*Systematic Review and Meta-analysis Article*
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

*Med J Islam Repub Iran.* 2025 (20 Aug);39.110. https://doi.org/10.47176/mjiri.39.110

# Diagnostic Accuracy of Deep Learning for Predicting Lymph Node Metastasis Based on Computed Tomography in Gastric Cancer: A Systematic Review and Meta-analysis

Armin Majd Gharamaleki[1], Arman Majd Gharamaleki[2], Alireza Amanollahi[3], Sarvin Tabibzadeh[4]*

Received: 1 Jun 2025 Published: 20 Aug 2025

## Abstract

**Background:** Early detection of lymphatic metastasis (LNM) in gastric cancer (GC) is essential to determine the treatment strategy. Conventional methods exhibit limited efficacy, highlighting the need for more reliable approaches. Deep learning (DL) models show promise for LNM detection in computed tomography (CT); their performance requires comprehensive evaluation. This systematic review and meta-analysis evaluate the diagnostic performance of CT-based DL models for detecting LNM in GC patients.

**Methods:** A systematic review and meta-analysis was conducted according to PRISMA-DTA guidelines. PubMed, Embase, and Web of Science were searched up to May 5, 2025. The focus was on studies that used DL models to detect LNM in CT in GC. Using a bivariate random effect model, Pooled estimates were calculated, heterogeneity and publication bias were assessed, and clinical utility was evaluated via Fagan plots and likelihood ratio matrices. Validation type, input data types, CT phases, segmentation techniques, and DL architectures stratified subgroup analyses. The quality was assessed with QUADAS-2.

**Results:** From the 14 included studies, 11 studies with 5296 patients were analyzed. In internal validation, DL feature-based models achieved a pooled area under the curve (AUC) of 0.91 (95% CI: 0.88-0.93), sensitivity of 0.86 (95% CI: 0.75-0.92), and specificity of 0.83 (95% CI: 0.67-0.92). Performance degraded in external validation, with specificity dropping to 0.59 (95% CI: 0.26-0.85). Models that integrated DL features with radiomics features showed similar overall performance but were noted to have a higher confirmatory power. In terms of clinical utility, although the models could significantly alter post-test probabilities, they ultimately lacked the certainty required to serve as standalone diagnostic tools.

**Conclusion:** CT-based DL models show high diagnostic accuracy but limited generalizability across external datasets, indicating overfitting. A key finding of this meta-analysis is that pervasive and asymmetric heterogeneity, particularly in specificity, suggests that technical standardization alone is insufficient. Integrating clinical variables reduces heterogeneity; however, prospective, multicenter studies are needed to further enhance reproducibility.

**Keywords:** Stomach Neoplasms, Deep Learning, Lymphatic Metastasis, Tomography, X-Ray Computed, Artificial Intelligence, Convolutional Neural Networks, Systematic Review, Meta-Analysis

***Conflicts of Interest:*** *None declared*
***Funding:*** *None*



***Cite this article as***: Majd Gharamaleki A, Majd Gharamaleki A, Amanollahi A, Tabibzadeh S. Diagnostic Accuracy of Deep Learning for Predicting Lymph Node Metastasis Based on Computed Tomography in Gastric Cancer: A Systematic Review and Meta-analysis. *Med J Islam Repub Iran.* 2025 (20 Aug);39:110. https://doi.org/10.47176/mjiri.39.110

***Corresponding author:*** *Dr Sarvin Tabibzadeh, sarvin.tabibzadeh@iau.ir*

1. Student Research Committee, School of Medicine, Ardabil University of Medical Sciences, Ardabil, Iran
2. Student Research Committee, Tabriz University of Medical Sciences, Tabriz, Iran
3. Trauma and Injury Research Center, Iran University of Medical Sciences, Tehran, Iran
4. School of Medicine, Islamic Azad University, Ardabil, Iran

***↑What is "already known" in this topic:***
Lymphatic metastasis is a pivotal determinant of the prognosis and treatment approach in gastric cancer (GC). Current diagnostic methods for lymphatic metastasis diagnosis offer limited accuracy. Deep learning (DL), as a novel technology, presents a promising alternative to traditional methods for diagnosing medical images.

***→What this article adds:***
This systematic review and meta-analysis comprehensively evaluate the diagnostic accuracy of DL models for detecting lymphatic metastasis in computed tomography in gastric cancer patients. Our findings highlight the significant potential of DL models, particularly in internal validation settings, but also reveal critical challenges related to poor generalizability and substantial between-study heterogeneity that currently limit their clinical applicability.

## Introduction

Gastric cancer (GC) ranks as the fifth most common malignancy globally and the third leading cause of cancer-related death (1). More than 50% of patients with GC present with lymph node metastases (LNM) at initial diagnosis or surgical resection, which decreases the 5-year survival rate to below 30% (2). Hochwald et al analyzed data from 5-year survivors of GC and concluded that LNM was the strongest prognostic factor for postoperative outcome. They further showed that the number of positive lymph nodes was the most important predictor of survival probability (3). The detection of LNM in GC is essential in determining the surgical approach for patients and the administration of chemotherapy (4). The National Comprehensive Cancer Network recommends the use of computed tomography (CT) to detect LNM, which has a limited diagnostic accuracy (50%-70%) (5). Human interpretation of medical images has several limitations, including subjectivity, interobserver variability, and fatigue. With the increasing volume of medical images and the time constraints of radiologists, the probability of missing findings, prolonged response times, and lack of quantitative analysis increases. These factors are serious obstacles to the development of personalized and evidence-based healthcare (6). Currently, computer-aided diagnosis (CAD) systems have been developed to enhance the performance of conventional imaging modalities and reduce the time required for image interpretation. Only machine learning (ML) based radiomics has been suggested to enhance the prediction of LNM in GC. Deep learning (DL), a novel technique, utilizes convolutional neural networks (CNNs) to learn internal patterns and profoundly represent medical imaging data. Compared with ML, DL can extract high-level contextual patterns and meticulous (7). In recent years, DL methods, particularly CNNs, have been widely used in medical image analysis. These models demonstrate high performance in image analysis because of their ability to recognize spatial patterns and learn hierarchically from image features. Other DL models, such as recurrent neural networks (RNNs) for sequential data and generative adversarial networks (GANs) for generating new data based on learned data distribution, have also been applied. For evaluating the performance of DL models in medical image recognition, metrics such as the receiver operating characteristic (ROC) curve, area under the curve (AUC), and confusion matrix are employed. The ROC curve indicates the balance between sensitivity and specificity of the model, and AUC, as a numerical index, summarizes the overall performance of the model (8). Despite the increasing utilization of DL in the detection of LNM from radiological images, the existing evidence is sparse and methodologically heterogeneous. Our search strategy revealed a lack of comprehensive reviews that rigorously evaluate the diagnostic accuracy of DL algorithms in detecting LNM from CT images in GC. Our aim in this systematic review and meta-analysis was to investigate the diagnostic accuracy of DL algorithms applied to CT images for predicting LNM in patients with GC. The goal of this study was to fill the gap in the current literature by conducting an evaluative assessment of the available evidence and elucidating the clinical applicability and methodological quality of these models.

## Methods

### *Data Sources*

This review was reported according to the PRISMA-2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses-2020) guidelines and its extension for PRISMA-DTA (Diagnostic Test Accuracy Studies) (9, 10). Up to May 5, 2025, a comprehensive search was conducted across 3 major databases: PubMed, Embase, and Web of Science. For each database, specific search strategies were performed, which included the following terms: ("Stomach Neoplasm*" OR "Neoplasm, Stomach" OR "Gastric Neoplasm*" OR "Neoplasm, Gastric" OR "Neoplasms, Gastric" OR "Neoplasms, Stomach" OR" Cancer of Stomach" OR "Stomach Cancer*"OR "Cancer of the Stomach" OR "Gastric Cancer*" OR "Cancer, Gastric" OR "Cancers, Gastric" OR "Cancers, Stomach" OR "Cancer, Stomach" OR "Gastric Cancer, Familial Diffuse") AND ("Lymphatic Metastasis" OR "Lymphatic Metastases" OR "Lymph Node Metastasis" OR "Lymph Node Metastases" OR "Metastasis, Lymph Node") AND ("Deep Learning" OR "Learning, Deep" OR "Hierarchical Learning" OR "Learning, Hierarchical" OR "Neural Networks, Computer" OR "Machine Learning" OR "Artificial Intelligence" OR "Convolutional Neural Networks" OR " CNN" OR "Transformer Models" OR "Vision Transformer"). The search aimed to ascertain the diagnostic accuracy and reliability of DL models in detecting LNM utilizing CT images in patients diagnosed with GC. The search strategy is detailed in the Appendix.

### *Study Selection*

The articles were reviewed independently by 2 authors (A.M. and A.M.). The titles and abstracts of the articles were evaluated, and the articles were included according to the eligibility criteria. The full text of the screened articles was reviewed. All full-length papers were reviewed by 2 authors (A.M. and A.M.), and discrepancies were resolved by the intermediacy of a corresponding author (ST). Reasons for exclusion were meticulously documented.

### *Including Criteria*

Articles employing DL models, such as CNNs, deep convolutional neural networks (D-CNNs), transformers, fully connected neural networks (FCNNs), and residual networks (ResNets), which focused on predicting LNM in GC patients using CT images, were included in this study. The reference standard for diagnosing LNM was histopathological assessment after surgery. No restrictions were imposed on the date of publication, country of origin, type of study design, or language of the article. No specific criteria were predefined for including hybrid models (DL combined with radiomics features, or clinical variables, such as patient demographics or tumor charac-





[ DOI: 10.47176/mjiri.39.110 ]

teristics) or traditional models (e.g., radiologist assessments or non-DL machine learning methods like logistic regression). However, studies reporting such comparator models were included post hoc if they met the primary inclusion criteria (DL-based LNM detection using CT in GC).

### Exclusion Criteria

Studies that focused on non-gastric cancers, predicted outcomes other than LNM, or utilized methods other than CT scan images as input data type were excluded. Additionally, studies that only used traditional algorithms or did not disclose performance metrics and raw data were excluded from this study. Case reports, conference abstracts, review articles, meta-analyses, retracted studies, and studies with a sample size of less than 100 people were excluded from this review. The exclusion of studies with fewer than 100 participants was implemented to reduce potential selection bias and enhance the statistical power and generalizability of the evaluation.

