Artificial intelligence models in predicting lymph node metastasis in early gastric cancer: a systematic review and meta-analysis

XiaoPeng Chen, ZhengGuo Yi, JinPing Ye

Department of General Surgery, Shangrao Municipal Hospital, Shangrao, Jiangxi, China

KEY WORDS

artificial intelligence, diagnostic accuracy, early gastric cancer, lymph node metastasis

ABSTRACT

INTRODUCTION Accurate preoperative assessment of lymph node metastasis (LNM) is a key determinant of treatment selection in early gastric cancer (EGC), particularly when choosing between endoscopic resection and minimally-invasive gastrectomy with lymphadenectomy. Although artificial intelligence (AI)-based models have been increasingly developed for LNM prediction, their overall diagnostic performance and clinical relevance to minimally-invasive treatment decision-making remain unclear.

AIM This systematic review and meta-analysis aimed to evaluate the diagnostic accuracy of AI-based models for predicting LNM in EGC, and to clarify their potential role in guiding minimally-invasive and endoscopic treatment strategies.

MATERIALS AND METHODS A comprehensive literature search of PubMed, Embase, and Web of Science databases was conducted through August 2025. Studies applying machine learning or deep learning algorithms to predict LNM in EGC were included. Study quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. Pooled sensitivity, specificity, and area under the curve (AUC) were calculated using a bivariate random-effects model. Subgroup analyses, meta-regression, and publication bias assessment were performed.

RESULTS A total of 18 studies involving 41 505 patients were included. In the internal validation cohorts, AI-based models demonstrated a pooled sensitivity of 0.81 and specificity of 0.82, with the AUC of 0.88. Comparable performance was observed in the external validation cohorts (sensitivity, 0.81; specificity, 0.84; AUC, 0.9), indicating good generalizability. In the studies directly comparing AI with clinician assessment, AI models consistently achieved higher sensitivity and overall diagnostic accuracy.

CONCLUSIONS AI-based models show robust performance for predicting LNM in EGC and outperform clinician assessment. Importantly, these models have the potential to serve as clinically meaningful decision-supporting tools for minimally-invasive and endoscopic management by assisting in the selection between endoscopic resection and gastrectomy with lymphadenectomy, thereby optimizing surgical extent while preserving oncological safety.

Correspondence to:
ZhengGuo Yi, MD, Department of General Surgery, Shangrao Municipal Hospital, 182 Wusan Avenue, Shangrao, 334001 Jiangxi, China, phone: +86 18376590219, email: dxllpwan@126.com
Received: December 4, 2025.
Revision accepted: December 23, 2025.
Published online: January 21, 2026.
Wideochir Inne Tech Maloinwazyjne. 2026; 21 (1): 1-12
doi:10.20452/wiitm.2026.18006

INTRODUCTION Gastric cancer (GC) remains one of the leading causes of cancer-related deaths globally, with a particularly high incidence in East Asia.[1] Early GC (EGC), which is confined to the mucosa or submucosa, offers a significantly better prognosis if detected and treated appropriately.[2] However, the detection of lymph node metastasis (LNM), which occurs in approximately 20%–30% of the patients with EGC, is crucial for determining the appropriate therapeutic approach, including the need for extended lymphadenectomy or more conservative treatment.[3] Therefore, accurate identification of LNM in the early stages of GC can help avoid over- or undertreatment, both of which are associated with poor prognosis.

Traditional diagnostic methods for assessing LNM in EGC, including computed tomography (CT), magnetic resonance imaging (MRI), and endoscopic ultrasound (EUS), have been widely

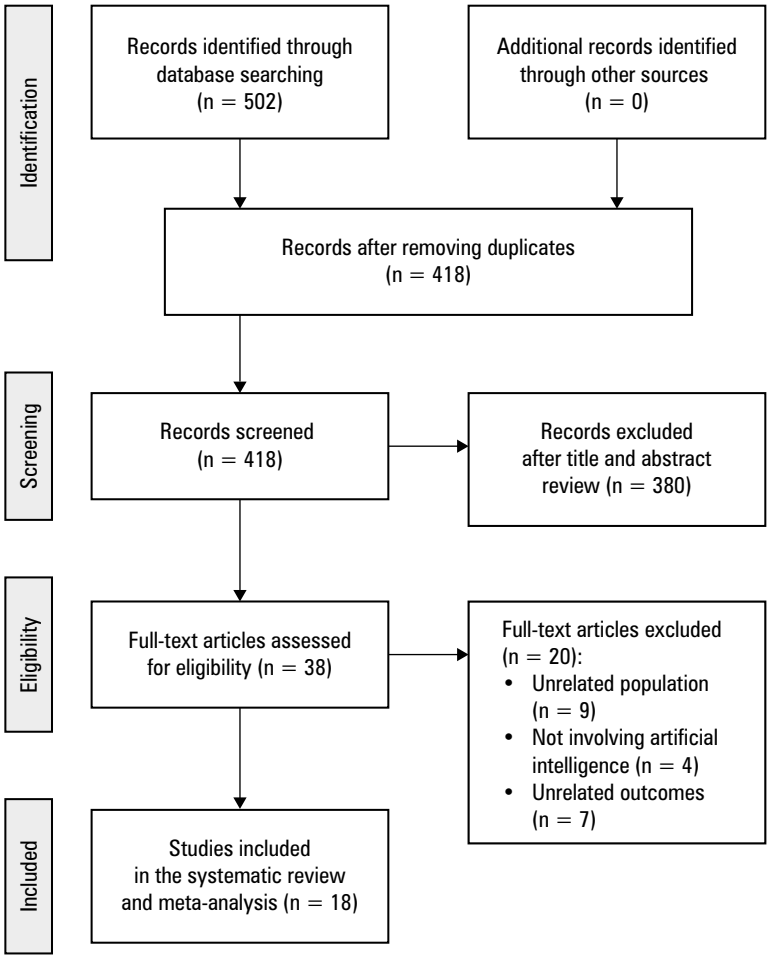

FIGURE 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart of literature selection

utilized. While these techniques provide valuable imaging data, their sensitivity and accuracy for detecting small or microscopic metastatic LNs, especially in EGC, remain limited.[4] CT and MRI are often incapable of detecting small LNs that may harbor microscopic metastasis, leading to either unnecessary surgeries or missed opportunities for more aggressive treatment.[5] Similarly, EUS, although beneficial in assessing larger LNs, is less effective in detecting micrometastasis, and its accuracy heavily depends on the operator's skill and experience.[6] These limitations underscore an urgent need for more effective and precise diagnostic methods to assess LNM in EGC.

In recent years, artificial intelligence (AI), particularly deep learning (DL) and machine learning (ML), has emerged as a promising tool for revolutionizing cancer diagnostics.[7,8] AI can analyze vast amounts of data from various sources, including medical images, histopathological slides, and clinical data, to identify complex patterns that may not be immediately apparent to the human eye.[9] This capability positions AI as an ideal technology for improving the accuracy of LNM detection in EGC.[10,11] Numerous studies have explored the application of AI in detecting LNM in GC, using a variety of imaging modalities, including CT, MRI, and EUS. For example, a recent study demonstrated that DL models could significantly improve the accuracy of histopathological diagnosis of EGC, achieving an overall area under the curve (AUC) of 0.75 in predicting LNM status.[12] Additionally, AI models trained on endoscopic images have shown promise in enhancing the detection of LNM, offering higher diagnostic performance than traditional methods.[10] Despite these advancements, the reported performance of AI-based LNM prediction models remains heterogeneous, with considerable variability in study design, sample size, model architecture, input features, and validation strategies. Furthermore, the generalizability and clinical utility of these algorithms have not been systematically evaluated.

AIM This systematic review and meta-analysis aimed to provide a comprehensive evaluation of AI applications for LNM detection in EGC. By reviewing the methodologies, performance metrics, and limitations of AI models used in current studies, we aimed to identify key trends, challenges, and gaps in the field. Furthermore, we discussed potential implications of AI for clinical practice and proposed future research directions to overcome existing barriers and improve the integration of AI into diagnostic workflows. Beyond diagnostic accuracy, accurate preoperative prediction of LNM is fundamental to minimally-invasive treatment planning in EGC. The decision to pursue endoscopic resection or gastrectomy with lymphadenectomy relies heavily on nodal status assessment, as inappropriate patient selection may result in noncurative endoscopic resection or unnecessary surgical morbidity. Therefore, AI-based prediction models should be evaluated not only as diagnostic tools, but also as decision-supporting instruments within minimally-invasive and endoscopic surgical workflows.

