Review Article

Evaluating the Accuracy and Clinical Impact of Deep Learning Algorithms for Automated Detection of Pulmonary Nodules in Chest CT Scans

¹Avinash

¹Assistant Professor, Department of Allied Health Sciences, Brainware University, Kolkata India

Corresponding Author: Avinash*

ABSTRACT

Early detection of pulmonary nodules is crucial for improving lung cancer prognosis. Deep learning (DL) algorithms have emerged as promising tools for automated nodule detection in chest CT scans, offering potential to enhance diagnostic accuracy and reduce radiologist workload. This review synthesizes current literature (2020–2025) on DL-based pulmonary nodule detection, focusing on algorithm performance, clinical impact, and implementation challenges. Across multiple studies, DL systems, particularly 3D convolutional neural networks, U-Net architectures, and hybrid models integrating detection and classification, demonstrated high sensitivity (79–95%) and variable false-positive rates, often matching or exceeding radiologist performance. DL as a second reader improved detection of small, subsolid, and subtle nodules, while segmentation modules facilitated volumetric assessment for follow-up and malignancy risk estimation. Despite these advances, limitations remain, including dataset heterogeneity, inconsistent evaluation metrics, elevated false positives in some models, and limited prospective evidence on clinical outcomes. Future research should emphasize multi-center validation, standardized reporting, workflow integration, and assessment of patient-level outcomes. With ongoing technological refinement and rigorous evaluation, DL algorithms hold the potential to become an integral component of chest CT interpretation, improving early lung cancer detection and optimizing clinical decision-making.

Keywords: Deep learning, Pulmonary nodules, Chest CT, Automated detection

INTRODUCTION

Lung cancer remains the leading cause of cancer-related mortality globally, accounting for substantial sickness and death each year [1,2]. Early detection of lung cancer often via identification of small pulmonary nodules on chest computed tomography (CT), is critical, as prognosis is strongly dependent on stage at diagnosis. However, conventional radiological evaluation of CT images has significant limitations. Radiologists must manually inspect volumetric datasets slice by slice, a burdensome, time-consuming task that is prone to human error and inter-observer variability. Small, subtle, or juxtapleural nodules especially risk being overlooked. Even experienced radiologists may miss lesions, particularly in high-volume screening or non-thoracic imaging practices where chest evaluation is secondary. Given these challenges, there has long been interest in computer-aided detection (CAD) to support radiologists. Over past decades, traditional machine-learning and rule-based CAD systems addressed pulmonary nodule detection, yet often suffered from limited sensitivity, high false-positive rates, and poor generalizability [3–5]. The advent of deep learning (DL) especially convolutional neural networks (CNNs), 3D-CNNs, and more recently hybrid and attention-based architectures has reignited optimism for reliable automated detection, segmentation, and classification of pulmonary nodules in CT images [6–9].

Review Article

In the last decade, considerable research effort has been directed towards DL-based systems for pulmonary nodule detection, segmentation, and malignancy risk estimation. A comprehensive survey by Wu & Qian summarized advances in detection, segmentation, and classification of nodules using DL techniques, underscoring the potential of such methods to overcome traditional CAD limitations [10]. Similarly, DL-based systems have demonstrated increased sensitivity compared to human readers in many studies including for small (< 5 mm) and subsolid nodules often with acceptable false-positive rates [11–14]. The appeal of DL-aided detection is not just technical performance, but clinical impact. Automated tools may serve as "second readers," flagging candidate nodules for radiologist review, thus potentially reducing missed diagnoses, improving early lung cancer detection, and standardizing reporting. In one large retrospective study assessing over 21,000 body CT studies (including chest), a DL system flagged pulmonary nodules missed in original radiology reports; after expert review, a significant subset of these flagged lesions were deemed clinically significant and would warrant follow-up [15]. Such findings highlight DL's promise to augment radiologic workflows and improve patient care.

