102022] [Medline: 33685585]
26. Lyu S, Wang Z, Mu Y, Wang L, Pei X. Deep learning algorithms versus radiologists in digital breast tomosynthesis for breast cancer detection: systematic review and meta-analysis. J Med Internet Res. May 6, 2026;28:e91659. [doi: 10.2196/91659] [Medline: 42090319]

27. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. Aug 14, 2020;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
28. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. Mar 24, 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
29. Gopalakrishna G, Mustafa RA, Davenport C, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. Jul 2014;67(7):760-768. [doi: 10.1016/j.jclinepi.2014.01.006] [Medline: 24725643]
30. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008;28(5):621-638. [doi: 10.1177/0272989X08319957] [Medline: 18591542]
31. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res. Dec 2023;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
32. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res. Jun 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
33. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. Jun 15, 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
34. Brini S, Leung TI. Value and credibility of meta-analysis: tutorial on enhancing methodological rigor and AI-powered efficiency. J Med Internet Res. Jul 2, 2026;28:e92132. [doi: 10.2196/92132] [Medline: 42390911]
35. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. Sep 2005;58(9):882-893. [doi: 10.1016/j.jclinepi.2005.01.016] [Medline: 16085191]
36. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. Jul 22, 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
37. Chen L, Ouyang Y, Liu S, et al. Radiomics analysis of lymph nodes with esophageal squamous cell carcinoma based on deep learning. J Oncol. 2022;2022:8534262. [doi: 10.1155/2022/8534262] [Medline: 36147442]
38. Ding M, Cui H, Li B, et al. Integrating preoperative computed tomography and clinical factors for lymph node metastasis prediction in esophageal squamous cell carcinoma by feature-wise attentional graph neural network. Int J Radiat Oncol Biol Phys. Jul 1, 2023;116(3):676-689. [doi: 10.1016/j.ijrobp.2022.12.050] [Medline: 36641040]
39. Geng X, Zhang Y, Li Y, et al. Radiomics-clinical nomogram for preoperative lymph node metastasis prediction in esophageal carcinoma. Br J Radiol. Feb 28, 2024;97(1155):652-659. [doi: 10.1093/bjr/tqae009] [Medline: 38268475]
40. Huang YL, Yan C, Lin X, et al. The development of a nomogram model for predicting left recurrent laryngeal nerve lymph node metastasis in esophageal cancer based on radiomics and clinical factors. Ann Transl Med. Dec 2022;10(23):1282. [doi: 10.21037/atm-22-5628] [Medline: 36618793]
41. Li X, Liu Q, Hu B, Xu J, Huang C, Liu F. A computed tomography-based clinical-radiomics model for prediction of lymph node metastasis in esophageal carcinoma. J Cancer Res Ther. Dec 2021;17(7):1665-1671. [doi: 10.4103/jcrt.jcrt_1755_21] [Medline: 35381737]
42. Ou J, Wu L, Li R, et al. CT radiomics features to predict lymph node metastasis in advanced esophageal squamous cell carcinoma and to discriminate between regional and non-regional lymph node metastasis: a case control study. Quant Imaging Med Surg. Feb 2021;11(2):628-640. [doi: 10.21037/qims-20-241] [Medline: 33532263]
43. Peng G, Zhan Y, Wu Y, et al. Radiomics models based on CT at different phases predicting lymph node metastasis of esophageal squamous cell carcinoma (GASTO-1089). Front Oncol. 2022;12:988859. [doi: 10.3389/fonc.2022.988859] [Medline: 36387160]
44. Qu J, Shen C, Qin J, et al. The MR radiomic signature can predict preoperative lymph node metastasis in patients with esophageal cancer. Eur Radiol. Feb 2019;29(2):906-914. [doi: 10.1007/s00330-018-5583-z] [Medline: 30039220]
45. Shen C, Liu Z, Wang Z, et al. Building CT radiomics based nomogram for preoperative esophageal cancer patients lymph node metastasis prediction. Transl Oncol. Jun 2018;11(3):815-824. [doi: 10.1016/j.tranon.2018.04.005] [Medline: 29727831]
46. Tan X, Ma Z, Yan L, Ye W, Liu Z, Liang C. Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma. Eur Radiol. Jan 2019;29(1):392-400. [doi: 10.1007/s00330-018-5581-1] [Medline: 29922924]
47. Wang ZL, Zhou ZG, Chen Y, Li XT, Sun YS. Support vector machines model of computed tomography for assessing lymph node metastasis in esophageal cancer with neoadjuvant chemotherapy. J Comput Assist Tomogr. 2017;41(3):455-460. [doi: 10.1097/RCT.0000000000000555] [Medline: 27879527]