MATERIALS AND METHODS This study was designed as a systematic review and meta-analysis to evaluate the diagnostic performance of AI-based models for predicting LNM in patients with EGC. The review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies guidelines.[13] The study protocol was prospectively registered in the Open Science Framework (https://osf.io/grf8q/). Since all data were obtained from previously published studies, ethical approval and patient informed consent were not required.

Search strategy To ensure a comprehensive review of the literature, a systematic search was conducted across 3 electronic databases: PubMed, Embase, and Web of Science. The search was performed from the inception of each database until August 2025 without any language limitation. The search terms included combinations of the following key words: "artificial intelligence," "machine learning," "deep learning,"

TABLE 1 Characteristics of the included studies on using artificial intelligence to detect lymph node metastasis in patients with early gastric cancer

Author, year	Country	Study type	Study design	Patient inclusion period	GC staging	Algorithm used in AI models	Variables used in AI models	Training set, n	Test set, n	Validation set, n
Wang et al, 2021[15]	China	Single-center	Retrospective cohort	2012–2017	T1–T2	mRMR	Station 3 lymph nodes and primary tumor radiomics	80	79	NA
Zhu et al, 2022[20]	China	Multicenter	Retrospective cohort	NA	T1	GBM, XGBoost, RF, DT, and NNET	Clinical features	1878	470	NA
Tian et al, 2022[14]	China	Multicenter	Case control study	2010–2015	T1a–T1b	GLM, RPART, RF, GBM, SVM, RDA, and NNET	Clinical features	1839	458	227
Na et al, 2022[16]	Korea	Single-center	Retrospective cohort	2005–2021	T1a	LR, SVM, and RF	Clinical features	10332	NA	4428
Zeng et al, 2022[19]	China	Single-center	Retrospective cohort	2016–2021	T1a–T1b	Pretrained deep learning networks	Deep transfer learning, radiomics, and clinical features	388	167	79
Yang et al, 2022[18]	China	Multicenter	Retrospective cohort	2012–2021	T1a–T1b	Linear SVC, LR, XGBoost, LightGBM, and Gaussian process classification model	Clinical features	305	NA	35
Wei et al, 2022[17]	China	Multicenter	Retrospective cohort	2015–2021	EGC	RFC, DT, SVM, XGBoost, GLM, and ANN	MRI parameters	368	158	NA
Lee et al, 2023[23]	Korea	Single-center	Retrospective cohort	2012–2020	EGC	GBM and LR	Clinical features	2044	512	548
Hayashi et al, 2023[22]	Japan	Single-center	Prospective cohort	2013–2018	T1b	XGBoost	Clinical and pathological variables	382	NA	140
Dong et al, 2023[21]	China	Single-center	Retrospective cohort	2017–2022	T1–T2	10-lncRNA risk--prediction model	Genome-wide expression profiles of lncRNA	20	98	127
Seo et al, 2024[25]	Korea	Multicenter	Retrospective cohort	2007–2017	T1b	Logistic regression, RF, XGBoost, and SVM	Clinical and pathological variables	2426	NA	1042
Kato et al, 2024[11]	Japan	Multicenter	Retrospective cohort	2010–2021	EGC	Neural network	Clinical and pathological variables	3506	NA	536
Yang et al, 2024[10]	China	Single-center	Retrospective cohort	2016–2023	EGC	CNN	Endoscopic images	54	24	30
Lee et al, 2024[24]	Korea	Multicenter	Retrospective cohort	2018–2023	EGC	CNN	Endoscopic images and videos	4336 images and 153 videos	260 images and 10 videos	436 images and 89 videos
Sung et al, 2024[12]	Korea	Multicenter	Retrospective cohort	NA	EGC	DeepLabV3+ and XGBoost	Hematoxylin and eosin–stained images	NA	NA	NA
Kang et al, 2025[28]	Korea	Multicenter	Retrospective cohort	2010–2015	EGC	CNN and CNN with RF	Endoscopic images, demographic data, biopsy pathology, CT findings	2927	449	766
He et al, 2025[27]	China	Multicenter	Retrospective cohort	2006–2019	pT1N0	2.5D MIL-based model	Preoperative portal venous phase CT images	1953	NA	1211
Gao et al 2025[26]	China	Single-center	Retrospective cohort	NA	T1	VGG16, ResNet34, MobileNetV2, and PVTv2	Morphological features of collagen fibers from multiphoton microscopy	143	69	NA

Abbreviations: AI, artificial intelligence; ANN, artificial neural network; CNN, convolutional neural network; CT, computed tomography; DT, decision tree; EGC, early gastric cancer; FCNN, fully convolutional neural network; GBM, gradient boosting machine; GC, gastric cancer; GLM, generalized linear model; LR, logistic regression; MIL, multiple instance learning; MRI, magnetic resonance imaging; mRMR, minimum redundancy maximum relevance; NA, not applicable; NNET, neural network; PVT, pyramid vision transformer; RDA, regularized dual averaging; RF, random forest; RFC, random forest classifier; RPART, recursive partitioning and regression tree; SVC, support vector classifier; SVM, support vector machine; VGG, visual geometry group; XGBoost, extreme gradient boosting

TABLE 2 Quality Assessment of Diagnostic Accuracy Studies-2 evaluation of the risk of bias

Author, year	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Analysis	Patient selection	Index test	Reference standard
Wang et al, 2021[15]	High	Low	Unclear	Unclear	Low	Low	Low
Zhu et al, 2022[20]	Unclear	Low	Unclear	Unclear	Low	Low	Low
Tian et al, 2022[14]	High	High	Low	High	Low	Low	Low
Na et al, 2022[16]	High	High	Low	Low	Unclear	Low	Low
Zeng et al, 2022[19]	High	High	Low	Low	Low	Low	Low
Yang et al, 2022[18]	High	High	Unclear	Unclear	Low	Low	Low
Wei et al, 2022[17]	High	High	Low	High	Low	Low	Low
Lee et al, 2023[23]	High	High	Low	Low	Low	Low	Low
Hayashi et al, 2023[22]	High	High	Unclear	Unclear	Low	Low	Low
Dong et al, 2023[21]	High	Unclear	Low	Low	Low	Low	Low
Seo et al, 2024[25]	High	Unclear	Low	Low	Low	Low	Low
Kato et al, 2024[11]	High	Unclear	Low	Low	Low	Low	Low
Yang et al, 2024[10]	High	High	Low	High	Low	Low	Low
Lee et al, 2024[24]	Low	Low	Low	Low	Low	Low	Low
Sung et al, 2024[12]	High	Low	Low	Unclear	Low	Low	Low
Kang et al, 2025[28]	High	Low	Low	High	Low	Low	Low
He et al, 2025[27]	High	Low	Low	High	Low	Low	Low
Gao et al, 2025[26]	High	Low	Low	High	Low	Low	Low

