J Clin Res Paramed Sci. 2026 June; 15(1): e173819 https://doi.org/10.5812/jcrps-173819

Published Online: 2026 June 30 **Systematic Review**

Deep Learning-Based Lung Cancer Classification from CT Images: A Systematic Review of Morphological Features, Model Architectures, and Classification Performance

Mohsen Taherian [1,*], S Hadi Yaghoubyan [1,2,**], Karamollah Bagherifard [1,2], Razieh Malekhosseini [1]

[1] Department of Computer Engineering, Yas. C., Islamic Azad University, Yasuj, Iran
[2] Young Researchers and Elite Club, Yas., C., Islamic Azad University, Yasuj, Iran

[*] ***Corresponding Author:*** Department of Computer Engineering, Yas. C., Islamic Azad University, Yasuj, Iran. Email: mtaherian389@gmail.com
[**] ***Corresponding Author:*** Department of Computer Engineering, Yas. C., Islamic Azad University, Yasuj, Iran. Email: yaghoobian.h@gmail.com

Received: 25 May, 2026; **Revised:** 6 June, 2026; **Accepted:** 18 June, 2026

Abstract

Context: Lung cancer accounts for approximately 12.4% of new cancer diagnoses and 18.0% of cancer-related deaths worldwide, and most patients present at an advanced stage. Computed tomography (CT) is the primary modality for lung cancer screening and characterisation; however, manual interpretation varies and is increasingly burdened by workload. Deep learning offers automated, high-accuracy classification. This systematic review synthesises evidence on deep learning-based lung cancer classification using CT images, focusing on morphological features, model architectures, and classification performance.

Evidence Acquisition: This review followed the PRISMA 2020 guidelines. Four databases were searched on 5 July 2026, covering 2014 to 2026: Semantic Scholar via Elicit (n = 1,000), PubMed/MEDLINE (n = 1,204), IEEE Xplore (n = 421), and Scopus (n = 3,781). Eligible studies were peer-reviewed articles that applied deep learning with explicit morphological features (shape, margin, texture, density) to CT-based classification, used human CT datasets with a minimum of 50 CT examinations (the smallest included cohort comprised 96 patients), and reported at least one quantitative metric. Records were screened in Elicit and then independently double-checked by two human reviewers, the first author (M.T.) and the second author (S.H.Y.), at both the abstract- and full-text screening stages and during data extraction. Conference proceedings and preprints were excluded.

Results: Twenty-seven studies met all eligibility criteria. All included models used convolutional neural networks; ResNet or 3D-ResNet backbones were the most common (10 studies), whereas transformer or state-space components appeared in four recent (2024 to 2026) studies. Morphological information was handcrafted or radiomics-derived in 14 studies and learned end-to-end in 13; the most frequently used feature groups were texture, shape and margin signs (spiculation, lobulation), density, and lesion size. Private datasets predominated (approximately two-thirds of studies); LIDC-IDRI or LUNA16 were used in roughly a quarter. Accuracy ranged from 66.3% to 99.9% (median 89.7%; 25 studies), and AUC ranged from 0.71 to 0.99 (median 0.92; 21 studies). Because this accuracy range combines binary and multiclass tasks with different chance levels and class balance, performance is reported by task rather than as a single figure, and sensitivity and specificity are prioritized over accuracy where studies allow. Only 10 studies (37%) reported external validation, and 8 (30%) treated metastatic (secondary) lesions as a distinct class. Discussion: CNN-based and hybrid architectures that combine learned representations with explicit morphological descriptors demonstrate strong classification performance on CT; however, this performance is almost always reported in small, predominantly single-centre datasets with limited external validation. Key gaps include the paucity of models that distinguish primary from metastatic disease, inconsistent reporting of performance metrics, and heterogeneous reference standards.

Keywords: Lung Neoplasms, Deep Learning, Computed Tomography, Convolutional Neural Networks, Computer-Assisted Diagnosis, Computer-Assisted Radiographic Image Interpretation

1. Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide. GLOBOCAN 2020 attributed approximately 2.21 million new cases and 1.80 million deaths to lung cancer annually, representing 12.4% of new cancer diagnoses and 18.0% of all cancer deaths (1). Five-year survival remains below 20% in most settings, largely because the disease is typically detected at a late stage (2). Management therefore depends on accurate classification. Three distinctions are particularly important: Benign versus malignant nodules; histological subtypes of non-small cell lung cancer versus small cell lung cancer; and primary tumours versus secondary deposits. The last distinction is easily overlooked. The lung is a common site of metastasis, and metastatic lesions carry different prognoses and treatment pathways; however, on CT they can closely resemble a primary tumour (2).

Computed tomography (CT) is the standard imaging modality for lung cancer screening, characterisation, and staging. The National Lung Screening Trial showed

How to Cite: Taherian M, Yaghoubyan SH, Bagherifard K, Malekhosseini R. Deep Learning-Based Lung Cancer Classification from CT Images: A Systematic Review of Morphological Features, Model Architectures, and Classification Performance. J Clin Res Paramed Sci. 2026;15(1):e173819. doi: https://doi.org/10.5812/jcrps-173819

that low-dose CT reduces lung cancer mortality by approximately 20% compared with chest radiography (3). Radiologists assess malignancy largely on the basis of nodule morphology. Size, shape, margin (spiculation or lobulation), attenuation (solid, part-solid, or ground-glass), and internal texture all contribute, along with signs such as pleural indentation and vascular convergence. These same features underpin risk models used in routine practice. However, visual interpretation is time-consuming, varies among observers, and is under increasing pressure owing to rising screening volumes. Subtle findings in nodules below 6 mm are easy to miss. Reproducible automated tools are therefore needed to support, not replace, radiological judgement. In practice, such tools could be used at multiple points in care, from triage in low-dose CT screening, through characterisation of incidentally detected nodules, to pre-treatment staging; the tolerable error differs across these settings because a missed cancer at screening has far higher costs than a false alarm during work-up.