48. Wu H, Wu X, Miao S, et al. Deep learning-based prediction of enhanced CT scans for lymph node metastasis in esophageal squamous cell carcinoma. Jpn J Radiol. Aug 2025;43(8):1347-1356. [doi: 10.1007/s11604-025-01780-y] [Medline: 40214915]
49. Wu YP, Wu L, Ou J, et al. Preoperative identification of small metastatic lymph nodes in esophageal squamous cell carcinoma using CT radiomics of lymph nodes. Abdom Radiol (NY). Mar 2025;50(3):1123-1132. [doi: 10.1007/s00261-024-04585-1] [Medline: 39305294]
50. Xu YH, Lu P, Gao MC, et al. Nomogram based on multimodal magnetic resonance combined with B7-H3mRNA for preoperative lymph node prediction in esophagus cancer. World J Clin Oncol. Mar 24, 2024;15(3):419-433. [doi: 10.5306/wjco.v15.i3.419] [Medline: 38576593]
51. Yi Y, Wang J, Li Z, et al. Multi-task and multi-scale attention network for lymph node metastasis prediction in esophageal cancer. Med Biol Eng Comput. Nov 2025;63(11):3251-3262. [doi: 10.1007/s11517-025-03391-1] [Medline: 40488960]
52. Zhang C, Shi Z, Kalendralis P, et al. Prediction of lymph node metastases using pre-treatment PET radiomics of the primary tumour in esophageal adenocarcinoma: an external validation study. Br J Radiol. Feb 1, 2021;94(1118):20201042. [doi: 10.1259/bjr.20201042] [Medline: 33264032]
53. Zhang Y, Liu L, Han M, Li L, Wu Q, Wang X. Unraveling the power of radiomics: prediction and exploration of lymph node metastasis in stage T1/2 esophageal squamous cell carcinoma. NPJ Precis Oncol. Jun 16, 2025;9(1):176. [doi: 10.1038/s41698-025-00929-2] [Medline: 40523899]
54. Peng Q, Zhou S, Chen R, et al. CT-based habitat radiomics combining multi-instance learning for early prediction of post-neoadjuvant lymph node metastasis in esophageal squamous cell carcinoma. Bioengineering (Basel). Jul 28, 2025;12(8):813. [doi: 10.3390/bioengineering12080813] [Medline: 40868326]
55. Zhao Z, Xie Y, Lai D, Liang J, Okereke IC, Lin W. Transcriptome analysis and artificial intelligence for predicting lymph node metastasis of esophageal squamous cell carcinoma. J Thorac Dis. May 30, 2025;17(5):3283-3296. [doi: 10.21037/jtd-2025-662] [Medline: 40529736]
56. Chen H, Zhou X, Tang X, Li S, Zhang G. Prediction of lymph node metastasis in superficial esophageal cancer using a pattern recognition neural network. Cancer Manag Res. 2020;12:12249-12258. [doi: 10.2147/CMAR.S270316] [Medline: 33273861]
57. Mehrabi Nejad MM, Ghanbari Boroujeni MR, Hayati A, Dashti F, Udupa JK, Torigian DA. Performance of AI methods in PET-based imaging for outcome prediction in lymphoma: a systematic review and meta-analysis. Eur J Radiol. Sep 2025;190:112204. [doi: 10.1016/j.ejrad.2025.112204] [Medline: 40466216]
58. Huang X, Wang Q, Xu W, et al. Machine learning to predict lymph node metastasis in T1 esophageal squamous cell carcinoma: a multicenter study. Int J Surg. Dec 1, 2024;110(12):7852-7859. [doi: 10.1097/JS9.0000000000001694] [Medline: 38905510]
59. Varnava Y, Jakate K, Garnett R, Androutsos D, Tyrrell PN, Khademi A. Out-of-distribution generalization for segmentation of lymph node metastasis in breast cancer. Sci Rep. Jan 7, 2025;15(1):1127. [doi: 10.1038/s41598-024-80495-y] [Medline: 39775089]
60. Horie Y, Yoshio T, Aoyama K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest Endosc. Jan 2019;89(1):25-32. [doi: 10.1016/j.gie.2018.07.037] [Medline: 30120958]
61. Hong Y, Wang H, Zhang Q, et al. Machine learning and deep learning hybrid approach based on muscle imaging features for diagnosis of esophageal cancer. Diagnostics (Basel). Jul 8, 2025;15(14):1730. [doi: 10.3390/diagnostics15141730] [Medline: 40722480]
62. Lu C, Zhang J, Liu R. Deep learning-based image classification for integrating pathology and radiology in AI-assisted medical imaging. Sci Rep. Jul 25, 2025;15(1):27029. [doi: 10.1038/s41598-025-07883-w] [Medline: 40715157]
63. Rao VM, Hla M, Moor M, et al. Multimodal generative AI for medical image interpretation. Nature. Mar 2025;639(8056):888-896. [doi: 10.1038/s41586-025-08675-y] [Medline: 40140592]
64. Sarwar S, Majeed S, Nawaz A, Bibi R, Lee SW. MDL-CA: a multimodal deep learning approach with a cross attention mechanism for accurate brain cancer diagnosis. Front Public Health. 2026;13:1687335. [doi: 10.3389/fpubh.2025.1687335] [Medline: 41561860]
65. Naqvi RA, Haider A, Kim HS, Jeong D, Lee SW. Transformative noise reduction: leveraging a transformer-based deep network for medical image denoising. Mathematics. 2024;12(15):2313. [doi: 10.3390/math12152313]
66. Abidin ZU, Naqvi RA, Islam MZ, et al. Dual-task hierarchical feature refinement and fusion network for precise segmentation of surgical tools and polyps in endoscopy. Expert Syst Appl. Dec 2025;293:128618. [doi: 10.1016/j.eswa.2025.128618]