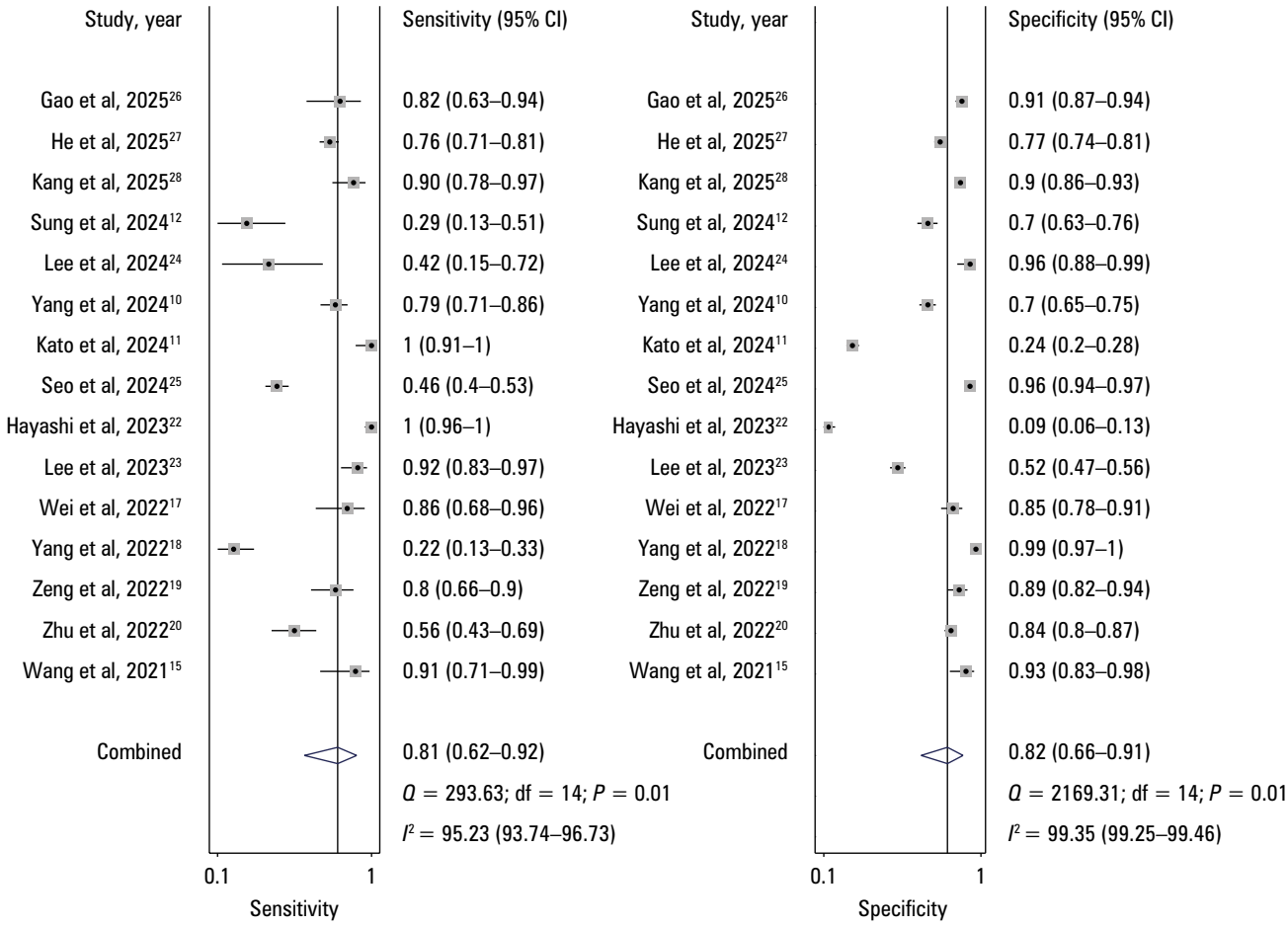

FIGURE 2 Forest plots of pooled sensitivity and specificity of artificial intelligence models for predicting lymph node metastasis in early gastric cancer in internal validation cohorts. Each horizontal line represents 95% CI for an individual study, while the diamond indicates the pooled estimate. Significant heterogeneity across the studies was observed.

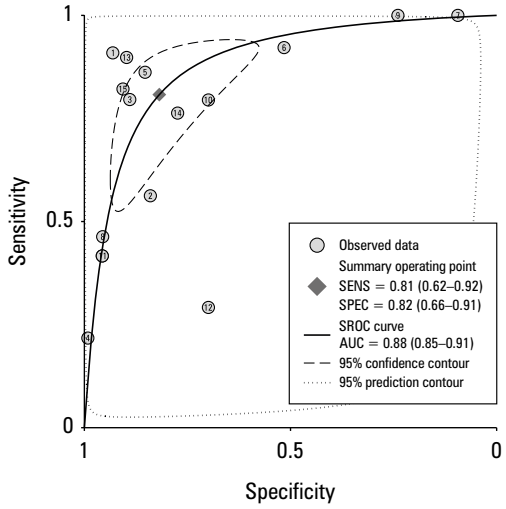

FIGURE 3 SROC curve illustrating the diagnostic performance of artificial intelligence models for lymph node metastasis prediction in internal validation datasets. The curve demonstrates strong overall discriminative ability, with the AUC approaching 0.88.

Abbreviations: AUC, area under the curve; SENS, sensitivity; SPEC, specificity; SROC, summary receiver operating characteristic

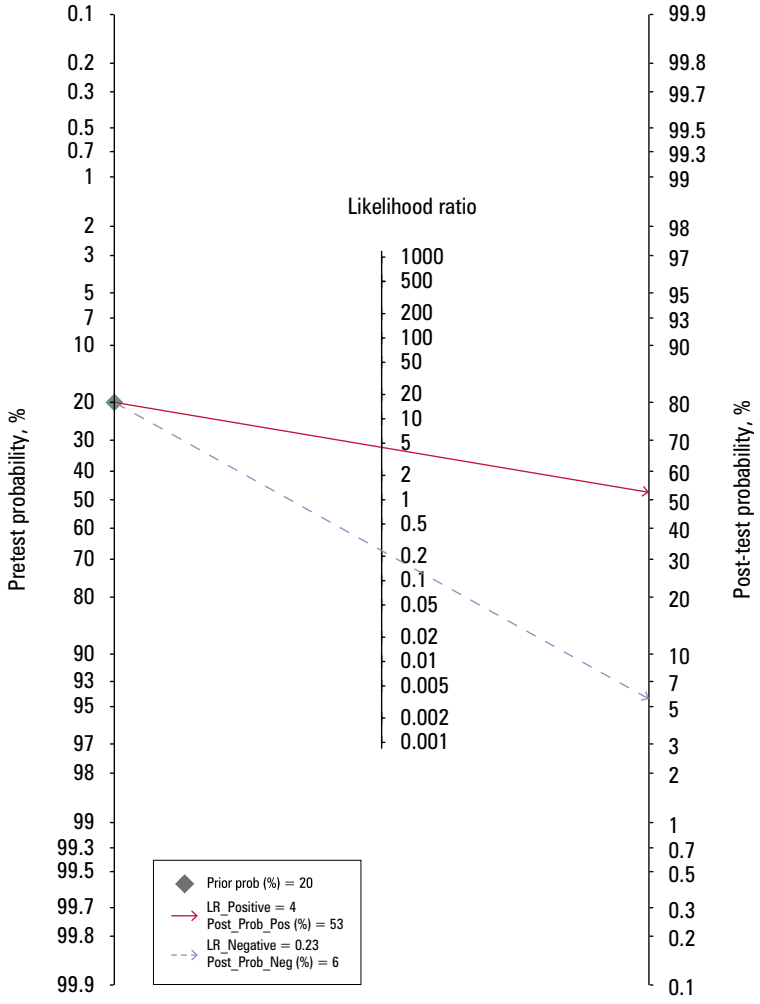

FIGURE 4 Fagan nomogram for pretest and post-test probability estimation based on artificial intelligence model performance in internal validation. A positive test result substantially increases the probability of lymph node metastasis, supporting clinical decision-making for surgical planning.

Abbreviations: LR, likelihood ratio; Post_Prob_Neg, post-test probability negative; Post_Prob_Pos, post-test probability positive; Prior prob, prior probability; others, see **TABLE 1**

"gastric cancer," "early gastric cancer," "lymph node metastasis," "diagnosis," and "detection." Boolean operators (AND, OR) were used to refine the search and maximize retrieval of relevant studies. Additionally, manual searches were conducted through the reference lists of included articles to identify additional studies that may have been overlooked during the database search.

Inclusion and exclusion criteria To be included in the analysis, the studies had to meet the following criteria: 1) study design: retrospective or prospective diagnostic accuracy studies; case reports, reviews, meta-analyses, and opinion pieces were excluded; 2) population: studies involving adult patients with EGC, including those with and without LNM; 3) index test: AI-based models, including ML, DL, or hybrid approaches, designed to predict the presence of LNM; and 4) outcome measures: studies that reported performance metrics, such as sensitivity, specificity, accuracy, AUC, or other relevant diagnostic performance metrics for AI-based methods in detecting LNM.

Studies were excluded if they focused on late-stage GC, AI was not used as the primary diagnostic tool for LNM detection, or the full text of the article was not available.

Data extraction Two independent reviewers (CXP and YJP) performed the data extraction using a predesigned, standardized form. Discrepancies between them were resolved through consensus or with the assistance of a third reviewer (YGZ). The following data were extracted from each study: 1) study characteristics: first author, publication year, country, study design, sample size, patient demographics; 2) AI model details: algorithm type (eg, convolutional neural network, support vector machine), input features (clinical, pathological, imaging, or multimodal), and validation strategy; 3) validation cohorts: type of validation (internal vs external) and whether physician assessment was included as a comparator; and 4) diagnostic performance: true positives (TPs), false positives (FPs), false negatives (FNs), true negatives (TNs), sensitivity, specificity, and AUC values.