### Data Extraction

Data were independently extracted by 2 authors (A.M. and A.M.). Fourteen included studies were recorded in preprepared tables. The index test was defined as any DL model applied to CT images for LNM prediction. The target condition was the diagnosis of LNM in patients with GC. The reference standard was postoperative histopathological assessment of resected lymph nodes. For the studies that compared different models of DL, we extracted data regarding the primary models that were identified based on criteria defined by the original authors. For each study, we extracted the following data: first author, year of study, study object, country of origin, type of study, input data type (deep learning features [DLF], deep leaning features + hand-crafted radiomics features [HCRF], deep leaning features + clinical variables, or DLF + HCRF + clinical variables), model architecture, segmentation method, clinical variables, sample size, validation method, and performance metrics. The performance of DL models was evaluated using the area under the ROC curve (AUC; plots sensitivity on the y-axis against 1 − specificity on the x-axis at varying thresholds), as well as sensitivity (true positive rate), specificity (true negative rate), accuracy (overall correct classification), and the 95% confidence interval of the AUC. For studies with complete lymph node metastasis (LNM) distributions, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were directly calculated; for studies with incomplete data, these values were estimated where possible. For studies reporting multiple validation datasets for a single model, only the validation dataset with the largest sample size was included in the meta-analysis. Ultimately, 3 studies were excluded from the quantitative synthesis due to insufficient data, and 11 studies were included in the meta-analysis.

### Statistical Analyses

A diagnostic test accuracy (DTA) meta-analysis was conducted using a bivariate random-effects model to synthesize the performance of CT-based DL models in predicting LNM. Analyses were stratified by validation type (internal vs. external validation cohorts) and performed hierarchically for subgroups based on input data type, CT phase, segmentation method, and model architecture, provided that at least three studies were available for each subgroup.

For subgroups with 3 or more studies, the analysis was performed in Stata (Version 17) using the metadta command and in the R environment (version 4.5.1), utilizing the mada package. This model was used to calculate pooled estimates of sensitivity, specificity, each with 95% CIs, diagnostic odds ratios (DOR), positive likelihood ratios (LR+), negative likelihood ratios (LR-), AUC, and its 95% CIs. SROC curves, forest plots, and bivariate box plots of sensitivity and specificity were generated to visualize pooled results and between-study heterogeneity. Heterogeneity was assessed using the chi-square test and the $I^2$ statistic, with $I^2$ values interpreted to indicate the degree of variability across studies. Publication bias was evaluated using Egger's tests, but was not assessed for subgroups with fewer than 4 studies, since tests like Egger's are underpowered and unreliable in small samples. To evaluate further diagnostic performance and clinical utility across all groups, additional analyses and visualizations were conducted. These included the generation of Fagan's nomograms and likelihood ratio matrices. Statistical analysis was performed by 2 authors (A.A. and S.T.).

### Subgroup Analyses

The meta-analysis was conducted in a stepwise, hierarchical manner to evaluate the different factors on diagnostic performance systematically. The analyses were structured as follows:

1. Input data types: models were categorized into 3 groups based on input data (DLF, DLF + HCRF, and DLF + HCRF + clinical variables). Pooled diagnostic performance estimates were calculated for internal validation cohorts (IVC) and external validation cohorts (EVC) to assess baseline model performance across these groups.

2. CT phases by input data type: models were further stratified by CT phase within the DLF-based model and analyzed separately in the internal validation set. While studies utilized arterial, venous, portal venous, parenchymal, and unenhanced CT imaging, only arterial and combined portal venous/venous phase models had sufficient data for analysis. Due to limited data for the venous phase, the portal phase was combined with the venous phase for analysis, as both are functionally similar.

3. Segmentation techniques by input data type: the models were stratified based on their segmentation technique (manual, semi-automatic, or automatic). A quantitative meta-analysis was performed only for the manual segmentation subgroup, as there was an insufficient number of studies employing semi-automatic or automatic methods to permit statistical pooling. Consequently, the results for these groups were summarized descriptively.

4. DL model architectures by input data type: DL models were categorized by algorithm type within each input



[ DOI: 10.47176/mjiri.39.110 ]

data type and analyzed separately in internal and external validation datasets. Only CNN-based models had sufficient studies for meta-analysis, as other architectures (e.g., Graph-based models and Vision Transformers) lacked adequate data.

5. CNN models by segmentation technique, CT phase, and input data type: a combined analysis evaluated CNN-based models within the DLF group, focusing on manual segmentation and portal/venous-phase CT imaging in IVC. This step integrated key methodological factors to assess their collective impact on diagnostic performance.

### *Quality Assessment*

The quality assessment of the included articles was conducted independently by 2 authors

*(*A.M. and A.M.) based on the Quality Assessment of Diagnostic Accuracy Studies tool-2 (QUADAS-2) framework (11). Disagreements were resolved through the intervention of the corresponding author (S.T.).

## Results

### *Screening and Selection of Articles*

A systematic search was conducted according to predetermined search strategies. A total of 554 articles were identified. After eliminating duplicate articles, 415 articles were selected for review based on title and abstract. After the initial review, 369 articles were excluded, and 44 articles were selected for further review. The full text was reviewed, and 14 articles were ultimately included in the study because they aligned with our objectives and met the specified criteria. Among the 14 included studies, 11 provided sufficient data for quantitative synthesis and were included in the meta-analysis, while 3 studies were excluded from the meta-analysis due to insufficient diagnostic data. None of the retrieved studies had a sample size below 100 participants; therefore, no articles were excluded based on this criterion. A flow diagram illustrating the selection process is presented in Figure 1.

### *Study and Patient Characteristics*

Fourteen studies, each employing DL models to detect LNM using CT scan images in patients with GC, were included in this review. All included studies were retrospective. Seven studies utilized multiple hospital centers to collect patients, while the remaining studies were conducted at a single center. All studies employed histopathological diagnosis (various stages from early to locally advanced) to diagnose GC, and none of the patients studied had received chemotherapy agents, except for 1 study conducted by Zheng et al (12). The total cohort comprised 8148 patients, with individual study sample sizes ranging from 170 to 1699. A total of 5296 patients were included in the meta-analysis. The gold standard for determining LNM status was commonly postoperative pathological assessment. Studies generally divided cohorts into training, internal validation, and external validation sets, with 1 international validation cohort from Italy in the study by Dong et al (13). The studies extracted features from the tumor, lymph nodes, or a combination of these structures, except the study by Shang et al, which utilized imaging of the spleen (14). Most studies aimed to predict the incidence of LNM in patients using DL models, except the study by Liu et al, which utilized the capability of DL to detect LNM to prevent surgical overtreatment (15). Dong et al and Zhao et al also examined the stages of LNM, and

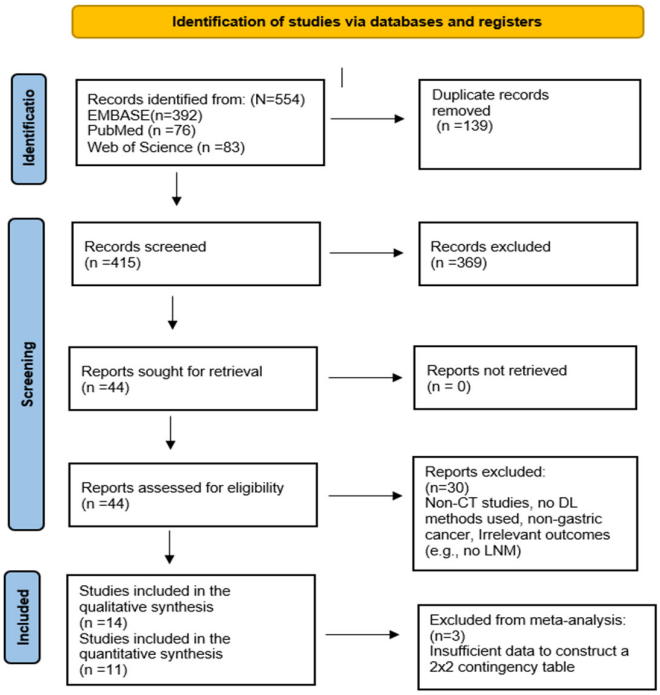


*Figure 1.* PRISMA flow chart illustrating the study selection process





[ DOI: 10.47176/mjiri.39.110 ]

*Table 1.* Characteristics of Included Studies

| Study | Country | Study Type | Reference Standard | Prediction target | Feature source | Segmentation method | Clinical variables integration | CT imaging phase |
|---|---|---|---|---|---|---|---|---|
| Liu (15) | China | Retrospective diagnostic accuracy study | Surgical histopathology | Discriminating D1 (pT1 + pN0) vs. D2 (≥ pT1 + ≥ pN1) lymphadenectomy candidates. | Tumor & lymph nodes | Manual by 2 radiologists | Age, sex, tumor location, T-stage | Arterial, parenchymal |
| Jin (18) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in 11 lymph nodes stations | Primary tumor & surrounding areas | Manual by 2 radiologists | tumor location, grade of differentiation, Laure´n histological type | Portal venous |
| Zeng (20) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in Early gastric cancer | Tumor | Manual by 2 radiologists | age, gender, tumor size, depth, grade, Lauren type, ulcer, LVI | Portal venous |
| Shang (14) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in GC | Spleen | Manual by 2 radiologists | Age, sex, clinical symptoms | Venous phase |
| Zhang (4) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in LAGC | Tumor | Manual by 2 radiologists | Tumor diameter, clinical T stage, and CT-reported LN | Venous phase |
| Zhang (7) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM with multicenter data & MSDA | Tumor & lymph nodes | Automatic with 3D IFPN & FDT module | Age & sex as auxiliary tasks | Unenhanced, enhanced |
| Zhu (19) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in 12 lymph nodes stations & overall metastasis | Tumor | Automatic with 3D Attention-UNet | No | Arterial phase |
| Gao (17) | China | Retrospective diagnostic accuracy study | Surgical histopathology | Peri-gastric metastatic lymph nodes | Metastatic lymph nodes | Manual by 3 radiologists | No | Arterial, venous, equilibrium |
| Dong (13) | China, Italy | Retrospective diagnostic accuracy study | Surgical histopathology | Pathologic N stage, discriminating non-N0 vs N0 LNM | Tumor | Manual by 1 radiologist | Clinical N stage | Unenhanced and arterial, venous (PC, VC1, VC2) arterial, venous (VC3, IVC) |
| Guan (21) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in GC | Tumor | Semiautomatic with CT thresholding, manual adjustment by 2 radiologists | CT-reported LN status | Arterial phase |
| Zheng (12) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in LAGC post-NAC | Tumor | Semiautomatic with AILEN, manual adjustment by 2 radiologists | Tumor location and cN stage | Portal venous |
| Zhao (16) | China | Retrospective diagnostic accuracy study | Surgical histopathology | N0, N1, N2, N3a, N3b Stages Binary LNM status | Tumor & WSI pathology | Manual by 2 radiologists | No | Portal venous |
| Li (22) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in GC | Tumor | Manual by radiologists | Tumor thickness, nICVP, and CT- CT-reported LN | Arterial, venous (dual energy CT) |
| Wan (23) | China | Retrospective diagnostic accuracy study | Surgical histopathology | LNM in non-enlarged lymph nodes | Lymph nodes | Manual by 2 radiologists | NO | Arterial phase |

GC: Gastric cancer CT: Computed tomography LN: Lymph node LNM: Lymphatic metastasis LVI: Lymph vascular invasion LAGS: Locally advanced gastric cancer IFPN: Improved Feature Pyramid Network PC: Primary cohort VC: Validation cohort IVC: International validation cohort NAC: Neoadjuvant chemotherapy VIT: Vision transformer DCNN: Deep convolutional neural network SAE: Sparse autoencoder WSI: Whole slide image

the study by Gao et al focused on perigastric lymph node metastasis (PGMLNs**)** (13, 16, 17). Two studies evaluated LNM at various lymphatic stations (18, 19). Some included studies reported comparator models, alongside DL models, including hybrid models integrating DL with radiomics or clinical variables. These comparator models were quantitatively analyzed based on input data types (e.g., DL features, radiomics, clinical variables) in the meta-analysis. Tables 1 and 2 provide a comprehensive summary of the characteristics of the included studies and their reported performance metrics.