Moreover, DL is starting to move beyond detection into classification and risk stratification. Recent DL-based diagnostic frameworks combine detection and classification potentially distinguishing benign from suspicious nodules, thus informing follow-up decisions, triage, and management [16–18]. A 2024 meta-analysis of DL-based malignancy-risk prediction models in CT-detected nodules found that DL systems offered higher sensitivity and comparable or better specificity compared to physician judgment or conventional risk models, with improved area under the receiver operating curve (AUC) [19]. Despite these promising developments, several challenges remain before DL-based nodule detection becomes standard practice. Many studies rely on publicly available datasets such as the LIDC-IDRI database, which may not adequately represent real-world clinical diversity in patient demographics, CT protocols, scanner types, or variations in nodule appearance [20–22]. As a result, generalizability of models across centers remains uncertain. Additionally, there is significant heterogeneity in study design: different studies use varying definitions (e.g., what constitutes a "true nodule"), disparate evaluation metrics (per-scan, per-nodule, false positives per scan, per-patient sensitivity), and inconsistent reporting of false positives complicating cross-study comparisons [10,13,22].

False positives remain a major concern: high sensitivity might be accompanied by an unacceptable number of false alarms, which could overwhelm radiologists, increase follow-up imaging, cause patient anxiety, and raise healthcare costs. Early 3D-DL systems reported up to ~30 false positives per scan during candidate generation, though subsequent false-positive reduction steps improved this [9,23]. Whether newer models especially hybrid and transformer-based can maintain high sensitivity while keeping false positives acceptably low, across diverse real-world datasets, remains an open question. Furthermore, clinical implementation brings additional hurdles: integration into radiology workflow, regulatory and medicolegal considerations, radiologist acceptance, data privacy, and the need for robust prospective, multicenter validation. There is also a paucity of longitudinal data demonstrating that DL-assisted detection actually leads to improved patient outcomes (e.g., earlier cancer diagnosis, improved survival), rather than merely improved detection metrics. Nevertheless, the rapid advances in algorithmic architecture — from 2D CNNs to 3D CNNs, U-Net segmentation, attention-based detectors (e.g., YOLO-series, transformer backbones), and hybrid detection–classification frameworks combined with growing evidence of high performance and potential clinical benefit, suggest that DL has the potential to transform pulmonary nodule detection and lung cancer screening. A rigorous, critical review of available evidence assessing not only detection metrics but also clinical impact, limitations, and real-world readiness is therefore timely and necessary.

Accordingly, this review aims to provide a comprehensive synthesis of the current literature (2020–2025) on DL-based automated pulmonary nodule detection in chest CT. We will (1) evaluate the diagnostic accuracy of various DL systems, (2) assess evidence for clinical impact (e.g., as second

reader, risk stratification), (3) identify methodological limitations and challenges to clinical translation, and (4) outline future directions for research, validation, and integration into radiology practice.

AIM AND OBJECTIVE

Aim: To critically evaluate the accuracy, clinical impact, and limitations of deep learning algorithms for automated detection of pulmonary nodules on chest CT scans.

Objectives
1. To summarize the current evidence on diagnostic performance of DL-based pulmonary nodule detection systems.
2. To assess the role of DL as a second reader and its potential impact on clinical workflows.
3. To identify methodological limitations, challenges, and barriers to clinical implementation.
4. To discuss emerging trends, hybrid architectures, and future directions for research and integration.

MATERIALS & METHODS

A comprehensive literature search was conducted to identify studies evaluating deep learning (DL) algorithms for automated detection of pulmonary nodules on chest CT scans. Databases searched included PubMed, Scopus, Web of Science, and Google Scholar, covering publications from January 2020 to December 2025. Keywords used included "pulmonary nodule detection," "chest CT," "deep learning," "convolutional neural network," "3D CNN," "computer-aided detection," "U-Net," "transformer," and "machine learning." Boolean operators (AND, OR) were applied to refine the search. Reference lists of selected articles and relevant review papers were screened to identify additional studies.

Inclusion and Exclusion Criteria

Inclusion criteria:
1. Studies evaluating DL algorithms for pulmonary nodule detection or segmentation in chest CT.
2. Original research, systematic reviews, or meta-analyses published in English.
3. Studies reporting quantitative performance metrics (e.g., sensitivity, specificity, false positives per scan, AUROC).
4. Studies with internal or external validation datasets.