Since most of the included studies did not explicitly report diagnostic contingency tables, we reconstructed the 2 × 2 tables using 2 complementary approaches. First, when sufficient data were available, we derived the numbers of TPs, FPs, FNs, and TNs based on the reported sensitivity, specificity, total sample size, and the number of cases confirmed as positive according to the reference standard. Second, in the studies lacking sufficient numerical details, we estimated these values by extracting the optimal sensitivity and specificity from the receiver operating characteristic (ROC) curve using the Youden index. It is important to acknowledge that the latter approach may introduce potential bias, as the cutoff point determined by the ROC curve may not precisely mirror clinical practice. This discrepancy could result in case misclassification and, consequently, influence the calculated contingency table values.

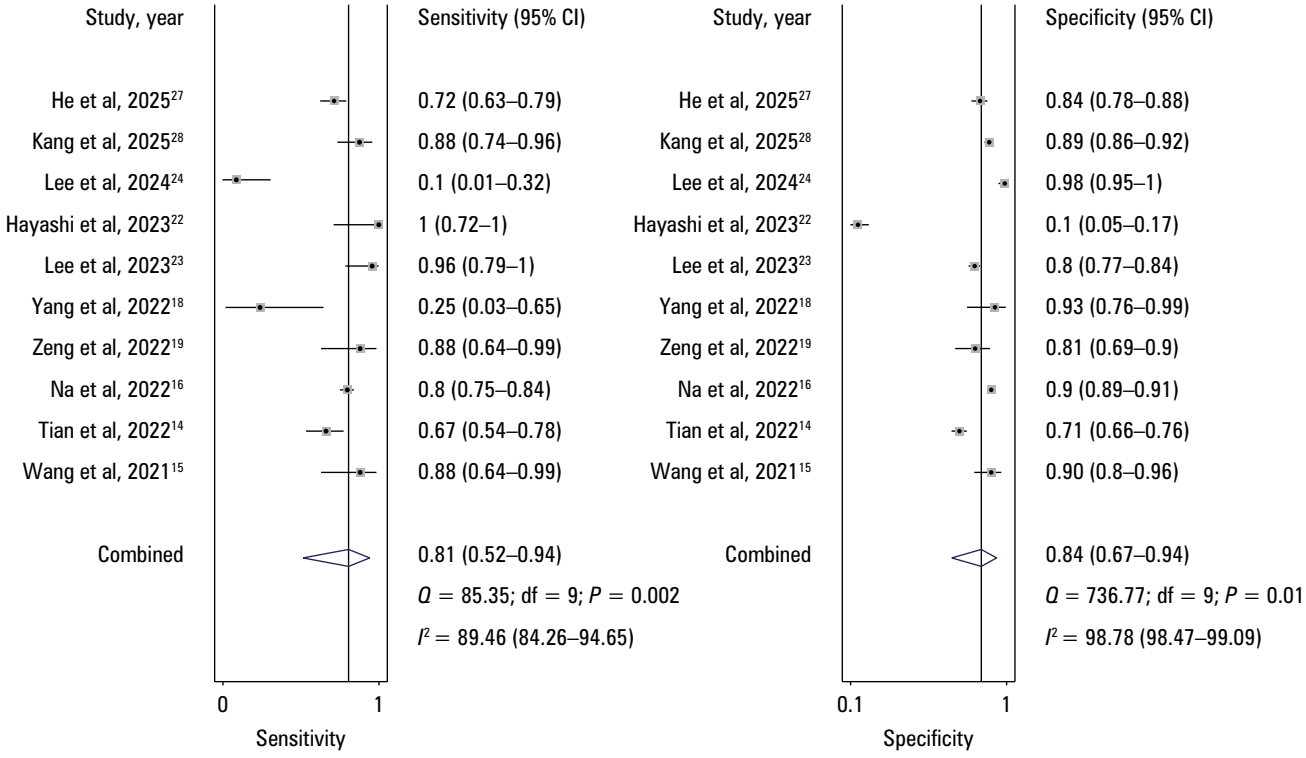

FIGURE 5 Forest plots of pooled sensitivity and specificity of artificial intelligence models in external validation cohorts. The pooled estimates demonstrate consistent diagnostic performance across independent patient populations.

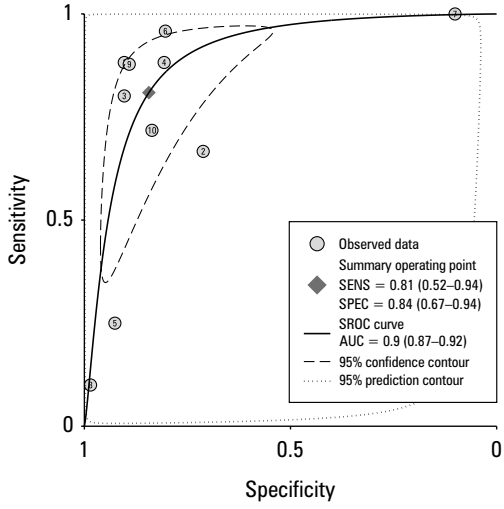

FIGURE 6 SROC curve for external validation datasets, showing robust discriminative performance of artificial intelligence–based prediction models for lymph node metastasis. The pooled AUC was approximately 0.9, confirming external generalizability.

Abbreviations: see **FIGURE 3**

Quality assessment The methodological quality of the included studies was evaluated using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.[14] Four domains were assessed: patient selection, index test, reference standard, and analysis. Each domain was rated as having a "low," "high," or "unclear" risk of bias, and concerns regarding applicability were similarly evaluated. The assessments were performed independently by 2 reviewers (CXP and YJP), with disagreements resolved by consensus.

Statistical analysis A bivariate random-effects model was employed to synthesize diagnostic accuracy estimates across the studies and evaluate the performance of AI-based models in predicting LNM in EGC. The pooled sensitivity, specificity, and corresponding 95% CIs were calculated separately for the internal validation datasets, external validation datasets, and physician assessments (endoscopic, radiological, or pathological), where applicable. Forest plots were constructed to visually summarize the pooled sensitivity and specificity estimates, while summary ROC (SROC) curves were generated to depict the overall diagnostic performance and provide combined AUC estimates with their 95% CIs and prediction intervals. Between-study heterogeneity was assessed using the Higgins I^2 statistic, with values of 25%, 50%, and 75% representing low, moderate, and high heterogeneity, respectively. In the instances where substantial heterogeneity was observed (I^2 >50%) and the number of included datasets exceeded 10, meta-regression analyses were performed to explore potential sources of heterogeneity. Covariates considered in the meta-regression included AI model type (eg, ML vs DL), validation strategy (internal vs external), type of input features (clinical, imaging, or multimodal), and study design (retrospective cohort vs case control). Additionally, univariate subgroup analyses were conducted to further examine their potential effects on diagnostic accuracy in the internal validation.

Potential publication bias was evaluated using the Deeks funnel plot asymmetry test, with a P value below 0.05 indicating significant small-study effects. Moreover, Fagan nomograms were constructed to estimate post-test probabilities and assess the clinical utility of AI-based models across different pretest probability scenarios.