[ DOI: 10.47176/mjiri.39.110 ]

*Table 2.* Summary of the DL Models' Performance Metrics for LNM Prediction in GC Using Preoperative CT Imaging. Combined models (integrating DL with radiomics/clinical variables) are included when DL was used for feature extraction

| Authors | Sample size | Deep learning model | Input data type | validation sets | Discrimination statistics of the main model | TP | TN | FP | FN |
|---|---|---|---|---|---|---|---|---|---|
| Dong (13) 2020 | 730 (PC: 225, VC1: 178, VC2: 145, VC3: 131, IVC: 51) | DenseNet-201 | DLF+HCRF+ Clinical variables730 | EVC1 | C-index: 0.777 (95%CI: 0.735-0.819) | 114 | 21 | 36 | 9 |
| | | | | EVC2 | C-index: 0.817 (95%CI: 0.775-0.860) | 111 | 12 | 16 | 6 |
| | | | | EVC3 | C-index: 0.787 95%CI:0.756-0.887 | 60 | 26 | 34 | 11 |
| | | | | International | C-index: 0.822 (95%CI: 0.737-0.838) | - | - | - | - |
| Zheng (12) 2024 | 1205 (TC: 361, IVC: 155, EVC1:319, EVC2:370) | Transformer-based DLN | DLF | EVC2 | AUC: 0.788 (95% CI: 0.735–0.835) Sen:0.785 Spe: 0.597 ACC: 0.715 | 157 | 71 | 48 | 43 |
| | | Transformer-based DLN | DLF + Clinical | EVC2 | AUC: 0.77 Sen:0.769 Spe: 0.618 ACC: 0.714 (95% CI: 0.713–0.818) | 180 | 84 | 52 | 54 |
| Guan (21) 2023 | 347 (TC: 242, IVC: 105) | ResNet50 | DLF | IVC | AUC: 0.9803 Sen:0.9839 Spe: 0.9767 ACC: 0.981 | 61 | 42 | 1 | 1 |
| | | | DLF+ HCRF | IVC | AUC: 0.9687 Sen:0.9839 Spe: 0.9535 ACC: 0.9714 | 61 | 41 | 2 | 1 |
| Zhang (4) 2022 | 523 (TC: 367, IVC: 156) | ResNet50 | DLF | IVC | AUC: 0.796 (95% CI: 0.715-0.865) Sen: 0.802 spe: 0.647 ACC: 0.752 | 87 | 30 | 17 | 22 |
| Li (22) 2020 | 204 (TC: 136, IVC: 68) | DCNN | DLF + HCRF + Clinical variables | IVC | AUC: 0.82 (95% CI: 0.72-0.92) Sen: 0.74 spe: 0.8 ACC: 0.76 | 22 | 30 | 8 | 8 |

DLF: Deep learning feature HCRF: Handcrafted radiomics feature DCNN: Deep Convolutional neural network PC: Primary cohort VC: Validation cohort IVC: International validation cohort TC: Training cohort EVC: External validation AUC: Area under the curve Sen: Sensitivity Spe: Specificity ACC: Accuracy CI: Confidence interval TP: True positive TN: True negative FP: False positive FN: False negative FRCNN: Faster region-based convolutional neural networks 3D IFPN: 3D improved feature pyramidal network UDC-GCN: Unsupervised domain selective graph convolutional network LMM-net: Lymph Node metastasis multitask learning network FRCNN: Faster region-based convolutional neural networks VGG19: Visual Geometry Group 19-layer network

### *Meta-analysis of Diagnostic Performance Based on Input Data Types*

The meta-analysis was conducted in a stepwise, hierarchical manner. For each subgroup, a bivariate random-effects model was used. The baseline pooled diagnostic performance estimates for DL models based on input data types across internal and external validation datasets. Forest plots in Figure 2 provide a visual summary of the pooled diagnostic performance. They also display individual study estimates with corresponding confidence intervals, highlighting both the consistency and variability of findings across studies.

### *DLF-Based Models*

In the IVC, DLF-based models demonstrated excellent overall diagnostic accuracy with a pooled AUC of 0.91 (95% CI: 0.88-0.93). The pooled sensitivity was 0.86 (95% CI: 0.75-0.92), and the pooled specificity was 0.83 (95% CI: 0.67-0.92). The models demonstrated strong discriminatory ability, with a DOR of 4.43 (95% CI: 2.63-6.86), a LR+ of 4.23, and an LR- of 0.24. We found substantial heterogeneity for both sensitivity ($I^2$ = 64.6%) and specificity ($I^2$ = 78.8%), and the test for publication bias was not significant ($P$ = 0.701).

In the EVC, the models showed good diagnostic accuracy with a pooled AUC of 0.83 (95% CI: 0.79-0.86). The pooled sensitivity remained high at 0.87 (95% CI: 0.63-





[ DOI: 10.47176/mjiri.39.110 ]

*Table 2.* Continued

| Authors | Sample size | Deep learning model | Input data type | Validation sets | Discrimination statistics of the main model | TP | TN | FP | FN |
|---|---|---|---|---|---|---|---|---|---|
| Shang (14) 2025 | 284 (TC: 202, IVC: 51, EVC:31) | MobileNetV2, NASNetMobile, EfficientNetB, ResNet50, ResNet101, ResNet152, VGG16, and VGG19. | DLF + HCRF | EVC | AUC:0.8152 (95% CI:0.60–0.96) Sen:0.9565 Spe:0.2500 ACC:0.7742 | 22 | 2 | 6 | 1 |
| | | | DLF + HCRF | EVC | AUC:0.853 (95% CI: 0.652–0.988) Sen:1 Spe:0.2500 ACC:0.8065 | 23 | 2 | 6 | 0 |
| Zhao (16) 2023 | 252 (TC:202, IVC:50) | ResNet-50, Vision Transformer | DLF | IVC | AUC:0.978 (95% CI: 0.912–1.000) Sen:1 Spe:0.9 ACC:0.98 | 13 | 33 | 4 | 0 |
| Gao (17) 2019 | 602 (Initial Group: 313, Precision Group: 189, Validation: 100) | FR-CNN | DLF | IVC | AUC: 0.9541 | - | - | - | - |
| Jin(18) 2021 | 1699 (TC+IVC: 1172, EVC:527) | ResNet-18 | DLF | EVC | Median AUC: 0.876 (95%CI: 0.856–0.893) Sen: 0.743 (median) Spe:0.936 (median) | - | - | - | - |
| Liu (15) 2019 | 557 (TC:371, IVC:186) | Autoencoder | DLF | IVC | AUC: 0.946 Sen: 0.896 Spe: 0.87 95% CI for AUC: 0.925-0.978 | - | - | - | - |
| Zeng (20) 2022 | 554 (TC: 388, IVC:167, EVC:79) | Resnet152 | DLF | EVC | AUC: 0.915 (95%CI: 0.850–0.981) Sen: 0.882 Spe: 0.806 ACC: 0.861 | 15 | 50 | 12 | 2 |
| | | | DLF + HCRF | EVC | AUC: 0.581 (95%CI: 0.415–0.746) Sens: 0.706 Spec: 0.541 ACC: 0.759 | 12 | 34 | 28 | 5 |
| | | | DLF + HCRF + Clinical variables | EVC | AUC: 0.915 (95%CI: 0.850–0.981) Sen: 0.882 Spe: 0.806 ACC: 0.861 | 15 | 50 | 12 | 2 |
| Zhang (7) 2022 | 211 Multi-center domain | UDS-GCN | DLF | EVC | AUC:0.6121 Sen: 0.9818 Spe:0.1053 ACC:0.7568 | 54 | 2 | 17 | 1 |
| Zhu (19) 2025 | 293 (TC:205, IVC: 58, EVC:30) | LMML-net | DLF | EVC | AUC: 0.805 (95%CI: 0.658-1) Sen:0.81 Spe:0.769 ACC:0.794 | 22 | 23 | 5 | 8 |
| Wan (23) 2021 | 170 (TC:119, IVC:51) | Sparse autoencoder | DLF | IVC | AUC: 0.735 (95% CI:0.59-0.٨٥) Sen:0.8485 Spe:0.5789 | 28 | 1 | 8 | 5 |
| | | | DLF + HCRF | IVC | AUC: 0.872 (95% CI: 0.751-0.949) Sen:0.7879 Spe:0.9474 | 26 | 18 | 1 | 7 |

0.96), while the pooled specificity was considerably lower at 0.59 (95% CI: 0.26-0.85). The DOR was 4.53, with an LR+ of 2.39 and an LR- of 0.26. Significant heterogeneity was observed for both sensitivity ($I^2$ = 71. ٦%) and specificity ($I^2$ = 87%), but Egger's test for publication bias was nonsignificant ($P$ = 0.221).

### *Combined Deep Learning and Handcrafted Radiomics Features*

This group was assessed only in the IVC. Models combining DLF with HCRF achieved strong performance. The pooled AUC was 0.91 (95% CI: 0.88-0.93), with a pooled sensitivity of 0.85 (95% CI: 0.66-0.94) and a pooled specificity of 0.83 (95% CI: 0.6-0.94). These models showed a strong ability to confirm and rule out LNM, with a DOR of 5.55, an LR+ of 5.32, and an LR- of 0.25. We noted significant heterogeneity for both sensitivity ($I^2$ = 76.8%) and specificity ($I^2$ = 76.3%), and publication bias was unlikely ($P$ = 0.882).