Two computational approaches have been applied to this problem. Radiomics extracts predefined descriptors of shape, texture, and intensity from a segmented lesion and inputs them into a classifier. These features remain interpretable but depend on manual feature design. Deep learning, in contrast, learns features directly from images and often outperforms handcrafted pipelines. Convolutional neural networks (CNNs) such as ResNet, DenseNet, Inception, and EfficientNet have been widely adapted to CT-based lung cancer classification through transfer learning. More recent work uses vision transformers and hybrid CNN-attention designs to capture longer-range spatial context (4). An increasing body of research deliberately combines explicit morphological features with learned representations, aiming to retain the accuracy of deep models without sacrificing the interpretability required for clinical adoption. Beyond lung imaging, neural network methods have been applied to tumour localisation and segmentation in other modalities, including noise-robust approaches to lesion detection in MRI, and machine learning is increasingly used across respiratory and diagnostic medicine more broadly (5, 6).

Despite a large and rapidly growing literature, studies vary widely in datasets, model designs, reference standards, and reported metrics. Consequently, it is difficult to determine which approaches perform best and how well they generalise. Public benchmarks such as LIDC-IDRI are frequently used in technical papers, whereas clinically grounded studies tend to rely on private single-centre cohorts. Few reviews have focused specifically on how morphological features are represented and integrated, or on the more challenging task of distinguishing primary disease from metastatic disease. This systematic review is the first phase of a doctoral thesis titled 'Providing Lung Cancer Classification Using Morphological Features Based on Deep Learning Pattern'. Conducted under PRISMA 2020 (7), it addresses one question: what is the classification performance of deep learning models that use morphological features on CT images to distinguish metastatic pulmonary lesions from primary lung cancer and benign lesions? Four objectives follow: First, to characterise the architectures and classification strategies used; second, to summarise the morphological features and preprocessing techniques employed; third, to synthesise reported performance and methods of model validation; and fourth, to identify methodological gaps in separating primary from metastatic lesions, so that later phases of the thesis can address them.

2. Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (7).

2.1. Eligibility Criteria

Studies were eligible if they were peer-reviewed journal articles that met all of the following conditions. They had to use CT imaging as the primary modality and apply at least one deep learning method, including a CNN, ResNet, DenseNet, ViT, or hybrid architecture. They had to use or explicitly report morphological features, including shape, margin, texture, or density. Explicit morphological features were defined as named and reported descriptors rather than implicit features, and they occurred in three distinct forms: radiologist-scored semantic signs such as a spiculation or lobulation grade; quantitative radiomics matrices such as GLCM texture or wavelet coefficients; and learned representations made interpretable through attention or saliency maps. Studies had to address a lung cancer classification task, including malignancy (benign versus malignant), histological subtype, invasiveness, staging, or nodule characterisation. Consistent with the review's focus on the benign–primary–metastatic distinction, studies that modelled a distinct metastatic pulmonary class were eligible and were examined as a subgroup (8 of the 27 included studies); however, the presence of a metastatic class was not itself an inclusion criterion. Datasets had to comprise human CT series; a minimum

of 50 CT examinations was required, and because patient-level and image-level counts are not interchangeable, the threshold was applied at the study level, with the smallest included cohort comprising 96 patients. Finally, articles had to be published in English from 2014 onwards and report at least one quantitative performance metric.

2.2. Search Strategy

The search was deliberately broad to maximise sensitivity. The narrower scope of the review—peer-reviewed journal articles that report explicit morphological features and address a lung cancer classification task—was applied through the eligibility criteria during screening rather than through the search string. Therefore, no re-searching was required.

2.3. Study Selection and Data Extraction

Screening proceeded in two stages. At each stage, the two human reviewers were the first author (M.T.) and the second author (S.H.Y.), who worked independently of each other and of the AI. At the title-and-abstract stage, the Elicit AI platform scored records against the predefined eligibility criteria; these scores were treated as soft recommendations to preserve sensitivity. Both reviewers (M.T. and S.H.Y.) then independently checked every record rather than relying solely on the AI ranking. Records that passed underwent independent full-text screening, with both authors screening each paper separately. Disagreements at either stage were resolved by consensus, with a third reviewer available for arbitration. When Elicit lacked a full text, the paper was screened manually by both reviewers. For data extraction, Elicit generated an initial structured extraction for each study, but both reviewers (M.T. and S.H.Y.) independently verified every field against the full text and corrected the automated values when they were incorrect, incomplete, or ambiguous. Thus, every screening and extraction step was double-checked by both human reviewers. Extracted items included study characteristics, imaging datasets, morphological features, preprocessing pipelines, deep learning architectures, classification tasks, evaluation metrics, and performance measures.