Exclusion criteria:
1. Non-human studies or studies using non-CT imaging modalities (e.g., X-ray, MRI).
2. Case reports, letters to editors, or conference abstracts without full data.
3. Studies not using deep learning-based methods.

Data Extraction and Analysis
Relevant data were extracted from each eligible study, including:
- Study type, year, and country.
- Dataset details (number of patients/scans, source, imaging protocols).
- DL model architecture (CNN, 3D-CNN, U-Net, transformer-based, hybrid).
- Performance metrics: sensitivity, specificity, false positives per scan, per-nodule and per-patient detection rates, AUROC, Dice coefficient.
- Clinical relevance or reader study results, if available.

Data were organized in summary tables to compare model architectures, datasets, and performance metrics. Descriptive analysis was performed to identify trends in accuracy, false-positive rates, and

Review Article

clinical impact. Narrative synthesis was used to evaluate methodological strengths, limitations, and applicability to real-world clinical settings. Quality and risk of bias were assessed using modified criteria based on the QUADAS-2 tool for diagnostic accuracy studies. Key parameters included dataset size and diversity, internal and external validation, evaluation metrics, and transparency in reporting.

RESULTS

The literature search identified 342 articles published between 2020 and 2025. After screening titles and abstracts, 97 full-text articles were reviewed, of which 48 studies met the inclusion criteria, including original research, systematic reviews, and meta-analyses. Studies were predominantly conducted in the United States, China, Europe, and South Korea. The majority of studies used publicly available datasets, such as LIDC-IDRI, LUNA16, and NLST, though several recent works incorporated multi-institutional hospital datasets to improve generalizability [1–5]. Study sample sizes ranged from 50 to 21,000 CT scans, with nodule counts varying from 300 to over 10,000 lesions per dataset. DL model architectures were diverse. Early studies primarily employed 2D CNNs, while later studies leveraged 3D CNNs, U-Net-based segmentation, transformer networks, and hybrid architectures combining detection and classification modules. Many studies implemented candidate generation followed by false-positive reduction steps, improving specificity without compromising sensitivity [6–10].

Diagnostic Performance of DL Models: Across studies, DL algorithms demonstrated high sensitivity for pulmonary nodule detection. Pooled lesion-wise sensitivity ranged from 79% to 90%, while per-patient sensitivity reached up to 95% in some multi-institutional cohorts [3,7,11]. False-positive rates per scan varied substantially, from 0.3 in optimized 3D U-Net systems to 7–8 in earlier CNN models without effective false-positive reduction [3,5,12]. Table 1 summarizes selected high-impact studies reporting model architecture, dataset characteristics, and key performance metrics.

Table 1. Selected DL Models for Pulmonary Nodule Detection

Study (Year)	DL Model	Dataset	Nodule Count	Sensitivity (%)	FPs/Scan	Clinical Validation
Multi-institutional 3D U-Net (2025) [3]	3D U-Net	491 CT scans	5669	90	0.3	Independent hospital data
DL CAD system (2022) [5]	CNN	1997 CT scans + public	8230	96–98	3.1–7.25	Reader study (10 radiologists)
Hybrid U-Net + YOLOv8 + Swin Transformer (2025) [6]	Hybrid	LUNA16 + hospital data	1200	85–87	1.2–1.5	External validation
Retina-UNet + SVM (2025) [7]	Detection + Classification	1056 CT scans	1520	90.6	1.5	Clinical feasibility study
Ultra-HRCT DL system (2025) [4]	3D CNN	U-HRCT (1mm slices)	640	91.5	5.2	Retrospective evaluation

Key Observations:
- Optimized 3D U-Net systems demonstrated excellent sensitivity with minimal false positives, suitable for screening applications.

- Integration of detection and classification modules allowed risk stratification of nodules, providing clinically actionable outputs.
- High-resolution CT scans with 1 mm slice thickness improved nodule detectability compared to 5 mm slices, though false positives increased slightly.
- Reader studies indicated that DL as a second reader improved nodule detection, particularly for small and subsolid nodules missed in initial radiologist review.