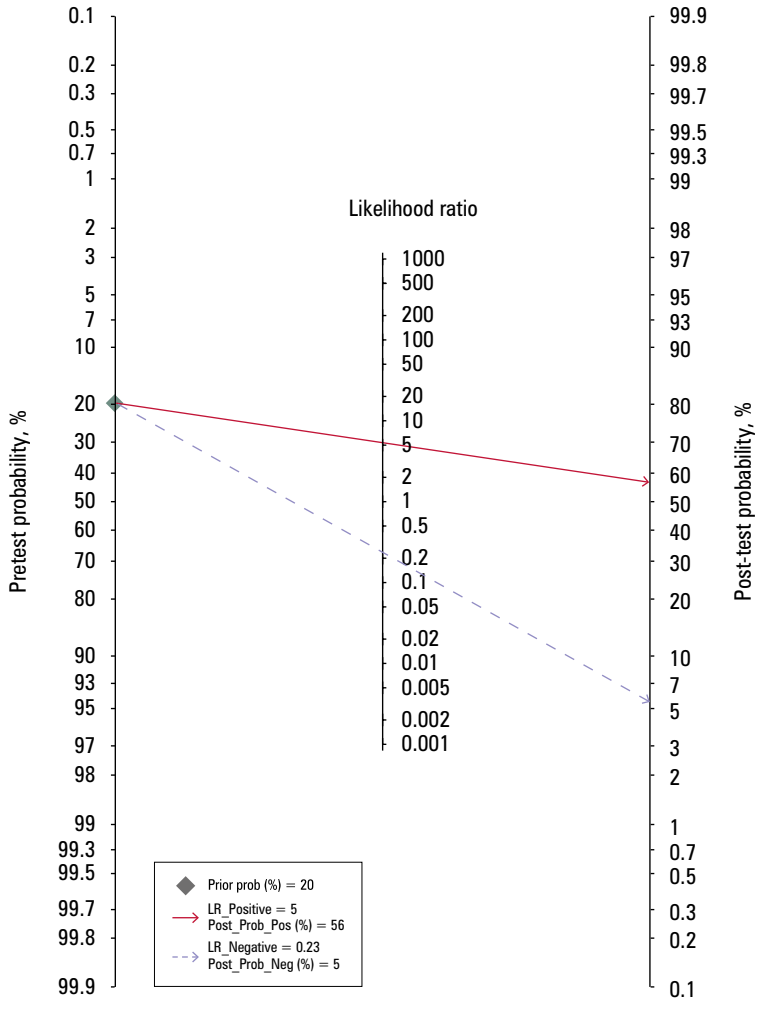

FIGURE 7 Fagan nomogram evaluating the clinical utility of artificial intelligence models in external validation. Post-test probabilities illustrate the potential of artificial intelligence–assisted prediction to stratify patients for tailored surgical approaches.

Abbreviations: see **TABLE 1** and **FIGURE 4**

All statistical analyses were conducted using the Midas and Metadata packages in Stata, version 14.0 (StataCorp, College Station, Texas, United States). All statistical tests were 2-sided, and a *P* value below 0.05 was considered significant. The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

RESULTS **Study selection** A total of 502 records were identified during the search phase. After eliminating 84 duplicates, 418 unique articles remained for screening. Upon careful evaluation of the titles and abstracts, 380 articles were excluded as not meeting the predefined criteria. Subsequent in-depth scrutiny of full texts led to the inclusion of 18 articles for analysis.[10-12,15-29] The included studies were conducted primarily in China, Korea, and Japan, spanning from 2021 to 2025. These investigations centered on the application of AI in predicting LNM in EGC. Various AI methodologies, encompassing ML and DL models, were utilized across the studies. The detailed literature screening process is presented in **FIGURE 1**.

Characteristics of included studies A total of 18 studies involving 41 505 patients were included, with all studies adopting retrospective or prospective cohort designs. All included studies utilized datasets from East Asian populations, reflecting the higher incidence of EGC in these regions. Input features for the AI models varied, including clinical characteristics, imaging data, histopathological features, and genomic profiles. ML techniques, such as random forest, gradient boosting machines, and support vector machines, were commonly used, alongside DL methods, such as convolutional neural networks. Sample sizes ranged widely, with training datasets containing between 20 and 10 332 cases. Many studies lacked external validation, limiting their generalizability, while others employed robust validation strategies, achieving higher levels of reliability. Reported performance metrics, such as sensitivity, specificity, accuracy, and AUC, highlighted the potential of AI to enhance LNM detection. The detailed characteristics of the included studies are outlined in **TABLE 1**.

Quality assessment The methodological quality of the included studies, evaluated using the QUADAS-2 tool, is summarized in **TABLE 2**. Overall, the risk of bias was deemed low to moderate across most domains. However, patient selection bias was frequently rated as high (13/17 studies), largely due to retrospective designs and nonconsecutive enrollment. Index test bias was low in most cases, reflecting consistent application of AI algorithms. Concerns regarding applicability were generally low across all domains. These findings suggest that while the methodological rigor was acceptable, future studies should prioritize prospective designs and standardized validation protocols to further reduce bias. The detailed risk bias of the included studies is illustrated in **TABLE 2**.

Diagnostic performance of artificial intelligence models Performance in internal validation cohorts In 15 studies reporting internal validation results, AI-based models demonstrated robust diagnostic accuracy for predicting LNM in EGC.[10-12,16,18-21,23-29] The pooled sensitivity was 0.81 (95% CI, 0.62–0.92), and the specificity was 0.82 (95% CI, 0.66–0.91; **FIGURE 2**). The corresponding summary area under the SROC curve was 0.88 (95% CI, 0.85–0.91; **FIGURE 3**), indicating strong discriminative capacity. Applying these values to a pretest probability of 20% increased the post-test probability to 53% following a positive result, and decreased it to 6% after a negative result (**FIGURE 4**). These results indicate that AI models can substantially refine clinical risk stratification beyond baseline estimates. Nevertheless, significant heterogeneity was observed (I^2 = 95.23% for sensitivity; I^2 = 99.35% for specificity), reflecting

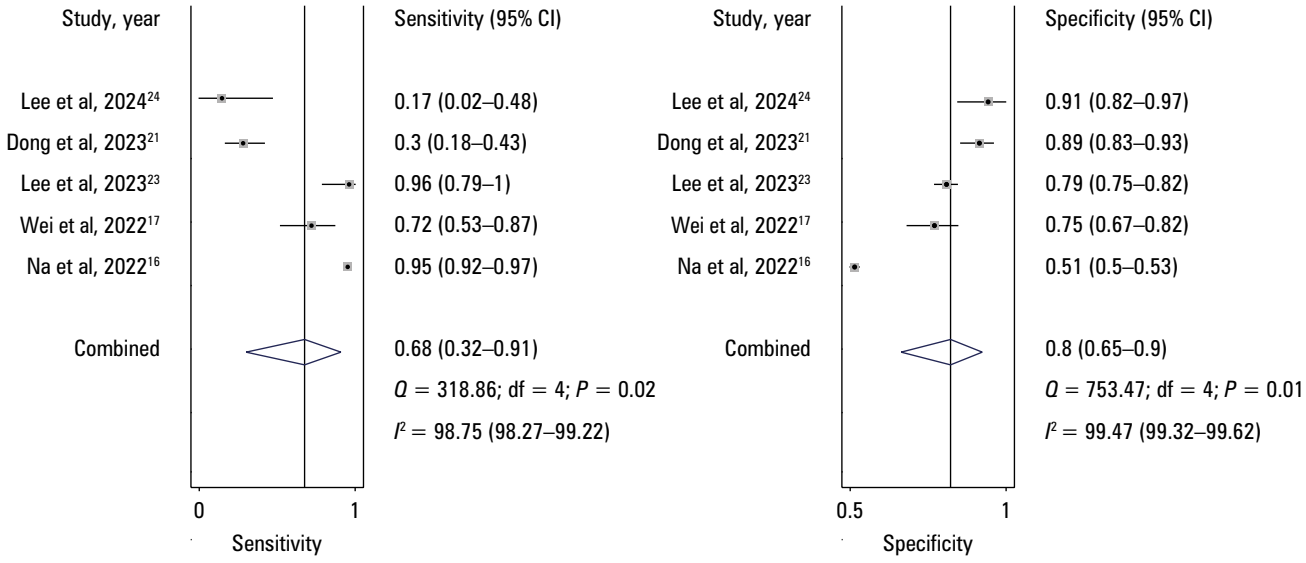

FIGURE 8 Forest plots comparing the diagnostic performance of experienced clinicians with artificial intelligence models in detecting lymph node metastasis. Artificial intelligence consistently showed higher pooled sensitivity than clinician assessment.

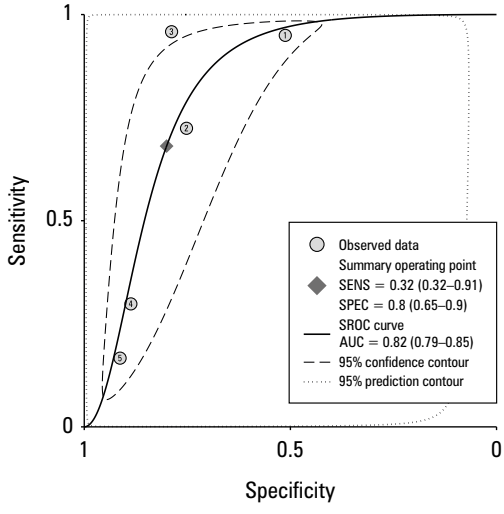

FIGURE 9 SROC curve comparing diagnostic accuracy of clinicians vs artificial intelligence–based models. Artificial intelligence models achieved superior discrimination, with a higher AUC than human assessments.