2.4. Quality Assessment

Two reviewers (M.T. and S.H.Y.) independently assessed the methodological quality and risk of bias of all 27 included studies; disagreements were resolved by consensus, with a third reviewer available for arbitration. Two complementary instruments were applied. Risk of bias and applicability were evaluated using QUADAS-2 (8) across its four domains—patient selection, index test (the deep learning classifier), reference standard, and flow and timing—with each domain rated as low, high, or unclear. Reporting completeness and adherence to good practice for AI diagnostic studies were assessed using the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) (9). The appraisal followed the CLAIM 2024 update (10), which is now the reference standard for AI medical-imaging studies, and the core items scored here are consistent with it. For each study, we recorded adherence to core CLAIM items covering data source and eligibility, definition and adequacy of the reference standard, data partitioning, model and training description, performance metrics and their reported uncertainty, internal and external validation, and comparison with clinicians, and expressed adherence as the percentage of applicable items met. Because most included studies were diagnostic prediction-model studies rather than classical diagnostic-accuracy studies, the QUADAS-2 signalling questions were interpreted with attention to machine-learning-specific sources of bias, particularly data leakage arising from image- or slice-level rather than patient-level partitioning and the presence or absence of a genuinely held-out or external test set. For the same reason, risk of bias was additionally considered against the domains of PROBAST (11), the instrument designed for prediction-model studies, which reinforced the patient-selection and analysis concerns identified by QUADAS-2. Domain-level judgements for every study are reported in Tables A2 in the Supplementary File (QUADAS-2) and A3 (CLAIM).

3. Results

3.1. Study Characteristics and Publication Trends

The database searches yielded 6,406 records. After removal of 1,413 duplicates, 4,993 records remained for screening. Of these, 47 reports were assessed in full text, and 27 met all eligibility criteria and were included in the review (12-38). The included studies were published between 2018 and 2026, and 16 (59%) were published between 2024 and 2026. This weighting toward recent years reflects the maturation of deep learning for thoracic CT rather than any single benchmark or challenge (Figure 1.

3.2. Imaging Datasets

Dataset selection was dominated by private, single- or multi-institution CT cohorts, used in approximately

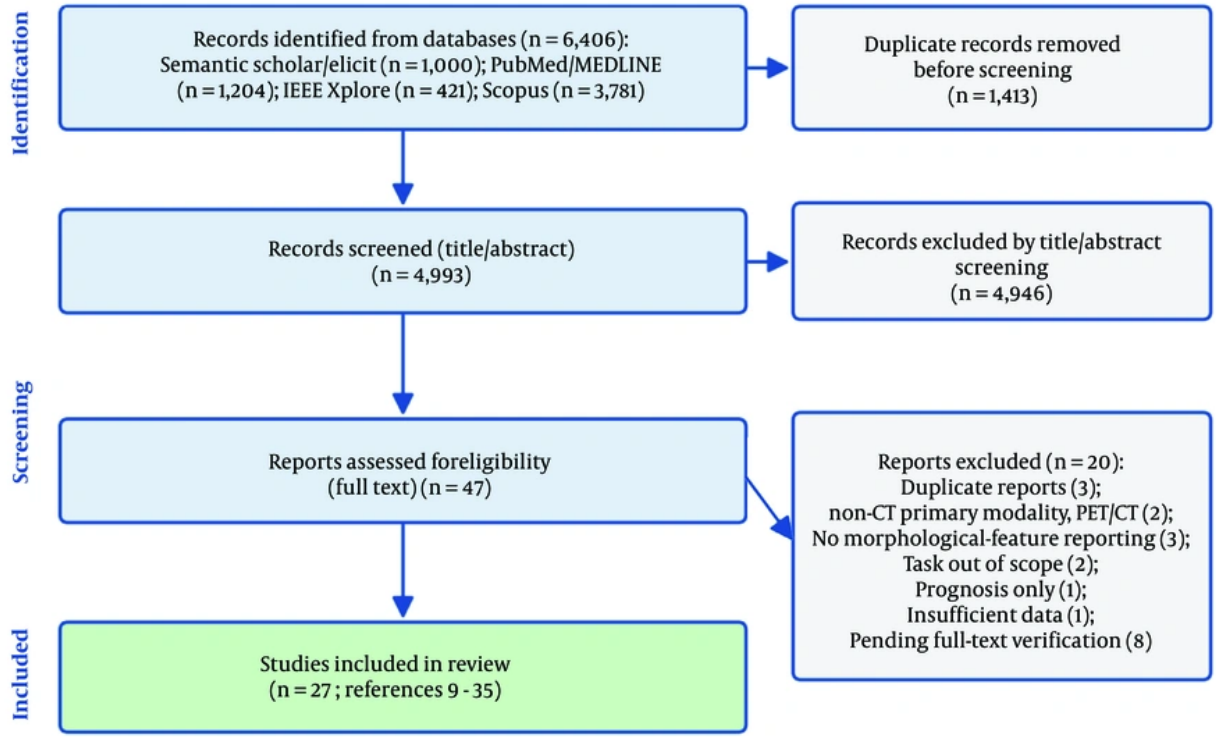

Figure 1. PRISMA 2020 flow diagram of the study selection process, from 6,406 records retrieved to 27 studies included.

two-thirds of studies (about 18 of 27), with histopathology or clinical follow-up as the reference standard. Public benchmarks, chiefly LIDC-IDRI and LUNA16, and occasionally NLST or the Kaggle Data Science Bowl 2017, appeared in approximately one-quarter to one-third of studies, sometimes alongside private data. Reported sample sizes ranged from 96 patients to several thousand lesions or images (median around 470). External validation on an independent cohort was reported in only 10 of the 27 studies (37%). Reference standards were heterogeneous: histopathology confirmed the diagnosis in 21 studies, whereas 6 relied on radiologist-defined labels or composite imaging follow-up, which are softer endpoints; therefore, these two types of ground truth should not be pooled uncritically. Scanner vendor, acquisition and reconstruction settings, and nodule annotation procedures were reported only sporadically, limiting judgement of model transferability. Most cohorts were retrospective and single-centre and may not reflect a consecutive screening or incidental-nodule population. Moreover, in several of the ten externally validated studies, the external set came from a different scanner within the same institution rather than a genuinely independent multicentre cohort; thus, external validation is weaker than the 37% figure alone suggests. Per-study reference standards, settings, and validation type are provided in Table A4 in the Supplementary File.