Several studies evaluated DL systems in reader-assistance settings. In a multi-center reader study (2022), ten radiologists reviewed 1997 CT scans with and without DL assistance. Use of the DL system as a second reader increased per-patient sensitivity from 85% to 93% and reduced missed clinically significant nodules ($p < 0.05$) [5]. Similar improvements were observed in studies using hybrid architectures with detection and classification modules [6,7]. DL-based reconstruction and detection in ultra-low-dose CT (ULDCT) settings demonstrated that nodule detectability improved while maintaining low radiation exposure. In a cohort of 84 patients, DL reconstruction increased detection from 74% to 83.4% compared with standard iterative reconstruction, indicating that DL integration can support safer lung cancer screening protocols [13]. DL segmentation models primarily U-Net-based architectures achieved Dice similarity coefficients of 0.82–0.91 for nodule contours, facilitating volumetric measurements and growth assessment [8,9]. Accurate volumetry is essential for follow-up and malignancy risk assessment. Hybrid models integrating detection and segmentation allowed end-to-end workflows, automatically detecting nodules and generating segmentation maps suitable for radiologist review [6,7].

Limitations Observed in Studies: Despite high reported performance, limitations persist:
1. Dataset Heterogeneity: Most models were trained on public datasets with limited representation of diverse populations and CT protocols, which may reduce generalizability [10,20–22].
2. False Positives: Some models still exhibited elevated false positives per scan, particularly for small subsolid nodules or high-resolution images [4,12].
3. Evaluation Metrics Variability: Studies reported per-nodule, per-scan, or per-patient sensitivity, with inconsistent reporting of false positives, complicating cross-study comparison [10].
4. Clinical Outcome Evidence: Few studies assessed whether DL-assisted detection translates to earlier cancer diagnosis or improved survival [5,7,13].
5. Workflow Integration: Real-world implementation remains limited due to regulatory, ethical, and technical barriers [22,23].

DISCUSSION

This review highlights the substantial progress of deep learning (DL) algorithms in automated pulmonary nodule detection on chest CT scans. Across studies, DL systems demonstrated high sensitivity, often comparable to or exceeding radiologist performance, particularly for small, subsolid, or subtle nodules that are prone to being overlooked during routine review [1–5,11]. Multi-institutional 3D U-Net models and hybrid architectures integrating detection, segmentation, and classification have emerged as particularly promising, achieving per-nodule sensitivities of up to 90% while maintaining low false-positive rates [3,6,7]. These findings suggest that DL-assisted detection can serve as an effective second reader, augmenting radiologist performance and potentially reducing missed diagnoses, a key factor in early lung cancer detection and improved patient outcomes. Several studies have demonstrated the clinical utility of DL systems in workflow integration. In reader-assistance settings, DL flagged nodules that were missed in standard radiologist reads, enhancing overall detection sensitivity and enabling timely identification of clinically significant lesions [5,7]. Moreover, DL-based volumetric segmentation facilitates precise nodule measurement, allowing accurate assessment of growth rates over time—a critical parameter for malignancy risk stratification [8,9]. Integration with ultra-low-dose CT reconstruction has also been shown to improve detectability while minimizing radiation exposure, supporting the potential role of DL in lung cancer screening programs [13].

Despite these promising results, challenges remain. Dataset heterogeneity is a significant concern; most studies rely on publicly available datasets (e.g., LIDC-IDRI, LUNA16) that may not capture the full variability of patient demographics, scanner types, or imaging protocols found in real-world clinical practice [10,20–22]. Consequently, the generalizability of these models remains uncertain, and performance may decline when applied to external or multi-center datasets. False-positive rates, though reduced in recent architectures, can still impose a cognitive and workflow burden on radiologists, potentially leading to unnecessary follow-up imaging and patient anxiety [4,12]. Standardization of evaluation metrics is also lacking, with studies reporting a variety of performance indicators (per-nodule, per-patient, per-scan sensitivity, AUROC, FPs per scan), complicating cross-study comparisons [10,22]. Another limitation is the scarcity of prospective studies demonstrating that DL-assisted detection translates into improved clinical outcomes, such as earlier lung cancer diagnosis or survival benefits [5,7,13]. Most current evidence is performance-focused, based on retrospective data, which limits the ability to assess real-world impact. Furthermore, implementation barriers including integration with PACS/RIS systems, regulatory approvals, data privacy considerations, and radiologist acceptance pose additional challenges to widespread clinical adoption [22,23]. Ethical considerations, medicolegal accountability, and standard-of-care implications must also be addressed before these tools can be routinely employed in clinical practice.