[ DOI: 10.47176/mjiri.39.110 ]

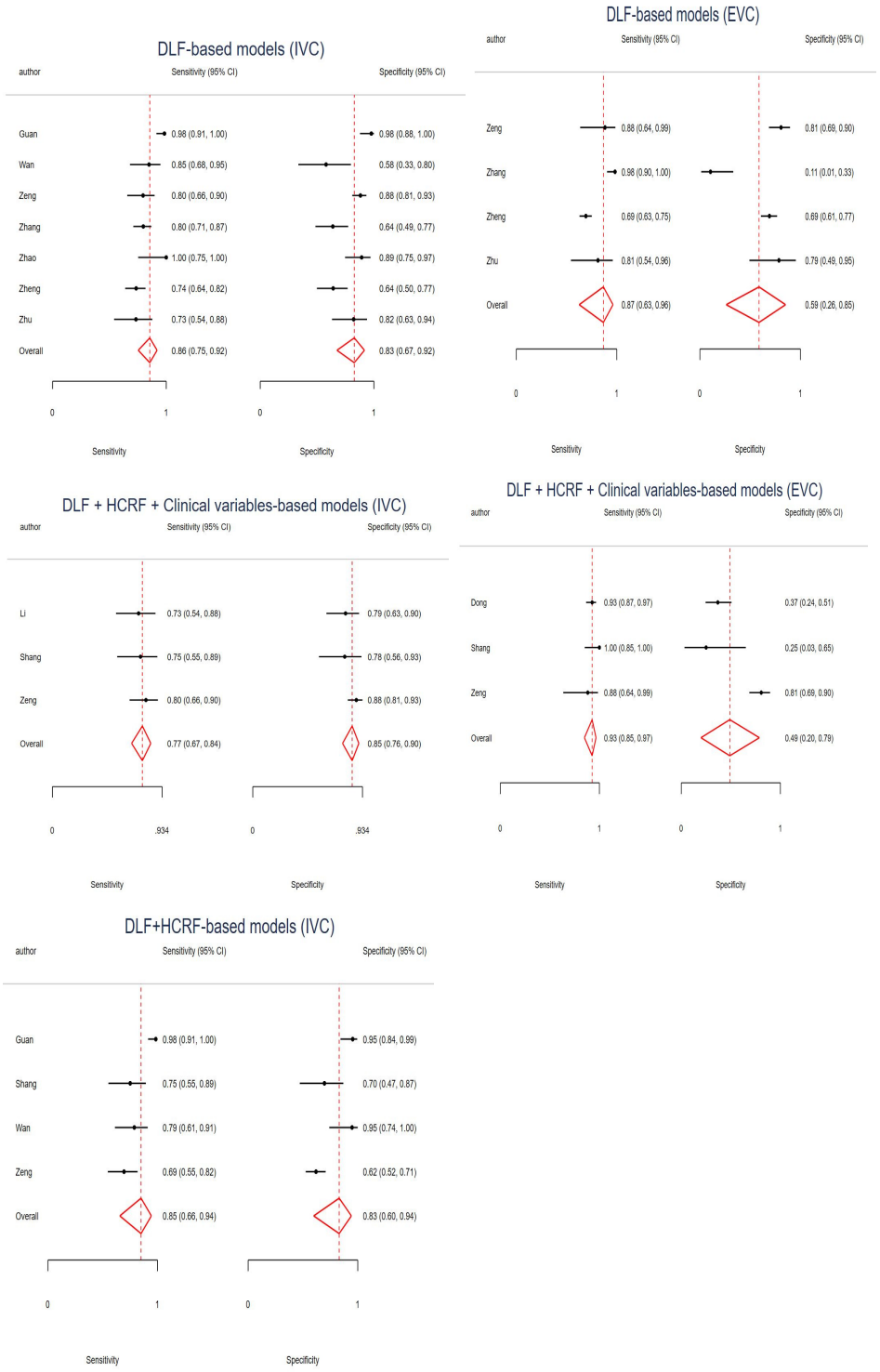


*Figure 2.* Forest Plots summarizing the pooled diagnostic performance estimates for the three primary subgroups, categorized by input data type in internal and external validations

### *Combined DL, handcrafted Radiomics Features, and Clinical Variables*

Adding clinical variables to the hybrid models yielded different outcomes between internal and external validation. In IVC, the pooled AUC was 0.75 (95% CI: 0.68-0.82); sensitivity and specificity were 0.77 (95% CI: 0.67-0.84) and 0.85 (95% CI: 0.76-0.90), respectively. The DOR was 3.59, with an LR+ of 4.77 and an LR- of 0.29. We found no significant heterogeneity for either sensitivity ($I^2$ = 2%) or specificity ($I^2$ = 8%). In contrast, EVC





[ DOI: 10.47176/mjiri.39.110 ]

showed a higher pooled AUC of 0.88 (95% CI: 0.86-0.9) and higher sensitivity at 0.93 (95% CI: 0.85-0.97), but specificity dropped to 0.49 (95% CI: 0.2-0.79). The DOR was 5.34 (95% CI: 2.86-8.95), with an LR+ of 2.07 (95% CI: 1.18-4.38) and LR- of 0.2 (95% CI: 0.11-0.35). Heterogeneity was low for sensitivity ($I^2$ = 25%) but high for specificity ($I^2$ = 79.8%). Notably, all studies in this subgroup utilized CNN-based architectures; therefore, a separate CNN subgroup analysis was not necessary.

### Subgroup Analyses

To investigate sources of heterogeneity and assess the impact of imaging protocols, DL architecture, and segmentation method, subgroup analyses were conducted.

### CT Phase in DLF-Based Models

Models trained on arterial phase images had high diagnostic performance in internal validation, with a pooled AUC of 0.91 (95% CI: 0.84-0.97). Sensitivity reached 0.90 (95% CI: 0.67-0.98) and specificity was 0.86 (95% CI: 0.54-0.97). The DOR was 12.1 (LR+, 9.77; LR-, 0.19). Substantial heterogeneity was present for both sensitivity ($I^2$ = 71.6%) and specificity ($I^2$ = 72.١%). Models using combined portal and venous phase images achieved more balanced diagnostic performance in internal validation with a pooled AUC of 0.85 (95% CI: 0.82-0.88), a sensitivity of 0.80 (95% CI: 0.72-0.87), and a specificity of 0.79 (95% CI: 0.64-0.89). The DOR was 3.56, with LR+ of 3.66 and LR- of 0.29. Heterogeneity was low for sensitivity ($I^2$ = 21.8%) but substantial for specificity ($I^2$ = 78%). We observed no publication bias. ($P$ = 0.247).

### Segmentation Method

Models employing manual segmentation performed well with a pooled AUC of 0.84 (95% CI: 0.81-0.87). Sensitivity and specificity were 0.82 (95% CI: 0.76-0.87) and 0.78 (95% CI: 0.62-0.89), respectively. The DOR was 3.87, supported by an LR+ of 3.66 and an LR- of 0.27. Heterogeneity was negligible for sensitivity ($I^2$ = 3%) but was high for specificity ($I^2$ = 72.6%).

### DL Architecture

The analysis of DL architectures was limited to the internal validation cohort. Within this group, we evaluated CNN-based models in both DLF-only and hybrid (DLF + HCRF) frameworks.

In the DLF-only models, CNN architectures demonstrated excellent performance with a pooled AUC of 0.94 (95% CI: 0.92-0.960. Sensitivity was 0.89 and specificity was 0.88, yielding a DOR of 6.15 and robust likelihood ratios. Heterogeneity was substantial for both metrics ($I^2$ = 66.7% and 74%, respectively). In hybrid models, the CNN-based subgroup yielded a higher DOR of 10.4, but this estimate was imprecise due to a wide confidence interval (0.91-45.7). This group also had a lower AUC of 0.87 and remarkably high heterogeneity for both sensitivity ($I^2$ = 79.3%) and specificity ($I^2$ = 82.8%).

### Combined Subgroup Analysis

To assess the combined effect of key methodological choices, a subgroup analysis was performed on models that used a CNN architecture, manual segmentation, and portal/venous phase images. These models, evaluated in IVC, demonstrated good diagnostic performance with a pooled AUC of 0.79 (95% CI: 0.78-0.8), a sensitivity of 0.82 (95% CI: 0.74-0.89), and a specificity of 0.83 (95% CI: 0.67–0.92). The DOR was 4.06. Heterogeneity was low for sensitivity ($I^2$ = 6.4%) but substantial for specificity ($I^2$ = 73.4%).

The complete meta-analytic results, including all subgroup findings, are summarized in Tables 3 and 4. Figure 3 illustrates the SROC curves for the 3 main groups, analyzed and categorized based on input data type. Figure 4 presents a bivariate box plot that compares sensitivity and specificity across model subgroups.

### Clinical Utility Assessment

To comprehensively evaluate the clinical utility and diagnostic impact of the different modeling approaches, Fagan nomograms and likelihood ratio matrix plots were generated for the main subgroups (based on input data type). The Fagan nomograms illustrate how a test result modifies the post-test probability of having LNM from a baseline pre-test probability. The LR matrices categorize the overall diagnostic power of each model based on its positive and negative LR, placing them into quadrants representing their value for LNM confirmation, exclusion, or both.

In the IVC, the DLF-based model demonstrated considerable clinical utility. The Fagan nomogram (Figure 5A) shows that a positive test increases the post-test probability of LNM to a convincing 82.2%. Despite this significant probability shift, the LR matrix (Figure 6A) places the summary estimate in the lower-right quadrant (RLQ), indicating that while the model is useful, it does not meet the stringent criteria for use as a definitive test. In the EVC, the model's utility diminished. The Fagan plot (Figure 5B) shows a more modest increase in probability to 72.9% after a positive test. The LR matrix (Figure 4B) also demonstrates this weaker confirmatory power, as the summary estimate stays consistently in the RLQ, which visually confirms the performance drop in external data.

The hybrid model combining DL and HCRF showed the strongest performance in confirming the disease. The Fagan nomogram (Figure 5C) reveals that a positive test (LR+ of 5.32) elevates the post-test probability to 85.1%, the highest among all groups. However, the LR matrix plot (Figure 6C) shows that even with this strong performance, the summary estimate still resides in the RLQ, failing to cross the threshold for a high-value confirmatory test, although 1 study did achieve this. Models integrating DL features with radiomics and clinical variables in IVC demonstrated a notable ability to rule out LNM. The Fagan plot (Figure 5E) shows that a negative test (LR = 0.29) reduces the probability to 17.8%. Even so, the LR matrix (Figure 6D) kept this estimate in the RLQ. In the EVC, this model's utility for confirming the disease was the weakest. The Fagan plot (Figure 5D) shows that a





[ DOI: 10.47176/mjiri.39.110 ]

*Table 3.* Pooled Effect Size Based on Input Data Type for Deep Learning Models in Detecting Lymph Node Metastasis

| Input data type | N. studies | Validation | AUC (95% CI) | Sen (95% CI) | Spe (95% CI) | DOR (95% CI) | LR+ (95% CI) | LR- (95% CI) | $I^2$ | *P*-Value | Publication bias |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DLF-based models | 7 | IVC | 0.91 (0.88-0.93) | 0.86 (0.75-0.92) | 0.83 (0.67-0.92) | 4.43 (2.63-6.86) | 4.23 (2.19-7.74) | 0.24 (0.146-0.38) | Sen:64.62 Spec:78.83 | <0.001 | 0.701 |
| DLF-based models | 4 | EVC | 0.83 (0.79-0.86) | 0.87 (0.63-0.96) | 0.59 (0.26-0.85) | 4.53 (1.91-9.85) | 2.38 (1.22-5.36) | 0.26 (0.1-0.52) | Sen:71.65 Spe:87.07 | <0.001 | 0.221 |
| DLF + HCRF-based models | 4 | IVC | 0.91 (0.88-0.93) | 0.85 (0.66-0.94) | 0.83 (0.6-0.94) | ۵/۵۵ (۱/۵۶-۱۴/۱) | 5.32 (1.47-14.5) | 0.25 (0.07-0.64) | Sen:76.8 Spe:76.34 | <0.001 | 0.882 |
| DLF + HCRF + Clinical variables-based models | 3 | IVC | 0.75 (0.68-0.82) | 0.77 (0.67-0.84) | 0.85 (0.76 - 0.90) | 3.59 (2.43-5.24) | 4.77 (2.88-7.61) | 0.29 (0.19-0.41) | Sen: 2.05 Spe:8.05 | 0.998 | - |
| DLF + HCRF + Clinical variables-based models | 3 | EVC | 0.88 (0.86-0.90) | 0.93 (0.85-0.97) | 0.49 (0.2 - 0.79) | 5.34 (2.86-8.95) | 2.07 (1.18-4.38) | 0.20 (0.11-0.35) | Sen:25.09 Spe:79.81 | <0.001 | - |