3.3. Morphological Features

Morphological features were incorporated into models via two approximately equal approaches. Fourteen studies used handcrafted or radiomics-derived descriptors, often extracted using PyRadiomics and then fused with deep features. The other 13 learned features end-to-end, and several used attention or multi-view designs to expose the morphological basis of the prediction. Texture descriptors (GLCM, LBP, wavelet) were the most common group. Shape and margin signs were also frequently included, such as spiculation, lobulation, pleural indentation, and sphericity, together with density (ground-glass versus solid) and lesion size.

Table 1. Deep Learning Architectures Used Across the 27 Included Studies (Categories Are Not Mutually Exclusive) [a]

Architecture Category	Studies (n)	% of 27	Notes
CNN-based (all variants)	27	100	A CNN component in every included study
ResNet / 3D-ResNet backbone	10	37	Most common single backbone
DenseNet	3	11	
VGG	2	7	
Hybrid (CNN + radiomics / GNN / attention)	13	48	Non-exclusive with CNN rows
Transformer / ViT / state-space	4	15	2024 - 2026; +1 CNN–transformer hybrid
Graph neural network (GNN)	2	7	
3D volumetric input	13	48	Per-study input dimensionality (Table A4)
2D slice input	9	33	Remaining 5 multi-view or not stated
ImageNet transfer learning	5	19	Other pretraining sources in ~7 more
Public code or model weights released	0	0	Not reported in any included study

[a] Studies may report more than one architecture; counts are not mutually exclusive. Abbreviation: NR, not reported.

Table 2. Reported Classification Performance Across the 27 Included Studies [a]

Metric	Studies (n)	Range	Mean	Median
Accuracy (%)	25	66.3 - 99.9	87.1	89.7
AUC-ROC	21	0.71 - 0.99	0.89	0.92
Sensitivity (%)	most studies	not pooled	–	–
Specificity (%)	most studies	not pooled	–	–
F1-score (%)	some studies	not pooled	–	–

[a] Accuracy and AUC summarise the studies reporting each metric among the 27 included; values were verified against the full text by both reviewers (M.T. and S.H.Y.). Sensitivity, specificity, and F1 were reported too heterogeneously to pool.

3.4. Deep Learning Architectures

All included models used a convolutional neural network in some form. ResNet and 3D-ResNet backbones were the most common single choice, appearing in about 10 studies, with DenseNet and VGG used in a handful more. Approximately half the studies used hybrid designs combining a CNN with handcrafted radiomics, a graph neural network, or an attention module. Transformer components remained a minority. Four studies published between 2024 and 2026 used a vision transformer, a multi-scale transformer, or a state-space (Mamba-style) module, and one additional study used a CNN-transformer hybrid. Graph neural networks were used in two studies. A recent survey indicates that the broader field is shifting toward transformer architectures (39), suggesting that their relative scarcity here may reflect the 2014 to 2026 window and journal-only scope rather than the state of the art. Inputs were three-dimensional volumetric in 13 studies and two-dimensional slice-based in 9, with the remaining 5 using multi-view stacks or not stating the approach. Some form of transfer learning was used in about half the studies, but ImageNet pretraining specifically appeared in only 5 (Nishio, Paul, El-Bana, He, and Liufu); given the domain gap between natural photographs and CT, this scarcity is noteworthy rather than assumed. No included study released public code or trained model weights, and hyperparameters or training hardware were frequently omitted; this lack of reporting is itself a limitation. Per-study input dimensionality, transfer learning, and code availability are listed in Table A4 in the Supplementary File (Table 1).

3.5. Classification Performance

Because classification tasks differed in the number of classes and class balance, accuracy is not comparable across studies and is not pooled into a single estimate; sensitivity and specificity, which distinguish missed cancers from false alarms, are prioritised where reported. Performance varied by task. For binary benign-versus-malignant nodule classification (about 14 studies; chance level 50%), accuracy clustered at high levels, ranging from approximately 74% to 99.9%, with

sensitivity and specificity, when reported, mostly in the high 80s to high 90s. For binary tasks framed around metastatic disease (primary or metastatic versus benign; 5 studies), accuracy ranged from about 85% to 96% and AUC from 0.78 to 0.95. Multiclass tasks that separated benign, primary, and metastatic lesions, or used a four-way malignancy scheme (chance level 33% or 25%), showed lower and more variable performance, ranging from 66.3% to 97.1%; studies of adenocarcinoma subtyping and ground-glass invasiveness were among the weakest (66.3% and 76.9%). Reported class balance was frequently skewed toward benign lesions or a single dominant class, which can inflate accuracy for more common categories; per-study task, class composition, and metrics are provided in Table A4 in the Supplementary File. The pooled ranges provided for context only—accuracy 66.3 - 99.9% (25 studies) and AUC 0.71 - 0.99 (21 studies)—should be interpreted as a coarse envelope rather than a like-for-like comparison. Given heterogeneity in tasks, datasets, and reference standards, results are synthesised narratively rather than pooled (Table 2).