Abbreviations: see **FIGURE 3**

variability in algorithm type, input features, and reference standards.

Performance in external validation cohorts External validation, available in 10 studies, further supported the generalizability and robustness of AI-based approaches.[15-17,10,20,23-25,28,29] Pooled sensitivity and specificity were 0.81 (95% CI, 0.52–0.94) and 0.84 (95% CI, 0.67–0.94), respectively (**FIGURE 5**). The summary AUC was 0.9 (95% CI, 0.87–0.92; **FIGURE 6**), slightly outperforming the internal validation results and confirming consistent diagnostic power across independent patient populations. The positive likelihood ratio and negative likelihood ratio were 5 and 0.23, respectively, translating into a post-test probability of 56% after a positive test, and 6% after a negative test (**FIGURE 7**). These metrics underscore the clinical applicability of AI algorithms in diverse clinical settings and patient cohorts.

Comparison with clinician performance Five studies directly compared the diagnostic performance of AI models with that of experienced clinicians, including endoscopists, radiologists, and pathologists.[17,18,22,24,25] AI algorithms consistently outperformed human experts, particularly in sensitivity. The pooled sensitivity and specificity for clinician assessments were 0.68 (95% CI, 0.32–0.91) and 0.8 (95% CI, 0.65–0.9), respectively (**FIGURE 8**), with the AUC of 0.82 (95% CI, 0.79–0.85; **FIGURE 9**). Similarly, the post-test probability after a positive clinician assessment increased to only 46%, as compared with 53%–56% for AI predictions (**FIGURE 10**).

Subgroup analyses and meta-regression in internal validation cohorts To explore potential sources of heterogeneity, subgroup analyses were conducted based on the type of study (single-center or multicenter), study design, and variable type (**TABLE 3**). Stratification by type of study suggested that AI models developed in single-center studies achieved slightly higher sensitivity (0.9; 95% CI, 0.75–0.96) than those from multicenter studies (0.69; 95% CI, 0.42–0.87; P = 0.33). Similarly, case-control studies demonstrated higher sensitivity (0.99; 95% CI, 0.38–1) than retrospective cohorts (0.72; 95% CI, 0.55–0.84; P = 0.12). Stratification by input variables indicated that models incorporating clinical and pathological variables showed slightly higher sensitivity (0.84; 95% CI, 0.54–0.96) than those using endoscopic and imaging features (0.78; 95% CI, 0.66–0.87). These results suggest that none of these factors may contribute to heterogeneity, although none of the subgroup differences were significant. Subgroup analyses by type of study, study design, and variable type showed no significant differences. Notably, within most subgroups, residual heterogeneity remained significant (P <0.05), indicating that the examined factors did not fully account for variability across the studies. Furthermore, meta-regression analyses indicated that none of the tested covariates explained the substantial

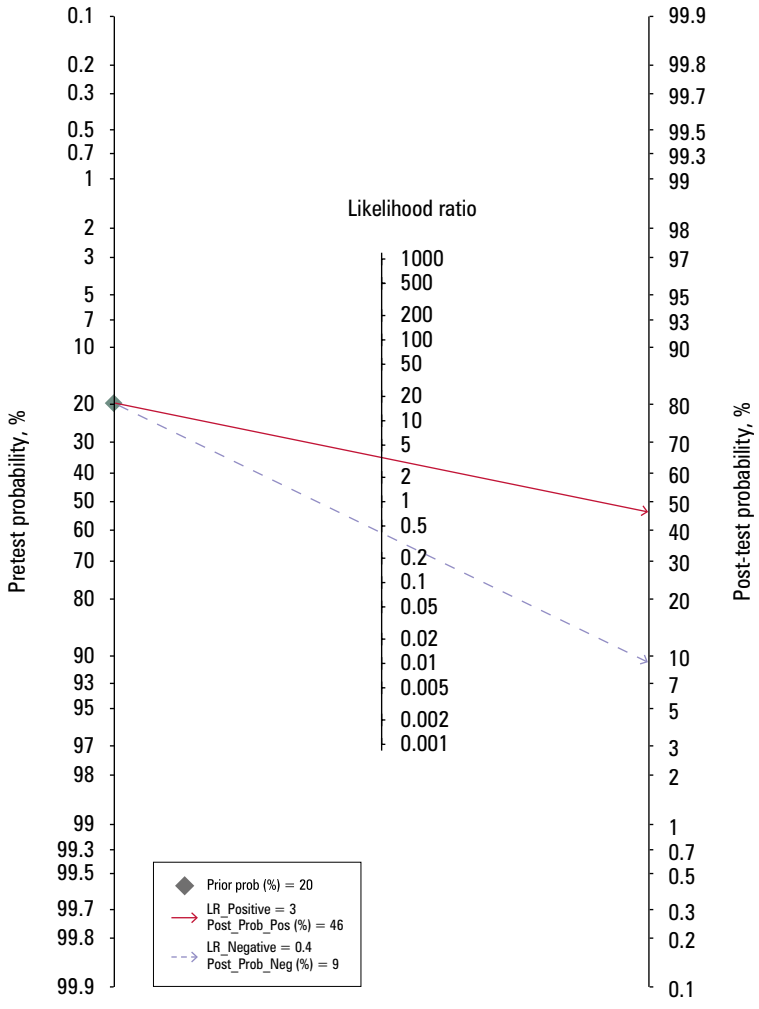

FIGURE 10 Fagan nomogram depicting clinical post-test probability shifts based on clinician diagnostic performance. As compared with artificial intelligence, clinician assessments resulted in a lower probability increase following a positive test result.

Abbreviations: see **FIGURE 4**

heterogeneity. Specifically, AI model type (P = 0.54), validation strategy (P = 0.32), type of input features (P = 0.27), and study design (P = 0.25) were not significant moderators.

Publication bias Potential publication bias was evaluated using the Deeks funnel plot asymmetry test for the included diagnostic accuracy studies. The test indicated no evidence of publication bias for the internal validation analysis (P = 0.38; **FIGURE 11**), external validation analysis (P = 0.35), or clinician comparison studies (P = 0.48; **FIGURES 12** and **13**). The symmetrical distribution of effect sizes further supports the reliability of the pooled estimates.

DISCUSSION **Principal findings** This systematic review and meta-analysis comprehensively evaluated the diagnostic performance of AI-based models for predicting LNM in EGC. By synthesizing evidence from 18 studies involving 41 505 patients, we demonstrated that AI algorithms achieved high diagnostic accuracy for LNM detection. In internal validation cohorts, the pooled sensitivity and specificity were 0.81 and 0.82, respectively, with the AUC of 0.88. Importantly, similar performance was maintained in external validation datasets (sensitivity, 0.81; specificity, 0.84; AUC, 0.9), suggesting promising generalizability. Moreover, AI models consistently outperformed experienced clinicians, particularly in sensitivity, highlighting their potential to augment or even surpass human diagnostic capabilities. Overall, these findings suggest that AI-based prediction tools could play a pivotal role in preoperative risk stratification and clinical decision-making for EGC.

Comparison with previous studies Our findings align with and extend previous literature demonstrating the potential of AI to improve oncological diagnostic accuracy. Conventional imaging modalities, such as CT, MRI, and EUS, have historically shown limited sensitivity for detecting micrometastatic disease, often resulting in FNs and suboptimal treatment decisions. For example, previous meta-analyses reported sensitivities of approximately 60% for CT and EUS in detecting NM in EGC—significantly lower than the pooled estimates observed in our analysis for AI models.[4,30] This suggests that, by integrating multidimensional data—including imaging, clinical, and pathological features—AI can extract latent patterns beyond the capabilities of conventional diagnostic approaches.

Moreover, our results corroborate earlier evidence from individual studies indicating that AI can outperform clinicians in nodal assessment. As compared with experts, the superior sensitivity of AI models observed in this meta-analysis reflects their ability to consistently detect subtle imaging and histopathological cues that may be overlooked due to human cognitive limitations or interobserver variability. Notably, several DL-based models achieved AUCs exceeding 0.9 in the external validation, underscoring their potential utility as reliable tools in routine clinical practice.