DLF: Deep learning feature HCRF: Handcrafted radiomics feature Validation cohort IVC: International validation cohort EVC: External validation AUC: Area under the curve Sen: Sensitivity Spe: Specificity CI: Confidence interval DOR: Diagnostic odds ratio LR+: positive likelihood ratio LR-: Negative likelihood ratios

*Table 4.* Subgroup analyses for deep learning models in detecting lymph node metastasis

| Subgroup | N. studies | Validation | AUC (95% CI) | SEN (95% CI) | Spe (95% CI) | DOR (95% CI) | LR+ (95% CI) | LR- (95% CI) | $I^2$ | *P*-Value | Publication bias |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CT phases DLF-based models in the arterial Phase | 3 | IVC | 0.91 (0.84 - 0.97) | 0.90 (0.67-0.98) | 0.86 (0.54 - 0.97) | 12.1 (1.29-48) | 9.77 (1.21-40.8) | 0.19 (0.02-0.78) | Sen:71.63 Spe:72.75 | <0.001 | - |
| DLF-based models in the combined (portal venous/venous) Phase | 4 | IVC | 0.85 (0.82-0.88) | 0.80 (0.72-0.87) | 0.79 (0.64-0.89) | 3.56 (2.37-4.97) | 3.66 (1.91-6.79) | 0.29 (0.2-0.42) | Sen:21.81 Spe:78.01 | 0.006 | 0.247 |
| Segmentation method DLF-based models with manual Segmentation | 4 | IVC | 0.84 (0.81-0.87) | 0.82 (0.76-0.87) | 0.78 (0.62-0.89) | 3.87 (2.64-5.33) | 3.66 (1.89-6.95) | 0.27 (0.19-0.38) | Sen:3.09 Spe:72.6 | 0.026 | 0.595 |
| DL architectures CNN-based algorithms in DLF-based models | 5 | IVC | 0.94 (0.92-0.96) | 0.89 (0.74-0.96) | 0.88 (0.73-0.95) | 6.15 (2.68-12.4) | 6.08 (2.6-12.5) | 0.19 (0.08-0.37) | Sen:66.66 Spe:74.01 | <0.001 | 0.742 |
| CNN-based algorithms in DLF + HCRF-based models | 3 | IVC | 0.87 (0.72-0.96) | 0.88 (0.56-0.98) | 0.8 (0.51-0.94) | 10.4 (0.91-45.7) | 5.74 (0.92-19.2) | 0.26 (0.02-0.1.1) | Sen:79.25 Spe:82.79 | <0.001 | -_ |
| Combined analysis CNN-based with manual Segmentation in combined (venous/Portal) Phases in DLF-based models | 3 | IVC | 0.79 (0.78-0.80) | 0.82 (0.74-0.89) | 0.83 (0.67-0.92) | 4.06 (2.71-5.74) | 4.69 (3.44-6.39) | 4.61 (2.09-9.57) | Sen:6.42 Spe:73.43 | 0.088 | __-_ |

DLF: Deep learning feature HCRF: Handcrafted radiomics feature IVC: Internal validation EVC: External validation AUC: Area under the curve Sen: Sensitivity Spe: Specificity CI: Confidence interval DOR: Diagnostic odds ratio LR+: positive likelihood ratio LR-: Negative likelihood ratios

positive test (LR+ of 2.07) only raises the post-test probability to 71.4%. This is visually reinforced by its position in the LR matrix (Figure 6E), which had the lowest LR+ of all subgroups.

### *Quality Assessment*

The assessment of the methodological quality of the included studies revealed multiple sources of bias and concerns regarding applicability. In the patient selection domain, the retrospective design of all studies was rated as having "unclear" regarding bias. The index test domain presented significant issues; the common failure to prespecify a diagnostic threshold was rated as a high risk of bias, while the lack of reported radiologist blinding during segmentation also raised "unclear." Similarly, in the reference standard domain, the lack of reported pathologist blinding was a source of "unclear" for interpretation bias.

Regarding applicability, the predominance of single-center studies was a primary concern, raising doubts about the generalizability of the findings to diverse clinical settings. Details of the study assessment are provided in Table 5.

## Discussion

A systematic review and meta-analysis assessed the potential of CT-based DL models in predicting LNM in GC





[ DOI: 10.47176/mjiri.39.110 ]

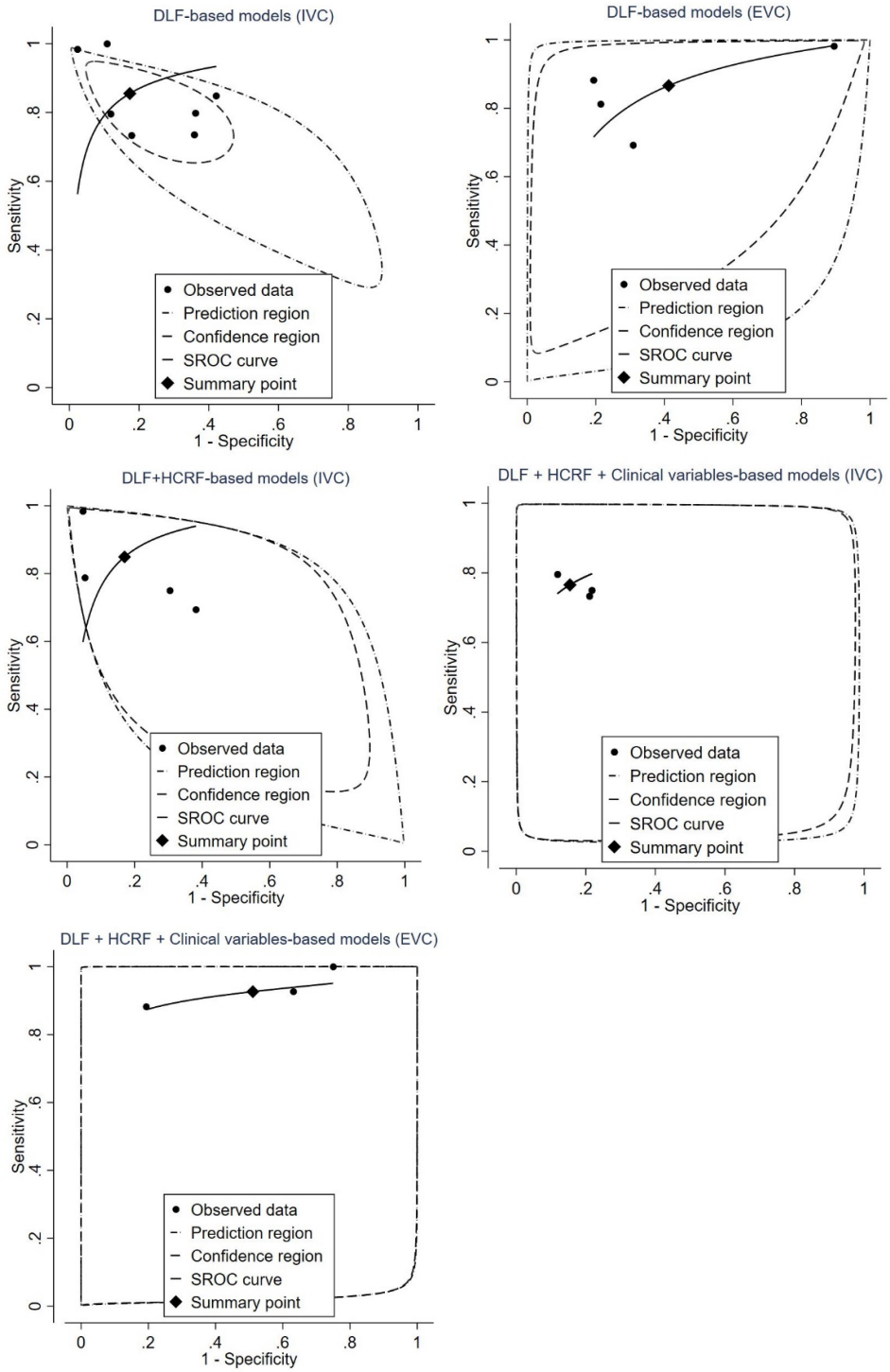


*Figure 3.* Illustrates the SROC curves for three key models based on input data type, highlighting differences in diagnostic performance between internal (IVC) and external validation (EVC) sets

patients. LNM is a critical prognostic factor influencing treatment planning and patient survival. Early detection of LNM is vital for designing appropriate treatments for patients. Our study's findings demonstrated the transformative potential of DL models in enhancing diagnostic accuracy for LNM prediction from CT images.

DL is a subfield of ML characterized by its capability to learn features from existing data autonomously. Recent research has demonstrated that DL-based methodologies, particularly CNN models, exhibit enhanced accuracy in diagnosing and identifying various types of cancer (24). DNNs are elegant algorithms that possess substantial





[ DOI: 10.47176/mjiri.39.110 ]

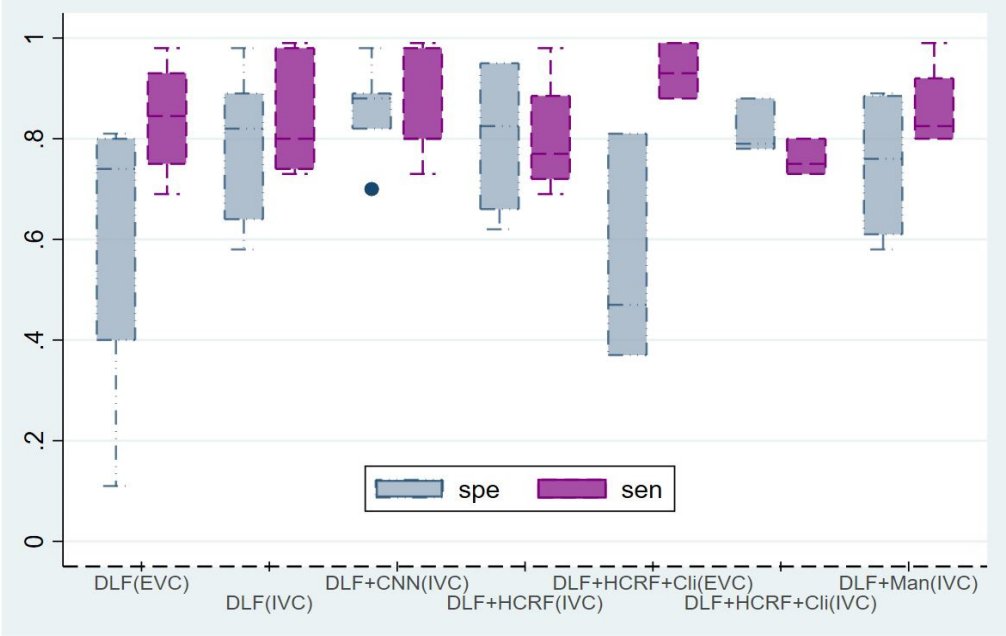


*Figure 4.* Bivariate box plot. Compare the analysis of sensitivity and specificity across model subgroups. This box plot illustrates the distribution of pooled sensitivity (sen) and specificity (spe) for key subgroups. The plot highlights the performance gap between internal and external validation; while sensitivity remains high in both settings, specificity drops significantly and shows greater variability in external validation cohorts compared to internal validation cohorts.