3.6. Quality Assessment

Risk of bias was substantial across the evidence base. Under QUADAS-2, 20 of the 27 studies (74%) were at high overall risk of bias and the remaining 7 (26%) at unclear risk; none was at low risk across all four domains. Patient selection was the weakest domain: no study clearly met the criterion for an unbiased consecutive or random sample (11 high, 16 unclear), reflecting the predominance of retrospective, single-centre cohorts and selected populations. Risk of bias in the index test was high in 11 studies (41%), driven mainly by cross-validation without a held-out test set and, in several cases, image- or slice-level partitioning that risks patient-level data leakage. Reference-standard bias was high in 8 studies (30%) and unclear in 7, most often when labels were derived from radiologist annotation rather than tissue confirmation or from composite imaging-based endpoints. Flow and timing were frequently unclear (13 studies) because partitioning and case flow were incompletely described. Applicability concerns were concentrated in patient selection: 16 of 27 studies (59%) drew on narrow or non-representative populations (e.g., single-disease, subtype-only, or endemic-region cohorts), and 8 studies raised applicability concerns regarding the reference standard because radiologist-defined labels do not correspond to confirmed malignancy.

CLAIM adherence was moderate to good overall (mean 81% of core items; 23 studies rated high adherence and 4 moderate), but gaps clustered in areas critical for reproducibility and clinical translation. Only 14 of 27 studies (52%) clearly reported train/validation/test partitioning, only 18 (67%) reported any uncertainty (confidence interval or standard error) around headline metrics, only 11 (41%) compared model performance with that of radiologists, and only 10 (37%) included any external or independent-cohort validation. Histopathology served as the reference standard in 21 studies (78%), with the remainder relying on radiologist labels or composite endpoints. Taken together, these two instruments indicate that although reporting of architectures and headline metrics is generally adequate, the principal weaknesses are limited external validation, incomplete reporting of data partitioning with attendant leakage risk, heterogeneous reference standards, and narrow study populations. These limitations temper confidence in the high reported accuracies and should be considered in any pooled interpretation. Full per-study judgements are provided in Tables A2 and A3 in the Supplementary File (Table 3).

4. Discussion

4.1. Principal Findings

This review synthesised 27 studies that met a strict requirement for explicit morphological-feature reporting, and three findings stand out. First, CNN-based architectures, particularly ResNet and 3D-ResNet backbones (although ImageNet transfer learning was used in only five studies), achieve high accuracy, with a median of 89.7% and a median AUC of 0.92. However, this accuracy is demonstrated predominantly on small, mostly private datasets. Second, models that combine explicit morphological descriptors with deep representations perform well and are more interpretable than purely end-to-end pipelines. This supports the thesis that morphological priors complement learnt features. Third, the clinically most important tasks are also the most difficult and the least studied. Separating primary from metastatic disease and sub-typing adenocarcinoma both fall into this group, and both show the lowest performance.

Despite these headline accuracies, several structural weaknesses recur. Most studies rely on private, single-institution data. External validation appears in only about one-third of the studies. Metric reporting is inconsistent, and heavy reliance on pooled accuracy across tasks with different chance levels can be misleading; consensus guidance on task-appropriate metric selection now exists and should be followed (40). Reference standards are heterogeneous. Collectively,

Table 3. Summary of Methodological Quality and Risk of Bias Across the 27 Included Studies (QUADAS-2 Domains and Applicability) [a]

QUADAS-2 Risk of Bias	Low	High	Unclear
Patient selection	0 (0)	11 (41)	16 (59)
Index test (DL classifier)	9 (33)	11 (41)	7 (26)
Reference standard	12 (44)	8 (30)	7 (26)
Flow and timing	8 (30)	6 (22)	13 (48)
Overall risk of bias	0 (0)	20 (74)	7 (26)
QUADAS-2 applicability	Low n ()	High n ()	Unclear n ()
Patient selection	11 (41)	16 (59)	0 (0)
Index test	27 (100)	0 (0)	0 (0)
Reference standard	13 (48)	8 (30)	6 (22)

[a] Values are expressed as No. (%). Judgements were made independently by two reviewers (M.T. and S.H.Y.) with consensus resolution. Percentages are of the 27 included studies. CLAIM adherence results and full per-study judgements are reported in the text above and in Tables A2 and A3 in the Supplementary File.

these issues complicate cross-study comparisons and may inflate the apparent state of the art.

4.2. Comparison with Prior Reviews

Previous systematic reviews in this domain have typically been narrower in scope, focusing on specific datasets, architecture families, or time periods predating 2022, and have rarely required explicit morphological-feature reporting. By making this requirement central, this review highlights how morphology is represented and integrated in CT-based classification models, a dimension that prior syntheses have largely left implicit.

4.3. Implications for Future Research

High benchmark accuracy does not guarantee clinical utility. Achieving clinical utility requires external validation on multi-centre, multi-scanner, prospective datasets. For the next phases of this thesis, the evidence supports three design choices. The first is a CNN backbone, either ResNet or EfficientNet, with ImageNet pretraining as the deep feature extractor. The second is the explicit integration of texture, shape, and density features through a feature fusion branch. The third is evaluation on LIDC-IDRI together with at least one additional institutional dataset, to enable assessment of generalisation. Broader experience with introducing artificial intelligence into clinical and educational practice shows that implementation is rarely straightforward (41), and adoption ultimately depends on demonstrated usefulness to end users, as reflected in the uptake of other digital-health services (42).