Despite the encouraging diagnostic performance, substantial heterogeneity was observed across the included studies, which likely arises from several methodological differences, including variations in study design, data sources, sample size, and feature selection. Our subgroup and meta-regression analyses, however, did not identify any single covariate—such as algorithm type, validation strategy, or input modality—as a significant source of heterogeneity. This suggests that multiple interacting factors contribute to the observed variability, underscoring the complexity of AI model development and evaluation in this context. One noteworthy finding is that models incorporating clinical and pathological variables tended to achieve slightly higher sensitivity than those relying primarily on imaging data. This observation reflects the multifactorial nature of NM, which is influenced not only by tumor morphology but also by biological aggressiveness and host–tumor

TABLE 3 Subgroup analyses results based on internal validation data

Variable		Sensitivity			Specificity		
		Summarized results	95% CI	I^2	Summarized results	95% CI	I^2
Overall pooled effect (n = 15)		0.18	0.62–0.92	95.23	0.82	0.66–0.91	99.35
Type of study	Single-center (n = 6)	0.9	0.75–0.96	95.12	0.71	0.39–0.9	99.25
	Multicenter (n = 9)	0.69	0.42–0.87	94.82	0.87	0.71–0.95	99.42
Study design	Retrospective cohort (n = 11)	0.72	0.55–0.84	94.22	0.87	0.76–0.93	97.72
	Case control study (n = 4)	0.99	0.38–1	97.29	0.62	0.15–0.94	99.74
Variables used in artificial intelligence models	Clinical and pathological variables (n = 10)	0.84	0.54–0.96	95.86	0.79	0.54–0.93	99.51
	Endoscopic and imaging variables (n = 5)	0.78	0.66–0.87	84.43	0.85	0.75–0.92	94.69

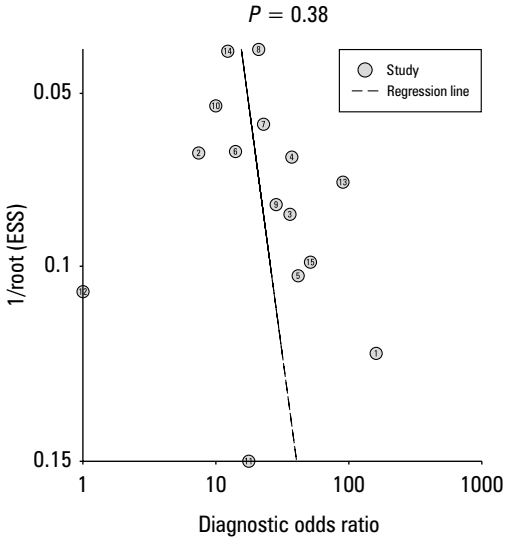

FIGURE 11 Deeks funnel plot asymmetry test assessing publication bias among internal validation studies. The slope of the regression line indicates no small-study effects.

Abbreviations: ESS, effective sample size

interactions—features that are often captured in clinicopathological and genomic data. Therefore, future AI approaches may benefit from multimodal data integration, combining radiological, pathological, and molecular features to enhance predictive performance and clinical utility. Additionally, most included studies employed retrospective, single-center designs, and only 11 studies performed external validation. This limited diversity in patient populations and clinical settings may lead to spectrum bias and restrict generalizability of the findings. While the pooled performance in the external validation cohorts was similar to that in the internal datasets, larger multicenter prospective studies are needed to confirm these results across diverse clinical scenarios.

Implications for minimally-invasive and endoscopic surgery Accurate assessment of LNM risk is a cornerstone of treatment selection in EGC, particularly in the era of minimally-invasive and endoscopic therapies. Patients with a low risk of nodal involvement may safely undergo endoscopic resection, thereby avoiding unnecessary gastrectomy and preserving gastric function, whereas those with a high risk of LNM require surgical resection with appropriate lymphadenectomy to ensure oncological safety. The present meta-analysis demonstrates that AI-based models achieve robust diagnostic accuracy and consistently outperform clinician assessment, highlighting their potential value as preoperative decision-supporting tools in this context.

From a minimally-invasive surgery perspective, AI-based LNM prediction may function as an effective triage instrument to stratify patients before treatment selection. By providing individualized probability estimates of NM, AI models may help identify optimal candidates for endoscopic resection and reduce the risk of noncurative procedures. This is particularly relevant given the clinical challenges of salvage surgery following noncurative endoscopic resection, which is associated with increased operative complexity, higher morbidity, and increased patient burden. Improved preoperative risk stratification using AI could therefore contribute to reducing unnecessary secondary surgery and optimizing initial treatment strategies.

Furthermore, AI-assisted prediction may support surgical planning beyond the binary choice between endoscopic and surgical treatment. For patients proceeding to minimally-invasive gastrectomy, preoperative estimation of LNM risk may inform the extent of LN dissection and facilitate more tailored operative strategies. In this regard, AI models align closely with the objectives of minimally-invasive surgery—minimizing surgical trauma while maintaining adequate oncological clearance.

Importantly, AI-based tools should be viewed as complementary to, rather than replacements for, clinical judgment. Integration of AI predictions into multidisciplinary decision-making processes may enhance diagnostic confidence, reduce interobserver variability, and standardize treatment selection, particularly in settings with variable expertise. As minimally-invasive and endoscopic techniques continue to evolve, AI-driven risk stratification holds promise for refining patient selection and improving the overall quality of surgical care in EGC.

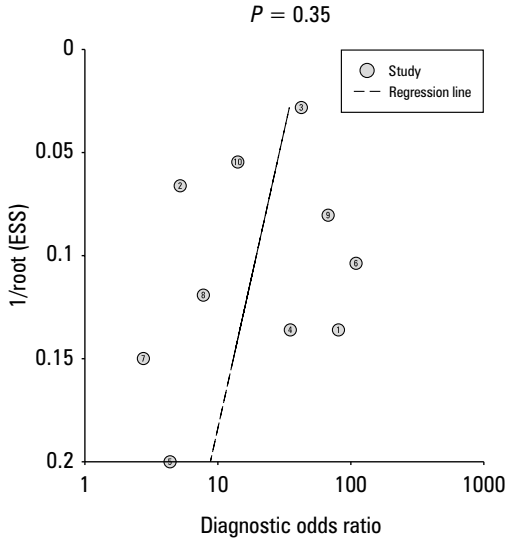

FIGURE 12 Deeks funnel plot assessing publication bias in external validation studies. No asymmetry was observed, indicating minimal publication bias.

Abbreviations: see **FIGURE 11**

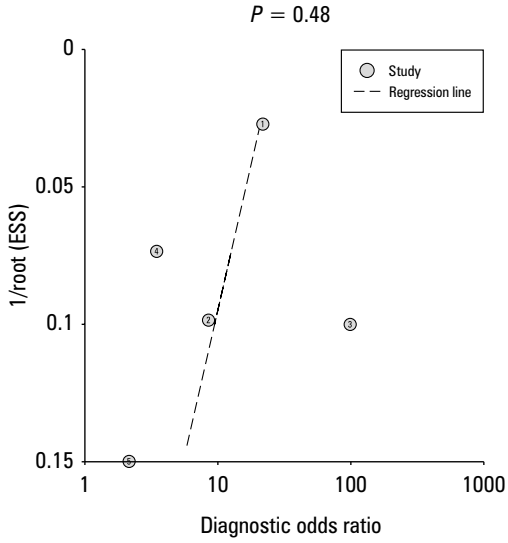

FIGURE 13 Deeks funnel plot evaluating publication bias in clinician comparison studies. The regression line suggests no small-study effects.

Abbreviations: see **FIGURE 11**

Limitations Several limitations of this meta-analysis should be acknowledged. First, the substantial heterogeneity across the studies limits the certainty of pooled estimates, despite the efforts to explore potential moderators. Second, the predominance of retrospective designs introduces risks of selection bias and confounding, while nonconsecutive patient enrollment in many studies may further compromise external validity. Third, the lack of standardized reference standards for nodal status—particularly regarding micrometastasis detection—may have contributed to variability in model performance and outcome definitions. Fourth, most models were developed and validated in East Asian populations, raising questions about their applicability to Western cohorts with differing disease epidemiology, histopathological features, and treatment strategies. Fifth, while AI models demonstrated high accuracy, few studies reported calibration metrics or clinical utility analyses (eg, decision-curve analysis), which are critical for translating diagnostic performance into meaningful clinical outcomes. Finally, the absence of an open-source model code and publicly available datasets in most studies limits reproducibility and hinders external validation by independent researchers.

Future directions To fully harness the potential of AI in EGC management, several research priorities must be addressed. Prospective, multicenter studies with standardized patient inclusion criteria and outcome definitions are essential to validate model performance and enhance generalizability. Future efforts should also focus on integrating multimodal data—including radiomics, histopathomics, genomics, and proteomics—to capture the full biological complexity of LNM. Advances in federated learning and privacy-preserving AI architectures could facilitate collaborative model development across institutions without compromising patient data security.