computational power to analyze large images, including hematoxylin and eosin-stained whole-slide images obtained from biopsies or surgically resected tissues. DNNs extend their capabilities beyond pathology images, effectively examining other medical imaging modalities such as CT scans, magnetic resonance imaging (MRI), and mammograms (25). A 2017 study employed artificial intelligence (AI) for skin lesion classification, utilizing 129,450 clinical images. The researchers compared the performance of DCNNs with 21 board-certified dermatologists. The findings revealed that artificial intelligence achieved comparable accuracy in classifying skin cancer to dermatologists (26). Zhou et al used MRI images from patients with hepatocellular cancer for grading, and they reached an AUC of 0.83 (27). The study by Jin et al. offers valuable insights into the diagnostic potential of DL models. They developed a ResNet-18-based DL model to predict LNM across 11 nodal stations in GC patients. The model demonstrated excellent performance in the external validation set, achieving a median AUC of 0.876 (range, 0.856-0.893), which substantially outperformed conventional clinicopathological models (median AUC, 0.652). Importantly, Grad-CAM visualizations revealed that the DL model focused on specific intra- and peritumoral regions during prediction, indicating that it could capture subtle imaging patterns even without direct lymph node segmentation. This supports the potential of DL approaches to identify metastasis by recognizing complex spatial features (18).

Due to the heterogeneity of the biological characteristics of GC, different treatment methods are used based on the histology, morphology, and depth of tumor invasion. Since the selection of the appropriate treatment method is mainly limited to preoperative imaging findings, there is a significant increase in overdiagnosis of the disease in the early stages and underdiagnosis in advanced stages (15). Hasegawa et al evaluated 315 patients with GC to determine the accuracy of multidetector row CT images in the prediction of serosal invasion and nodal metastases. The findings demonstrated an overall diagnostic accuracy of 75.9% for N staging (28). Liu et al developed ML models to analyze preoperative CT images and clinical data of patients with locally advanced gastric cancer to predict D1 versus D2 lymphadenectomy. The AE model reduced overtreatment by 14% to 20%. Given the high prevalence of advanced GC in Eastern countries, relatively few patients with early-stage disease were included in this study. Also, radiomic features were not analyzed in this study (15). Rathore et al used multiparametric MRI, incorporating pathology images, to develop RadPath signatures in patients with glioblastoma (29). Radiomics, a novel technology, has significant potential in the field of oncology. Classifying images using DL and integrating it with radiomics is a challenging approach; hence, limited studies have employed this method in the field of GC.

Our systematic review and meta-analysis of 14 studies demonstrate that CT-based DL models have a significant potential for detecting LNM in GC. The most critical finding of this study is the contrast between high internal accuracy and compromised external performance. While DL models achieve excellent diagnostic accuracy on internal validation sets, with DLF-based models reaching a pooled AUC of 0.91, their performance degrades when applied to external datasets. This was most evident in the sharp de-



[ DOI: 10.47176/mjiri.39.110 ]

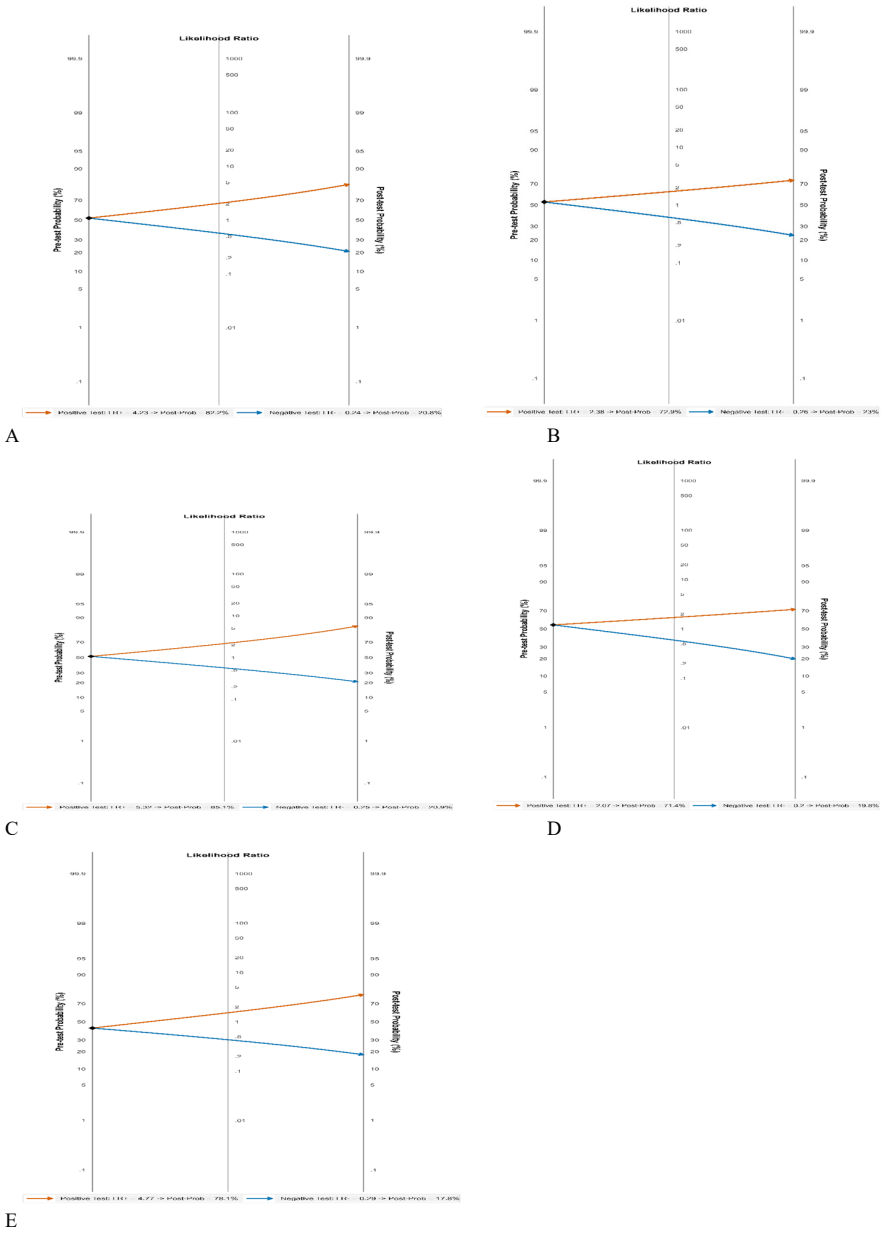


*Figure 5.* Fagan nomograms, illustrating the change in post-test probability of LNM based on positive and negative test results. (A) DLF-based model in internal validation cohort (IVC), (B) DLF-based model in external validation cohort (EVC), (C) DLF + HCRF model in IVC, (D) DLF + HCRF + Clinical Variables model in EVC, (E) DLF + HCRF + Clinical Variables model in IVC

cline in specificity (from 0.83 to 0.59), indicating a high rate of false positives in new data. This generalizability gap suggests that models are overfitting to site-specific characteristics of the training data (e.g., scanner protocols, patient populations) rather than learning robust, universal biological markers of metastasis. Our results indicate that combining multiple data modalities does not necessarily improve all performance metrics uniformly. In internal validation, while hybrid models (DLF + HCRF) and DLF-based models achieved an identical pooled AUC of 0.91, the hybrid approach demonstrated a notable improvement in confirmatory power. Specifically, the favorable likelihood ratio increased from 4.23 in DLF-based models to 5.32 in the hybrid models, suggesting that the inclusion of handcrafted radiomic features can enhance the model's ability to correctly confirm metastasis. Therefore, rather than offering limited added value, radiomics appears to specifically increase the model's confirmatory utility, even if it does not alter the overall discrimination.

The inclusion of clinical variables presented a mixed outcome: while sensitivity improved (0.93) in external validation, specificity declined sharply (0.49 vs. 0.85 in





[ DOI: 10.47176/mjiri.39.110 ]