Transformer, state-space, and graph-based components have appeared mainly in studies from 2024 to 2026. Their emergence suggests growing interest in modelling long-range spatial context and the relationship between a lesion and its surroundings. Nevertheless, CNNs remain the backbone of the field, and evidence of a clear transformer advantage in this setting remains limited.

4.4. Strengths and Limitations

The review has clear strengths. The search spanned biomedical, engineering, and interdisciplinary databases; the first and second authors (M.T. and S.H.Y.) independently screened studies at both the abstract and full-text stages; and every extracted item was double-checked field by field against the full text by both reviewers. The limitations must be stated plainly. First, the final evidence base is small: the strict eligibility criteria, and above all the requirement for explicit morphological-feature reporting, reduced a large initial corpus to 27 studies. Accordingly, the synthesis is narrative rather than quantitative, and the review is restricted to English-language publications. Second, a more fundamental limitation concerns the primary literature itself. The evidence base is dominated by private, single-institution data; two-thirds of the included studies (18 of 27) used private institutional datasets, most of them single-centre, and only 10 of 27 (37%) reported any external or independent-cohort validation. This is not a shortcoming of the review but of the field it summarises, and it is a serious one: a model fitted to one institution's patients, scanners, and acquisition protocols is subject to spectrum bias and may not transfer to other settings. Therefore, the high accuracies reported here are best interpreted as within-cohort performance, with generalisability remaining largely untested. We highlight this limitation explicitly because it, more than any headline metric, determines

how far these tools are from clinical deployment. A further consequence of restricting inclusion to peer-reviewed journal articles is the exclusion of conference and preprint venues, which may undercount transformer-based work that is expanding rapidly in medical imaging (39).

4.5. Concluding Remarks

This review of 27 studies shows that deep learning is a capable approach to CT-based lung cancer classification. The field is overwhelmingly CNN-based, increasingly hybrid, and beginning to adopt transformers. Among the studies reporting them, median accuracy was near 90% and median AUC above 0.9. Models that combined learned representations with explicit morphological descriptors performed particularly well and were easier to interpret.

Four priorities stand out for future work: larger multi-centre cohorts with shared test sets, routine external validation, standardised, task-appropriate metric reporting (40), and explicit classification of primary versus metastatic disease. These changes are most likely to move the field from strong benchmark performance toward genuine clinical usefulness. The formal quality appraisal reported here (QUADAS-2 and CLAIM) confirms that these gaps, rather than model design, are the primary constraints on the field.

Supplementary Material

Supplementary material(s) is available here [To read supplementary materials, please refer to the journal website and open PDF/HTML].

Footnotes

AI Use Disclosure: The authors declare that no generative AI tools were used in the creation of this article.

Authors' Contribution: Study concept and design: M. T., S. H. Y., and K. B.; Acquisition of data: M. T. and R. M.; Analysis and interpretation of data: M. T. and S. H. Y.; Drafting of the manuscript: M. T., S. H. Y., K. B., and R. M.; Critical revision of the manuscript for important intellectual content: M. T., S. H. Y., K. B., and R. M.; Administrative, technical, and material support: M. T., K. B., and R. M.; Study supervision: M. T. and S. H. Y.

Conflict of Interests Statement: The authors declare that they have no conflict of interest.

Data Availability: All data analysed in this review are drawn from the published studies cited here. The search strategies are included in the appendices, and the data extraction records are available from the corresponding author on reasonable request.

Funding/Support: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

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):209-249. [PubMed ID: 33538338]. https://doi.org/10.3322/caac.21660.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin*. 2023;**73**(1):17-48. [PubMed ID: 36633525]. [PubMed Central ID: PMC12559696]. https://doi.org/10.3322/caac.21763.
3. National Lung Screening Trial Research Team; Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;**365**(5):395-409. [PubMed ID: 21714641]. [PubMed Central ID: PMC4356534]. https://doi.org/10.1056/NEJMoa1102873.
4. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2010.11929.
5. Norouzi S, Sistani S, Khoshkhui M, Faridhosseini R, Payandeh P, Ghasemian F, et al. Exploring common symptoms in patients with respiratory allergies using K-means algorithm in the north-east of Iran in 2012 - 2015. *Tanaffos*. 2023;**22**(1):120-128. [PubMed ID: 37920309]. [PubMed Central ID: PMC10618594].
6. Tajvidi Asr R, Rahimi M, Hossein Pourasad M, Zayer S, Momenzadeh M, Ghaderzadeh M. Hematology and hematopathology insights powered by machine learning: shaping the future of blood disorder management. *Iran J Blood Cancer*. 2024;**16**(4):9-19. https://doi.org/10.61186/ijbc.16.4.9.
7. 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:n71. [PubMed ID: 33782057]. [PubMed Central ID: PMC8005924]. https://doi.org/10.1136/bmj.n71.
8. Whiting PF, Rutjes AWS, 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):529-536. [PubMed ID: 22007046]. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
9. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;**2**(2):e200029. [PubMed ID: 33937821]. [PubMed Central ID: PMC8017414]. https://doi.org/10.1148/ryai.2020200029.
10. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;**6**(4). e240300. [PubMed ID: 38809149]. [PubMed Central ID: PMC11304031]. https://doi.org/10.1148/ryai.240300.
11. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;**170**(1):51-58. [PubMed ID: 30596875]. https://doi.org/10.7326/M18-1376.