67. Jafar A, Abidin ZU, Naqvi RA, Lee SW. Unmasking colorectal cancer: a high-performance semantic network for polyp and surgical instrument segmentation. Eng Appl Artif Intell. Dec 2024;138:109292. [doi: 10.1016/j.engappai.2024.109292]
68. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology. Jun 2025;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]
69. Castelvecchi D. Can we open the black box of AI? Nature. Oct 6, 2016;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
70. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. Eur J Radiol. Dec 2023;169:111159. [doi: 10.1016/j.ejrad.2023.111159] [Medline: 37976760]
71. Baik YS, Lee H, Kim YJ, Chung JW, Kim KG. Early detection of esophageal cancer: evaluating AI algorithms with multi-institutional narrowband and white-light imaging data. PLoS ONE. 2025;20(4):e0321092. [doi: 10.1371/journal.pone.0321092] [Medline: 40184395]

Abbreviations

[^{18}F] FDG PET/CT: [^{18}F] fluorodeoxyglucose positron emission tomography/computed tomography
AUC: area under the curve
CECT: contrast-enhanced computed tomography
CEMRI: contrast-enhanced magnetic resonance imaging
DL: deep learning
DOR: diagnostic odds ratio
EAC: esophageal adenocarcinoma
EC: esophageal cancer
ESCC: esophageal squamous cell carcinoma
FN: false negative
FP: false positive
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LNM: lymph node metastasis
ML: machine learning
PI: prediction interval
PITROS: participants, index test, target condition, reference standard, outcomes, and setting
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool with AI
PROSPERO: International Prospective Register of Systematic Reviews
ROC: receiver operating characteristic
TN: true negative
TP: true positive

Edited by Stefano Brini; peer-reviewed by Fangzhou Liu, Hassaporn Thongdaeng, Sadra Behrouzieh, Seung Won Lee; submitted 23.Mar.2026; final revised version received 21.Sep.2026; accepted 22.Sep.2026; published 26.Nov.2026

Please cite as:
Tan C, Xia L, Gao X, Zhong Z, Zhang S, Deng W, Gui S, Shang Z
Diagnostic Performance of Imaging-Based AI vs Radiologists for Preoperative Lymph Node Metastasis in Esophageal Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e95893
URL: https://www.jmir.org/2026/1/e95893
doi: 10.2196/95893

© Cong Tan, Lingfeng Xia, Xiongying Gao, Zhilin Zhong, Shuang Zhang, Wenjun Deng, Shen Gui, Zhaojun Shang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.