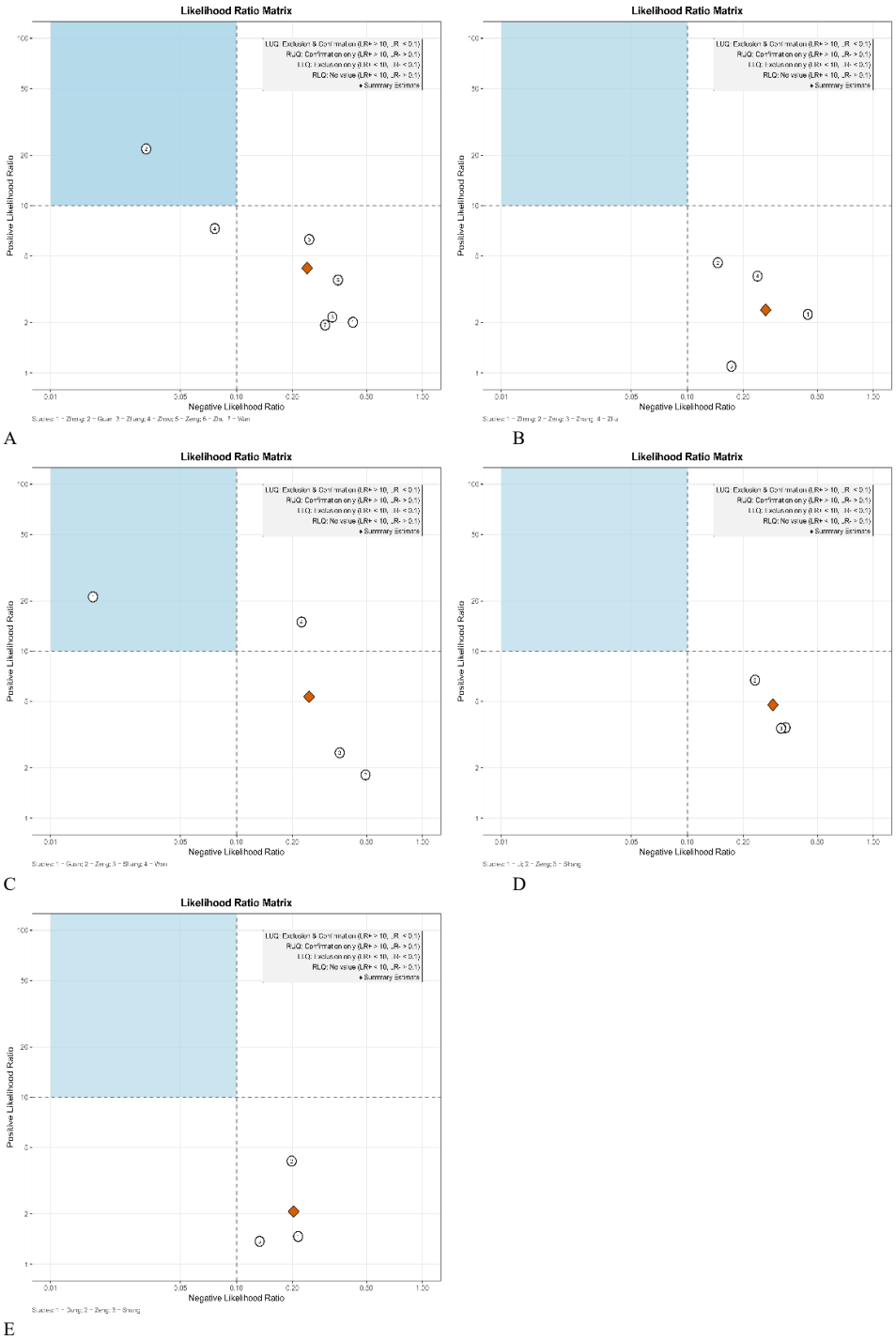


*Figure 6.* Likelihood ratio matrix plots for the three main subgroups, categorized by input data type. The plots display the positive likelihood ratio (LR+) against the negative likelihood ratio (LR−), with quadrants indicating diagnostic utility for confirming and/or excluding LNM.
A: DLF-based models in IVC
B: DLF-based models in EVC
C: DLF + HCRF-based models in IVC

internal), likely due to overfitting, rather than the effect of clinical data itself. This underscores the necessity for external validation, particularly for multimodal models.

One of the notable findings was the exceptional performance of the study by Guan et al., which, as a single study, met the criteria for a high-value diagnostic test (LR+ > 10 and LR- < 0.1). While the original paper's internal analysis might have reported nuanced enhancements from combining features, within our meta-analysis, this study stands out as a successful example of the poten-




[ DOI: 10.47176/mjiri.39.110 ]

*Table 5.* Quality Assessment of Included Studies Based on QUADAS2

| | Risk of bias | | | | Application concerns | | |
|---|---|---|---|---|---|---|---|
| Study | Patient selection | Index test | Reference standard | Flow and timing | Patient selection | Index test | Reference standard |
| Dong (13) | Unclear | High | Unclear | Low | Low | Low | Low |
| Zheng (12) | Unclear | Low | Low | Low | Low | low | Low |
| Zhang (7) | Unclear | Low | Unclear | Low | High | Low | Low |
| Gao (17) | Unclear | High | Unclear | Low | High | Low | Low |
| Guan (21) | Unclear | High | Low | Low | High | Low | Low |
| Shang (14) | Unclear | High | Unclear | Low | Low | Low | Low |
| Wan (23) | Unclear | High | Unclear | Low | High | Low | Low |
| Li (22) | Unclear | Unclear | Unclear | Low | High | Low | Low |
| Liu (15) | Unclear | Unclear | Unclear | Low | High | Low | Low |
| Jin (18) | Unclear | Low | Unclear | Low | Low | Low | Low |
| Zeng (20) | Unclear | High | Unclear | Low | Low | Low | Low |
| Zhu (19) | Unclear | High | Unclear | Low | Low | Low | Low |
| Zhao (16) | Unclear | High | Unclear | Low | High | Low | Low |
| Zhang (4) | Unclear | High | Unclear | low | High | low | low |

tial of hybrid models. Its high performance suggests that achieving excellent diagnostic accuracy is possible with careful feature selection and imaging protocols. Therefore, rather than being viewed as an example of non-enhancement, this study should be considered a benchmark for future research (21). Due to the insufficient number of studies, a quantitative meta-analysis could not be performed for the subgroup of models that directly combined DLF with clinical variables in our study.

Our subgroup analyses highlight the critical impact of specific technical decisions on model performance. For instance, CNN-based models emerged as top performers in terms of raw AUC (0.94), but this finding was tempered by significant heterogeneity for both sensitivity ($I^2$ = 66.7%) and specificity ($I^2$ = 74%). Similarly, the arterial-phase subgroup demonstrated the highest diagnostic odds ratio (12.1), suggesting it may be the optimal imaging protocol. However, the considerable heterogeneity within this group for both sensitivity ($I^2$ = 71.6%) and specificity ($I^2$ = 72.8%) limits the reliability of this conclusion. It is noticeable that models using arterial phase CT showed a higher DOR than those using the combined portal/venous phase (12.1 vs. 3.56). In the study by Liu et al., an autoencoder was used to develop a decision-making model in the arterial and parenchymal phases. This model achieved the highest AUC (0.946) among the examined models (18). In a study conducted by Gao et al, they utilized triphasic CTs; the researchers initially trained a DL model on 1,371 CT scans labeled by radiologists for LNM (AUC, 0.89). Because the outcomes of the initial phase were unsatisfactory, based on the pathology reports of 250 patients, 3 senior radiologists relabeled 1,004 CT scans with exceptional accuracy. The improved model achieved an AUC of 0.9541, highlighting the significant impact of the quality and precision of training data on a DL model's ultimate performance (17).

Our meta-analysis revealed substantial heterogeneity in the diagnostic performance of DL models for detecting LNM, with $I^2$ values for sensitivity and specificity exceeding 64% in both DLF-only and DLF combined with HCRF models, likely reflecting methodological diversity across studies. Notably, incorporating clinical variables in the internal validation cohort markedly reduced heterogeneity for both sensitivity ($I^2$ = 2.1%) and specificity ($I^2$ = 8.1%), suggesting that clinical information enhances model standardization and reproducibility for real-world use. However, heterogeneity varied unevenly across performance metrics. In subgroups with technical standardization, such as manual segmentation or portal/venous-phase CT, sensitivity was relatively consistent (e.g., $I^2$ = 6.4% for CNN-based models with manual segmentation on portal/venous-phase scans). Still, specificity remained highly variable ($I^2$ = 73.4%). This persistent variability in specificity, even in the most homogeneous subgroup, likely stems from clinical factors, such as diverse patient populations or imaging appearances of benign lymph nodes, indicating that technical standardization alone cannot fully address performance inconsistencies. The Fagan nomograms and LR matrices offer complementary insights into the clinical utility of the evaluated models. While the Fagan plots indicate that all models can meaningfully alter the post-test probability of LNM, suggesting potential support in clinical decision-making, the LR matrices expose a key limitation: none of the strategies consistently achieve the diagnostic certainty needed to serve as standalone tools for ruling in or ruling out disease. Notably, both visualization methods consistently reflect a marked decline in diagnostic strength during external validation, emphasizing the ongoing challenge of generalizability.

The included studies used different segmentation techniques to delineate tumor regions on CT images. Most studies relied on manual or semi-automatic segmentation methods. Many studies have used 2-dimensional (2D) features from single slices rather than 3D volumetric features, which may not accurately represent the entire tumor and could impact feature accuracy and model robustness. In a semi-automated segmentation technique, a combination of automatic and manual algorithms was applied (e.g., Guan et al and Zheng et al). Zhu et al utilized a 3D Attention-UNet with Focal Tversky Loss for semi-automatic tumor segmentation. Initially, radiologists manually annotated tumors with 3D Slicer, and the model refined these annotations to create a probability map (Ptumor). A key



[ DOI: 10.47176/mjiri.39.110 ]

aspect of this method is the use of attention gates, which focus on areas relevant to the tumor while minimizing the effect of unrelated background information. This approach helps resolve challenges related to class imbalance and irregular tumor shapes. The dice scores achieved were 0.582 for training and 0.547 for testing, indicating moderate accuracy in segmentation (19). Remarkably, the study by Shang et al. employed fully automated U-mamba segmentation for the spleen to decrease time and inter-observer variability, which was a unique approach (14). These examples underscore that technical decisions at each stage of the modeling pipeline can significantly influence overall model performance.

In retrospective studies, where pathological assessments are predetermined based on archived records, initial review of CT images by radiologists is a crucial step in training DL models. This process relies heavily on preexisting pathological results as the gold standard. However, the lack of blinding reporting between the labeling process and subsequent model predictions, as observed in some of the reviewed studies, raises potential concerns about the risk of bias. Also, a critical methodological issue identified in the assessed studies was the common failure to prespecify a diagnostic threshold. This omission introduces a significant risk of bias, as it raises the possibility that the cutoff points were selected post-hoc, after the data had been analyzed.

The assessment of clinical utility through Fagan nomograms and likelihood ratio matrices provides context for these statistical findings. Our results show that these models are clinically useful for risk stratification; for example, a positive test from the best-performing hybrid model could increase the post-test probability of LNM to over 85%. However, the LR matrices clearly illustrate that none of the current modeling strategies, on average, meet the stringent criteria (LR+ > 10 or LR- < 0.1) required for a standalone test to confirm or rule out the disease definitively.

This review has several limitations. First, the small number of studies in many subgroups limited the statistical power of the meta-analysis. Additionally, subgroup analyses for models based on DLF + clinical variables, automatic and semi-automatic segmentation methods, unenhanced versus enhanced CT phases, and for other DL architectures could not be performed due to insufficient data, precluding quantitative comparison with manually segmented or arterial/venous imaging protocols and CNN-based models. Second, while our subgroup analyses explored sources of the substantial and complex heterogeneity found throughout our results, its persistence—remarkably the asymmetric variability in specificity—underscores the challenge of clinical translation. A few studies included EVC, and single-center designs reduce applicability to diverse populations. Third, all included studies were retrospective, and a potential lack of reporting on blinding during data annotation introduces a risk of bias. Our exclusive focus on CT-based DL models also excluded potentially informative multimodal approaches (e.g., MRI, pathology, or clinical-genomic integration), which narrows the scope of our conclusions. Finally, the predominance of studies originating from Eastern countries may limit their global applicability.

It is recommended that future studies be conducted with standardized protocols for validation and prospectively in a multi-center manner. For future research to yield more robust and reliable findings, studies must establish and register a clear, pre-specified diagnostic threshold before data collection and analysis. Retrospective findings must be validated in prospective clinical trials with large sample sizes to confirm their real-world utility and impact on patient outcomes. Different phases of CT scans should be systematically compared, and imaging protocols should be optimized, with reporting guidelines to reduce heterogeneity. Automated techniques should be used for segmentation to eliminate observer bias. Feature selection methods should be investigated to reduce redundancy, and optimal features should be used. Genomic, proteomic, imaging, pathological, and clinical data should be integrated to design robust predictive models. Furthermore, model interpretability should be enhanced using tools such as Grad-CAM and SHAP values. Concurrently, given our finding that specificity heterogeneity persists even after technical standardization, future research should aim to identify and model the clinical or biological factors that contribute to this variability, as this appears to be a critical barrier to improving model generalizability.