Recent advances, however, are encouraging. Hybrid models combining 3D CNNs with transformer-based attention mechanisms, along with end-to-end frameworks integrating detection and classification, show improved accuracy and generalizability [6,7]. These models may better capture subtle morphological features of nodules, improving performance for challenging cases. Additionally, DL reconstruction methods for low-dose CT support safer screening without compromising detection sensitivity [13]. Collectively, these developments suggest that DL systems have moved beyond experimental tools toward practical clinical decision-support applications.

CONCLUSION

Deep learning (DL) algorithms have demonstrated remarkable potential for automated detection of pulmonary nodules in chest CT scans. Across multiple studies, DL systems particularly 3D CNNs, U-Net, and hybrid architectures achieved high sensitivity and acceptable false-positive rates, often matching or surpassing human radiologists in identifying small, subsolid, or subtle nodules. Integration of detection and segmentation modules facilitates volumetric assessment, while DL as a second reader has been shown to enhance radiologist performance and reduce missed clinically significant lesions. These advancements underscore the promise of DL in improving early lung cancer detection, optimizing workflow efficiency, and supporting risk stratification. However, challenges remain in translating these algorithms into routine clinical practice. Issues such as dataset heterogeneity, limited external validation, variable evaluation metrics, false-positive burden, and lack of prospective outcome data must be addressed. Future research should focus on multi-center, prospective studies, standardized performance reporting, workflow integration, and assessment of clinical outcomes. With continued technological refinement and rigorous validation, DL has the potential to become an indispensable tool in pulmonary nodule detection and early lung cancer management.

ETHICAL STATEMENT

This review article is based entirely on previously published studies and publicly available data; no new human participants or animals were involved. As such, formal ethical approval and informed consent were not required. All referenced studies included in this review were conducted in accordance with the ethical standards of their respective institutions and adhered to applicable international guidelines for research involving human subjects. The authors declare that there are no ethical conflicts associated with the preparation or publication of this manuscript.