12. Nishio M, Sugiyama O, Yakami M, Ueno S, Kubo T, Kuroda T, et al. Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. *PLoS One*. 2018;**13**(7):e0200721. [PubMed ID: 30052644]. [PubMed Central ID: PMC6063408]. https://doi.org/10.1371/journal.pone.0200721.
13. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. *J Med Imaging (Bellingham)*. 2018;**5**(1):1. [PubMed ID: 29594181]. [PubMed Central ID: PMC5862127]. https://doi.org/10.1117/1.JMI.5.1.011021.
14. Zhang G, Yang Z, Gong L, Jiang S, Wang L. Classification of benign and malignant lung nodules from CT images based on hybrid features. *Phys Med Biol*. 2019;**64**(12):125011. [PubMed ID: 31141794]. https://doi.org/10.1088/1361-6560/ab2544.
15. EL-Bana S, Al-Kabbany A, Sharkas M. A two-stage framework for automated malignant pulmonary nodule detection in CT scans. *Diagnostics (Basel)*. 2020;**10**(3):131. [PubMed ID: 32121281]. [PubMed Central ID: PMC7151085]. https://doi.org/10.3390/diagnostics10030131.
16. Venugopal VK, Vaidhya K, Murugavel M, Chunduru A, Mahajan V, Vaidya S, et al. Unboxing AI: radiological insights into a deep neural network for lung nodule characterization. *Acad Radiol*. 2020;**27**(1):88-95. [PubMed ID: 31623996]. https://doi.org/10.1016/j.acra.2019.09.015.
17. Lin X, Jiao H, Pang Z, Chen H, Wu W, Wang X, et al. Lung cancer and granuloma identification using a deep learning model to extract 3-dimensional radiomics features in CT imaging. *Clin Lung Cancer*. 2021;**22**(5):e756-e766. [PubMed ID: 33678583]. https://doi.org/10.1016/j.cllc.2021.02.004.
18. Li K, Liu K, Zhong Y, Liang M, Qin P, Li H, et al. Assessing the predictive accuracy of lung cancer, metastases, and benign lesions using an artificial intelligence-driven computer-aided diagnosis system. *Quant Imaging Med Surg*. 2021;**11**(8):3629-3642. [PubMed ID: 34341737]. [PubMed Central ID: PMC8245931]. https://doi.org/10.21037/qims-20-1314.
19. Wang C, Shao J, Xu X, Yi L, Wang G, Bai C, et al. DeepLN: a multi-task AI tool to predict the imaging characteristics, malignancy and pathological subtypes in CT-detected pulmonary nodules. *Front Oncol*. 2022;**12**. 683792. [PubMed ID: 35646699]. [PubMed Central ID: PMC9130467]. https://doi.org/10.3389/fonc.2022.683792.
20. Li L, Zhou X, Cui W, Li Y, Liu T, Yuan G, et al. Combining radiomics and deep learning features of intra-tumoral and peri-tumoral regions for the classification of breast cancer lung metastasis and primary lung cancer with low-dose CT. *J Cancer Res Clin Oncol*. 2023;**149**(17):15469-15478. [PubMed ID: 37642722]. [PubMed Central ID: PMC11797261]. https://doi.org/10.1007/s00432-023-05329-2.
21. Li R, Zhou L, Wang Y, Shan F, Chen X, Liu L. A graph neural network model for the diagnosis of lung adenocarcinoma based on multimodal features and an edge-generation network. *Quant Imaging Med Surg*. 2023;**13**(8):5333-5348. [PubMed ID: 37581061]. [PubMed Central ID: PMC10423350]. https://doi.org/10.21037/qims-23-2.
22. Yang X, Chu XP, Huang S, Xiao Y, Li D, Su X, et al. A novel image deep learning-based sub-centimeter pulmonary nodule management algorithm to expedite resection of the malignant and avoid over-diagnosis of the benign. *Eur Radiol*. 2023;**34**(3):2048-2061. [PubMed ID: 37658883]. https://doi.org/10.1007/s00330-023-10026-2.
23. Liu Y, Ren H, Pei Y, Shen L, Guo J, Zhou J, et al. Development of a CT-based comprehensive model combining clinical, radiomics with deep learning for differentiating pulmonary metastases from noncalcified pulmonary hamartomas: a retrospective cohort study. *Int J Surg*. 2024;**110**(8):4900-4910. [PubMed ID: 38759692]. [PubMed Central ID: PMC11326030]. https://doi.org/10.1097/JS9.0000000000001593.
24. Gao S, Xu Z, Kang W, Lv X, Chu N, Xu S, et al. Artificial intelligence-driven computer-aided diagnosis system provides similar diagnosis value compared with doctors' evaluation in lung cancer screening. *BMC Med Imaging*. 2024;**24**(1). 141. [PubMed ID: 38862884]. [PubMed Central ID: PMC11165751]. https://doi.org/10.1186/s12880-024-01288-3.
25. Huang J, Xie S, Huang J, Zheng Z, Lin Z, Lin J, et al. Imaging features and deep learning for prediction of pulmonary epithelioid hemangioendothelioma in CT images. *J Thorac Dis*. 2024;**16**(2):935-947. [PubMed ID: 38505025]. [PubMed Central ID: PMC10944745]. https://doi.org/10.21037/jtd-23-455.
26. Safta W, Shaffie A. Advancing pulmonary nodule diagnosis by integrating engineered and deep features extracted from CT scans. *Algorithms*. 2024;**17**(4):161. https://doi.org/10.3390/a17040161.
27. Song X, Duan X, He X, Wang Y, Li K, Deng B, et al. Computer-aided diagnosis of distal metastasis in non-small cell lung cancer by low-dose CT based radiomics and deep learning signatures. *Radiol Med*. 2024;**129**(2):239-251. [PubMed ID: 38214839]. https://doi.org/10.1007/s11547-024-01770-6.
28. Esha JF, Islam T, Pranto MAM, Borno AS, Faruqui N, Yousuf MA, et al. Multi-view soft attention-based model for the classification of lung cancer-associated disabilities. *Diagnostics (Basel)*. 2024;**14**(20):2282. [PubMed ID: 39451604]. [PubMed Central ID: PMC11506595]. https://doi.org/10.3390/diagnostics14202282.
29. Xiao D, Forero Y, Kammer MN, Chen H, Paez R, Heideman BE, et al. Radiomic 'stress test': exploration of a deep learning radiomic model in a high-risk prospective lung nodule cohort. *BMJ Open Respir Res*. 2025;**12**(1):e002687. [PubMed ID: 40579208]. [PubMed Central ID: PMC12207176]. https://doi.org/10.1136/bmjresp-2024-002687.
30. Gong W, Cui Q, Fu S, Wu Y. Application of contrast-enhanced CT-driven multimodal machine learning models for pulmonary metastasis prediction in head and neck adenoid cystic carcinoma. *Eur J Radiol*. 2025;**192**. 112377. [PubMed ID: 40857998]. https://doi.org/10.1016/j.ejrad.2025.112377.
31. Piskorski L, Debic M, von Stackelberg O, Schlamp K, Welzel L, Weinheimer O, et al. Malignancy risk stratification for pulmonary nodules: comparing a deep learning approach to multiparametric statistical models in different disease groups. *Eur Radiol*. 2025;**35**(7):3812-3822. [PubMed ID: 39747589]. [PubMed Central ID: PMC12165889]. https://doi.org/10.1007/s00330-024-11256-8.
32. Zhao X, Li J, Qi M, Chen X, Chen W, Li Y, et al. MSTD: a multi-scale transformer-based method to diagnose benign and malignant lung nodules. *IEEE Access*. 2025;**13**:16182-16195. https://doi.org/10.1109/ACCESS.2025.3531001.
33. Huang Y, Li Q, Chen H, Zhou X, Han Q, Lu H, et al. A novel approach to integrating vision transformers and machine learning for robust lung nodule classification using CT imaging. *J Radiat Res Appl Sci*. 2025;**18**(3). 101672. https://doi.org/10.1016/j.jrras.2025.101672.
34. Gao X, Ma X, Zhang Z, Yuan J, Li Q, Zheng P, et al. Differentiating sub-centimeter lung metastases in colorectal cancer by deep learning: a multicenter retrospective study. *BMC Med Imaging*. 2026. [PubMed ID: 42316047]. https://doi.org/10.1186/s12880-026-02508-8.
35. He L, Li Z, Duan Y, Gao Y, Zhan Y, Du Y, et al. Prediction of infiltration degree of ground-glass nodules using a fusion of CT radiomics and deep learning. *Sci Rep*. 2026;**16**(1). 19153. [PubMed ID: 42034781]. [PubMed Central ID: PMC13279961]. https://doi.org/10.1038/s41598-026-50328-1.
36. Nady G, Salem A, Badawy O, Abo-ElNour S. Explainable active reinforcement deep learning improves lung cancer detection from CT images. *Sci Rep*. 2026;**16**(1). 7510. [PubMed ID: 41735399]. [PubMed Central ID: PMC12932830]. https://doi.org/10.1038/s41598-026-38239-7.
37. Liufu Y, Su R, Wen Y, Guan Y, Mahmoud MA. A CT-based deep learning approach to distinguish multiple primary lung cancers, intrapulmonary metastases, and benign pulmonary lesions. *BMC*