Moreover, clinical implementation studies are needed to assess the real-world impact of AI-based LNM prediction tools on patient outcomes, health care utilization, and cost-effectiveness. Such studies should incorporate calibration analyses, net benefit evaluations, and decision-analytic modeling to quantify the added value of AI in clinical workflows. Finally, transparent reporting following Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or Diagnosis for AI and Prediction model Risk of Bias Assessment Tool for AI guidelines,[31,32] along with public release of code and anonymized datasets, will be crucial for advancing the field and fostering reproducibility.

CONCLUSIONS This systematic review and meta-analysis demonstrated that AI-based models achieve high diagnostic accuracy for predicting LNM in EGC and consistently outperform clinician assessment. Beyond diagnostic performance, these models show considerable potential as decision-supporting tools for minimally-invasive and endoscopic management by assisting clinicians in selecting appropriate treatment strategies and tailoring surgical extent. Prospective multicenter validation and implementation studies are warranted to facilitate the integration of AI-assisted prediction into minimally-invasive GC care.

ARTICLE INFORMATION

ACKNOWLEDGMENTS None.

FUNDING The work was supported by the Science and Technology Planning Project of Shangrao City (2023CZDX126).

CONTRIBUTION STATEMENT CXP, YZG, and YJP conceptualized, designed, and revised the manuscript; CXP and YZG searched the literature, collected the data, organized the data, and drafted the manuscript; CXP and YZG collected the data; CXP and YJP performed the statistical analyses. All authors read and approved the submitted version of the manuscript.

CONFLICT OF INTEREST None declared.

AI STATEMENT Artificial intelligence was not used in the preparation of this manuscript.

HOW TO CITE Chen XP, Yi ZG, Ye JP. Artificial intelligence models in predicting lymph node metastasis in early gastric cancer: a systematic review and meta-analysis. Wideochir Inne Tech Maloinwazyjne. 2026; 21: 1-12. doi:10.20452/wiitm.2026.18006

JOURNAL INFORMATION

Videosurgery and Other Miniinvasive Techniques is an official journal of the Videosurgery Foundation.

REFERENCES

1 Siegel RL, Kratzer TB, Giaquinto AN, et al. Cancer statistics, 2025. CA Cancer J Clin. 2025; 75: 10-45.

2 Elgenidy A, Alomari O, Hesn MM, et al. Relative survival, conditional survival, and causes of death in patients with early gastric cancer, with a focus on differences between cardia and non-cardia cancer. Cancers (Basel). 2024; 16: 4262.

3 Sun F, Huang Y, Sun Y, et al. Risk factors of additional surgery after non-curative endoscopic submucosal dissection for early gastric cancer. BMC Gastroenterol. 2023; 23: 383.

4 Liu S, Zhang M, Yang Y, et al. Establishment and validation of a risk score model based on EUS: assessment of lymph node metastasis in early gastric cancer. Gastrointest Endosc. 2024; 100: 857-866.

5 Lucey BC, Stuhlfaut JW, Soto JA. Mesenteric lymph nodes seen at imaging: causes and significance. Radiographics. 2005; 25: 351-365.

6 Wang Z, Liu J, Luo Y, et al. Establishment and verification of a nomogram for predicting the risk of lymph node metastasis in early gastric cancer. Rev Esp Enferm Dig. 2021; 113: 411-417.

7 Wang J, Zeng Z, Li Z, et al. The clinical application of artificial intelligence in cancer precision treatment. J Transl Med. 2025; 23: 120.

8 Gao P, Xiao Q, Tan H, et al. Interpretable multi-modal artificial intelligence model for predicting gastric cancer response to neoadjuvant chemotherapy. Cell Rep Med. 2024; 5: 101848.

9 Lei C, Sun W, Wang K, et al. Artificial intelligence-assisted diagnosis of early gastric cancer: present practice and future prospects. Ann Med. 2025; 57: 2461679.

10 Yang R, Zhang J, Zhan F, et al. Artificial intelligence efficiently predicts gastric lesions, Helicobacter pylori infection and lymph node metastasis upon endoscopic images. Chin J Cancer Res. 2024; 36: 489-502.

11 Kato M, Hayashi Y, Uema R, et al. A machine learning model for predicting the lymph node metastasis of early gastric cancer not meeting the endoscopic curability criteria. Gastric Cancer. 2024; 27: 1069-1077.

12 Sung YN, Lee H, Kim E, et al. Interpretable deep learning model to predict lymph node metastasis in early gastric cancer using whole slide images. Am J Cancer Res. 2024; 14: 3513-3522.

13 McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) Statement. JAMA. 2018; 319: 388-399.

14 Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011; 155: 529-536.

15 Tian H, Ning Z, Zong Z, et al. Application of machine learning algorithms to predict lymph node metastasis in early gastric cancer. Front Med (Lausanne). 2021; 8: 759013.

16 Wang X, Li C, Fang M, et al. Integrating No. 3 lymph nodes and primary tumor radiomics to predict lymph node metastasis in T1-2 gastric cancer. BMC Med Imaging. 2021; 21: 58.

17 Na JE, Lee YC, Kim TJ, et al. Machine learning model to stratify the risk of lymph node metastasis for early gastric cancer: a single-center cohort study. Cancers (Basel). 2022; 14: 1121.

18 Wei X, Yan XJ, Guo YY, et al. Machine learning-based gray-level co-occurrence matrix signature for predicting lymph node metastasis in undifferentiated-type early gastric cancer. World J Gastroenterol. 2022; 28: 5338-5350.

19 Yang T, Martinez-Useros J, Liu J, et al. A retrospective analysis based on multiple machine learning models to predict lymph node metastasis in early gastric cancer. Front Oncol. 2022; 12: 1023110.

20 Zeng Q, Li H, Zhu Y, et al. Development and validation of a predictive model combining clinical, radiomics, and deep transfer learning features for lymph node metastasis in early gastric cancer. Front Med (Lausanne). 2022; 9: 986437.

21 Zhu H, Wang G, Zheng J, et al. Preoperative prediction for lymph node metastasis in early gastric cancer by interpretable machine learning models: a multicenter study. Surgery. 2022; 171: 1543-1551.

22 Dong ZB, Xiang HT, Wu HM, et al. LncRNA expression signature identified using genome-wide transcriptomic profiling to predict lymph node metastasis in patients with stage T1 and T2 gastric cancer. Gastric Cancer. 2023; 26: 947-957.

23 Hayashi T, Takasawa K, Yoshikawa T, et al. A discrimination model by machine learning to avoid gastrectomy for early gastric cancer. Ann Gastroenterol Surg. 2023; 7: 913-921.

24 Lee HD, Nam KH, Shin CM, et al. Development and validation of models to predict lymph node metastasis in early gastric cancer using logistic regression and gradient boosting machine methods. Cancer Res Treat. 2023; 55: 1240-1249.

25 Lee S, Jeon J, Park J, et al. An artificial intelligence system for comprehensive pathologic outcome prediction in early gastric cancer through endoscopic image analysis (with video). Gastric Cancer. 2024; 27: 1088-1099.

26 Seo JW, Park KB, Lim ST, et al. Machine learning models for prediction of lymph node metastasis in patients with T1b gastric cancer. Am J Cancer Res. 2024; 14: 3842-3851.

27 Gao L, Liu W, Kang B, et al. AutoLNMNet: automated network for estimating lymph-node metastasis in EGC using a pyramid vision transformer and data derived from multiphoton microscopy. Microsc Res Tech. 2025; 88: 315-322.

28 He J, Xu J, Chen W, et al. Development of a deep learning model for T1N0 gastric cancer diagnosis using 2.5D radiomic data in preoperative CT images. NPJ Precis Oncol. 2025; 9: 249: 869.

29 Kang D, Jeon HJ, Kim JH, et al. Enhancing lymph node metastasis risk prediction in early gastric cancer through the integration of endoscopic images and real-world data in a multimodal AI model. Cancers (Basel). 2025; 17.

30 Cardoso R, Coburn N, Seevaratnam R, et al. A systematic review and meta-analysis of the utility of EUS for preoperative staging for gastric cancer. Gastric Cancer. 2012; 15 (Suppl 1): S19-S26.

31 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. 2025; 388: e082505.

32 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024; 385: e078378.