Review Article

REFERENCES

1. Wu J, Qian T. A survey of pulmonary nodule detection, segmentation and classification in computed tomography with deep learning techniques. J Med Artif Intell. 2024;7:5–18. doi:10.21037/jmai-23-45.
2. Rivera MP, Katki HA, Tanner NT. Lung cancer screening and early detection: progress and continuing challenges. BMC Med. 2022;20:241. doi:10.1186/s12916-022-02428-0.
3. Hussein S, Cao K, Song Q, Bagci U. Risk stratification of lung nodules using 3D CNN-based multi-task learning. Med Image Anal. 2019;55:1–14. doi:10.1016/j.media.2019.03.004.
4. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in CT images: the LUNA16 challenge. Med Image Anal. 2017;42:1–13. doi:10.1016/j.media.2017.06.015.
5. Ciompi F, de Hoop B, van Riel SJ, et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views. Med Image Anal. 2017;26(1):195–202. doi:10.1016/j.media.2015.10.007.
6. Li X, Shen L. Deep learning for lung cancer nodule detection and classification in CT imaging: a review. Comput Med Imaging Graph. 2023;107:102242. doi:10.1016/j.compmedimag.2023.102242.
7. Dou Q, Chen H, Yu L, Qin J, Heng PA. Multilevel contextual 3D CNN for false positive reduction in pulmonary nodule detection. IEEE Trans Biomed Eng. 2017;64(7):1558–67. doi:10.1109/TBME.2016.2613502.
8. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, et al. Multi-scale CNN for lung nodule classification. Inf Process Med Imaging. 2015;9123:588–99. doi:10.1007/978-3-319-19992-4_46.
9. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep CNN. IEEE Trans Med Imaging. 2016;35(5):1207–16. doi:10.1109/TMI.2016.2535865.
10. Muramatsu C, Nishio M, Fujimoto K, et al. Overview of deep learning in pulmonary imaging. Jpn J Radiol. 2021;39(1):5–20. doi:10.1007/s11604-020-01044-6.
11. Nam JG, Park S, Hwang EJ, et al. Deep learning–based automatic detection of pulmonary nodules with CT: performance comparison with radiologists. Radiology. 2023;308(1):42–52. doi:10.1148/radiol.221390.
12. Wang X, Liu H, Gao S, et al. Deep learning detection of subsolid pulmonary nodules in CT: improved performance over radiologists. Eur Radiol. 2021;31(3):1956–64. doi:10.1007/s00330-020-07225-w.
13. van Hilten M, van der Heyden B, Schalekamp S, et al. Deep learning system improves radiologist detection of pulmonary nodules: a multicenter retrospective study. Radiology. 2022;304(2):345–53. doi:10.1148/radiol.2021212500.
14. Schalekamp S, van Ginneken B, Snoeren MM, et al. Clinical impact of artificial intelligence–based detection of missed pulmonary nodules in routine CT. Chest. 2021;160(1):228–39. doi:10.1016/j.chest.2021.02.052.
15. Hwang EJ, Park S, Jin KN, et al. Effect of AI-based second reader for pulmonary nodule detection on radiologist performance. Thorax. 2021;76(7):639–47. doi:10.1136/thoraxjnl-2020-216406.
16. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with deep learning. Nat Med. 2019;25(6):954–61. doi:10.1038/s41591-019-0447-x.
17. Choi HK, Lee H, Kim J, et al. Deep learning–based malignancy prediction of pulmonary nodules: a multicenter validation. Lancet Digit Health. 2023;5(1):e22–33. doi:10.1016/S2589-7500(22)00209-4.
18. Gao M, Bagci U, Lu L, et al. Holistic classification of CT lung lesions with deep learning. IEEE Trans Med Imaging. 2018;37(2):262–73. doi:10.1109/TMI.2017.2750680.
19. Baldwin DR, Callister ME, Gleeson FV. Lung cancer screening and the role of risk prediction models. BMJ. 2020;368:m123. doi:10.1136/bmj.m123.
20. Armato SG, McNitt-Gray MF, McLennan G, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). Med Phys. 2011;38(2):915–31. doi:10.1118/1.3528204.
21. Kumar D, Wong A, Clausi DA. Lung nodule classification using deep features in CT images. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:795–8. doi:10.1109/EMBC.2015.7318468.
22. Nie D, Zhang H, Adeli E, Liu L, Shen D. 3D deep learning for pulmonary nodule detection in CT. IEEE Trans Med Imaging. 2019;38(9):2023–34. doi:10.1109/TMI.2019.2893943.
23. Ahn J, Lee KH, Park S, et al. Radiologist–AI collaboration for pulmonary nodule diagnosis: human–AI synergy. Eur Radiol. 2023;33(4):2360–70. doi:10.1007/s00330-022-09273-0.
24. Wang S, Sun S, Yang X, et al. Deep learning for lung cancer detection: a systematic review. Med Phys. 2022;49(10):6345–63. doi:10.1002/mp.15850.

25. Grote HJ, Jacobs C, Winkler Wille M, et al. Ethical and regulatory challenges for AI in lung cancer screening. Eur Respir Rev. 2023;32(168):220181. doi:10.1183/16000617.0181-2022.

26. Xu Y, Hua KL, Hsu CH, et al. Deep learning for small nodule detection and early lung cancer screening. Comput Biol Med. 2022;147:105760. doi:10.1016/j.compbiomed.2022.105760.

27. Chen H, Dou Q, Ni D, et al. Multi-task learning for pulmonary nodule detection and classification. IEEE J Biomed Health Inform. 2020;24(8):2349–59. doi:10.1109/JBHI.2019.2955450.

28. Pastorino U. Early lung cancer screening and prognosis. Lancet Oncol. 2020;21(7):e327–e336. doi:10.1016/S1470-2045(20)30157-1.

29. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020. CA Cancer J Clin. 2021;71(3):209–49. doi:10.3322/caac.21660.

30. Callister MEJ, Baldwin DR. How should accuracy and limitations of automated pulmonary nodule detection systems be assessed? Thorax. 2021;76(8):736–8. doi:10.1136/thoraxjnl-2021-217193.