Cancer. 2026;**26**(1). 177. [PubMed ID: 41485033]. [PubMed Central ID: PMC12870997]. https://doi.org/10.1186/s12885-025-15501-1.

38. Zhu Z, Hu G, Tan W, Gao K, Sun C, Zhou Z, et al. DeepFAN, a transformer-based model for human-artificial intelligence collaborative assessment of incidental pulmonary nodules in CT scans: a multireader, multicase trial. *Nat Cancer*. 2026;**7**(5):773-789. [PubMed ID: 42020549]. https://doi.org/10.1038/s43018-026-01147-w.
39. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. *Med Image Anal*. 2023;**88**. 102802. [PubMed ID: 37315483]. https://doi.org/10.1016/j.media.2023.102802.
40. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods*. 2024;**21**(2):195-212. [PubMed ID: 38347141]. [PubMed Central ID: PMC11182665]. https://doi.org/10.1038/s41592-023-02151-z.
41. Dehghani M, Pourasad MH, Khezri H. Challenges in implementing artificial intelligence for nursing education: a systematic review. *Educational Research in Medical Sciences*. 2025;**14**(1). https://doi.org/10.5812/ermsj-162371.
42. Galavi Z, Pourasad MH, Norouzi S, Jahani Y, Khajouei R. Public usage, perceived usefulness, and satisfaction with e-health services in the COVID-19 pandemic. *Journal of Clinical Research in Paramedical Sciences*. 2022;**11**(2). https://doi.org/10.5812/jcrps-133719.