## Conclusion

CT-based DL models show strong potential for predicting LNM in GC, with high diagnostic accuracy in internal validation. Yet, a notable drop in specificity during external validation highlights the risk of overfitting and limited generalizability. Model complexity had mixed effects: radiomic features added confirmatory strength, while clinical variables reduced heterogeneity but introduced variable trade-offs. Promising techniques, such as CNN architectures and arterial-phase imaging, require cautious interpretation due to persistent heterogeneity in the data. Extensive, prospective studies involving diverse populations, independent validation, and standardized protocols are necessary to confirm the clinical utility of this approach.

### *Declaration of Generative AI and AI-assisted Technologies in the Writing Process*

During the editing of this article, Gemini Pro 2.5 was used for spelling corrections and increased readability. It is important to emphasize that the authors bear full responsibility for the content written in this article.

### *Authors' Contributions*

Armin Majd Gharamaleki and Arman Majd Gharamaleki conducted the literature search, performed study screening, curated the raw data, and drafted the original manuscript. Alireza Amanollahi contributed to the statistical analysis and the preparation of figures. Sarvin Tabibzadeh conceptualized and designed the study, developed the methodology, performed the formal analysis, created the figures, and supervised the entire project. All authors read and approved the final manuscript.





[ DOI: 10.47176/mjiri.39.110 ]

### Ethical Considerations

This study is a systematic review and meta-analysis of previously published literature. Therefore, ethical approval was not required.

### Acknowledgment

The authors have no acknowledgements.

### Conflict of Interests

The authors declare that they have no competing interests.

## References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.
2. Deng JY, Liang H. Clinical significance of lymph node metastasis in gastric cancer. World J Gastroenterol. 2014;20(14):3967-75.
3. Hochwald SN, Kim S, Klimstra DS, Brennan MF, Karpeh MS. Analysis of 154 actual five-year survivors of gastric cancer. J Gastrointest Surg. 2000;4(5):520-5.
4. Zhang AQ, Zhao HP, Li F, Liang P, Gao JB, Cheng M. Computed tomography-based deep-learning prediction of lymph node metastasis risk in locally advanced gastric cancer. Front Oncol. 2022;12:969707.
5. Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Ann Oncol. 2017;28(6):1191-206.
6. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng. 2021;109(5):820-38.
7. Zhang Y, Yuan N, Zhang Z, Du J, Wang T, Liu B, et al. Unsupervised domain selective graph convolutional network for preoperative prediction of lymph node metastasis in gastric cancer. Med Image Anal. 2022;79:102467.
8. Zhang H, Qie Y. Applying Deep Learning to Medical Imaging: A Review. Applied Sciences. 2023;13(18):10521.
9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj. 2021;372:n71.
10. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. Jama. 2018;319(4):388-96.
11. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
12. Zheng Y, Qiu B, Liu S, Song R, Yang X, Wu L, et al. A transformer-based deep learning model for early prediction of lymph node metastasis in locally advanced gastric cancer after neoadjuvant chemotherapy using pretreatment CT images. EClinicalMedicine. 2024;75:102805.
13. Dong D, Fang MJ, Tang L, Shan XH, Gao JB, Giganti F, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Ann Oncol. 2020;31(7):912-20.
14. Shang H, Fang Y, Zhao Y, Mi N, Cao Z, Zheng Y. Deep Learning and Radiomics for Gastric Cancer Lymph Node Metastasis: Automated Segmentation and Multi-Machine Learning Study from Two Centers. Oncology. 2025:1-16.
15. Liu C, Qi L, Feng QX, Sun SW, Zhang YD, Liu XS. Performance of a machine learning-based decision model to help clinicians decide the extent of lymphadenectomy (D1 vs. D2) in gastric cancer before surgical resection. Abdom Radiol (NY). 2019;44(9):3019-29.
16. Zhao Y, Li L, Han K, Li T, Duan J, Sun Q, et al. A radio-pathologic integrated model for prediction of lymph node metastasis stage in patients with gastric cancer. Abdom Radiol (NY). 2023;48(11):3332-42.
17. Gao Y, Zhang ZD, Li S, Guo YT, Wu QY, Liu SH, et al. Deep neural network-assisted computed tomography diagnosis of metastatic lymph nodes from gastric cancer. Chin Med J (Engl). 2019;132(23):2804-11.
18. Jin C, Jiang Y, Yu H, Wang W, Li B, Chen C, et al. Deep learning analysis of the primary tumour and the prediction of lymph node metastases in gastric cancer. Br J Surg. 2021;108(5):542-9.
19. Zhu H, Yang Z, Zheng C, Jiang P, Fang Y, Xu Y, et al. A 3D end-to-end multi-task learning network for predicting lymph node metastasis at multiple nodal stations in gastric cancer. Biomedical Signal Processing and Control. 2025;108:107802.
20. Zeng Q, Li H, Zhu Y, Feng Z, Shu X, Wu A, 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. Guan X, Lu N, Zhang J. Computed Tomography-Based Deep Learning Nomogram Can Accurately Predict Lymph Node Metastasis in Gastric Cancer. Dig Dis Sci. 2023;68(4):1473-81.
22. Li J, Dong D, Fang M, Wang R, Tian J, Li H, et al. Dual-energy CT-based deep learning radiomics can improve lymph node metastasis risk prediction for gastric cancer. Eur Radiol. 2020;30(4):2324-33.
23. Wan Y, Yang P, Xu L, Yang J, Luo C, Wang J, et al. Radiomics analysis combining unsupervised learning and handcrafted features: A multiple-disease study. Med Phys. 2021;48(11):7003-15.
24. Tandon R, Agrawal S, Rathore NPS, Mishra AK, Jain SK. A systematic review on deep learning-based automated cancer diagnosis models. J Cell Mol Med. 2024;28(6):e18144.
25. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial Intelligence in Cancer Research and Precision Medicine. Cancer Discov. 2021;11(4):900-15.
26. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.
27. Zhou Q, Zhou Z, Chen C, Fan G, Chen G, Heng H, et al. Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. Comput Biol Med. 2019;107:47-57.
28. Hasegawa S, Yoshikawa T, Shirai J, Fujikawa H, Cho H, Doiuchi T, et al. A prospective validation study to diagnose serosal invasion and nodal metastases of gastric cancer by multidetector-row CT. Ann Surg Oncol. 2013;20(6):2016-22.
29. Rathore S, Nasrallah M, Mourelatos Z. NIMG-76. radiopathomics: integration of radiographic and histologic characteristics for prognostication in glioblastoma. Neuro-Oncology. 2019;21(Supplement_6):vi178-vi9.



[ DOI: 10.47176/mjiri.39.110 ]

*Appendix.* Literature searching strategies in PubMed, Web of science and Embase (The search conducted until 5 May 2025)

| NO. | Search query for Web of Science | Result |
|---|---|---|
| #1 | TS=("Stomach Neoplasms" OR "Neoplasm, Stomach" OR "Stomach Neoplasm" OR "Gastric Neoplasm*" OR "Neoplasm, Gastric" OR "Neoplasms, Gastric" OR "Neoplasms, Stomach" OR "Cancer of Stomach" OR "Stomach Cancer*" OR "Cancer of the Stomach" OR "Gastric Cancer*" OR "Cancer, Gastric" OR "Cancers, Gastric" OR "Cancers, Stomach" OR "Cancer, Stomach" OR "Gastric Cancer, Familial Diffuse") | 126094 |
| #2 | TS=("Lymphatic Metastasis" OR "Lymphatic Metastases" OR "Lymph Node Metastasis" OR "Lymph Node Metastases" OR "Metastasis, Lymph Node") | 75266 |
| #3 | TS=("Deep Learning" OR "Learning, Deep" OR "Hierarchical Learning" OR "Learning, Hierarchical" OR "Neural Networks, Computer" OR "Machine Learning" OR "Artificial Intelligence" OR "Convolutional Neural Networks" OR "CNN" OR "Transformer Models" OR "Vision Transformer") | 1012764 |
| #4 | #1 AND #2 AND #3 | 83 |

| NO. | Search query for PubMed | Result |
|---|---|---|
| #1 | Stomach Neoplasms[mh] OR Neoplasm, Stomach[tiab] OR Stomach Neoplasm[tiab] OR Gastric Neoplasm*[tiab] OR Neoplasm, Gastric[tiab] OR Neoplasms, Gastric[tiab] OR Neoplasms, Stomach[tiab] OR Cancer of Stomach[tiab] OR Stomach Cancer*[tiab] OR Cancer of the Stomach[tiab] OR Gastric Cancer*[tiab] OR Cancer, Gastric[tiab] OR Cancers, Gastric[tiab] OR Cancers, Stomach[tiab] OR Cancer, Stomach[tiab] OR Gastric Cancer, Familial Diffuse[tiab] | 149502 |
| #2 | Lymphatic Metastasis[mh] OR Lymphatic Metastases[tiab] OR Lymph Node Metastasis[tiab] OR Lymph Node Metastases[tiab] OR Metastasis, Lymph Node[tiab] | 136303 |
| #3 | Deep Learning[mh] OR Learning, Deep[tiab] OR Hierarchical Learning[tiab] OR Learning, Hierarchical[tiab] OR "Neural Networks, Computer"[mh] OR "Machine Learning"[mh] OR "Artificial Intelligence"[mh] OR "Convolutional Neural Networks"[tiab] OR "CNN"[tiab] OR "Transformer Models"[tiab] OR "Vision Transformer"[tiab] | 249615 |
| #4 | #1 AND #2 AND #3 | 79 |

| NO. | Search query for Embase | Result |
|---|---|---|
| #1 | Stomach Neoplasms OR Neoplasm, Stomach OR Stomach Neoplasm OR Gastric Neoplasm* OR Neoplasm, Gastric OR Neoplasms, Gastric OR Neoplasms, Stomach OR Cancer of Stomach OR Stomach Cancer* OR Cancer of the Stomach OR Gastric Cancer* OR Cancer, Gastric OR Cancers, Gastric OR Cancers, Stomach OR Cancer, Stomach OR Gastric Cancer, Familial Diffuse. | 317,006 |
| #2 | Lymphatic Metastasis OR Lymphatic Metastases OR Lymph Node Metastasis OR Lymph Node Metastases OR Metastasis, Lymph Node | 288241 |
| #3 | Deep Learning OR Learning, Deep OR Hierarchical Learning OR Learning, Hierarchical OR Neural Networks, Computer OR Machine Learning OR Artificial Intelligence OR Convolutional Neural Networks OR CNN OR Transformer Models OR Vision Transformer | 728374 |
| #4 | #1 AND #2 AND #3 | 392 |



[ DOI: 10.47176/mjiri.39.110 ]