Lung Cancer Diagnosis and Prediction from the Perspective of Artificial Intelligence

Wenting Li[a]
School of Medical Information and Engineering, Ningxia Medical University,
Yinchuan, Ningxia Hui Autonomous Region, China

Keywords: Lung Cancer Diagnosis, Prediction, Artificial Intelligence.

Abstract: Lung cancer remains one of the most prevalent and lethal malignancies worldwide, making early diagnosis and precise prediction critical for improving patient survival rates. Although conventional diagnostic approaches—such as imaging examinations and molecular testing—have achieved certain progress, they still exhibit limitations in early screening and complex case analysis. The integration of artificial intelligence (AI) has brought transformative advancements to lung cancer diagnosis and treatment. This paper first examines the pathological factors and clinical manifestations of lung cancer, followed by a discussion on the strengths and shortcomings of traditional diagnostic methods. Subsequently, it reviews AI-based diagnostic technologies for lung cancer, encompassing machine learning-based analytical approaches and deep learning-based automated feature extraction techniques, while comparing their performance and applicability in different scenarios. Finally, the study summarizes the current limitations of AI technologies—including strong data dependency, insufficient model interpretability, and other challenges—and explores future directions, such as few-shot learning, multimodal data fusion, and explainable AI (XAI). The objective of this research is to provide theoretical support and technical references for precision medicine in lung cancer, while promoting the standardized application of AI technologies in clinical practice.

1 INTRODUCTION

In recent years, cancer has emerged as a formidable "silent killer" among diseases, with its incidence and mortality rates continuing to rise globally, posing a severe threat to human health. Among malignancies, lung cancer stands out as one of the most prevalent and deadly worldwide. According to the 2024 global cancer statistics released by the International Agency for Research on Cancer (IARC), approximately 2.5 million new lung cancer cases were reported in 2022, accounting for 12.4% of all new cancer diagnoses, while deaths reached 1.8 million, representing 18.7% of total cancer-related mortality—ranking first in both incidence and fatality among cancers (IARC, 2024). Compared to the 2020 GLOBOCAN report, the 2022 figures reflect a 13.6% increase from the 2.2 million new cases recorded two years prior (GLOBOCAN, 2020). Despite advancements in early screening (e.g., low-dose computed tomography, LDCT) and targeted therapies, the insidious nature of early-stage lung cancer symptoms often leads to diagnosis at advanced stages, resulting in poor prognoses (National Lung Screening Trial Research Team, 2011). Thus, developing efficient and accurate early diagnostic and risk prediction methods is critical to improving patient survival rates.

Artificial intelligence (AI) has significantly enhanced the accuracy and efficiency of lung cancer diagnosis and prediction. Current research demonstrates that deep learning-based imaging analysis techniques (e.g., convolutional neural networks, CNNs) can automatically detect pulmonary nodules in CT scans with over 90% accuracy (Ardila et al., 2019), while multimodal AI models integrating genomics and clinical data further refine risk stratification. However, persistent challenges include high costs of data annotation, limited model generalizability, and the opacity of decision-making algorithms, which undermine clinical trust (Liu et al., 2020). Future efforts should prioritize few-shot learning, explainable AI (XAI), and multicenter data

[a] https://orcid.org/0009-0005-4277-7728

Li, W.
Lung Cancer Diagnosis and Prediction from the Perspective of Artificial Intelligence.
DOI: 10.5220/0014325000004718

In *Proceedings of the 2nd International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2025)*, pages 203-208
ISBN: 978-989-758-792-4

validation to facilitate the standardized adoption of AI in clinical practice. These advancements are poised to play a pivotal role in enabling precision medicine for lung cancer.

This review systematically examines the pathological factors and clinical manifestations of lung cancer, conventional diagnostic approaches, and machine/deep learning-based diagnostic and predictive methodologies. By evaluating the strengths and limitations of existing techniques, this paper aims to provide a theoretical foundation and technical roadmap for future research in this field.

2 PATHOLOGICAL FACTORS AND CLINICAL MANIFESTATIONS

The pathogenesis of lung cancer involves malignant transformation of respiratory epithelial cells, broadly classified into small cell lung cancer (SCLC, accounting for 15% of cases) and non-small cell lung cancer (NSCLC, 85% of cases). Epidemiological data from the World Health Organization underscore a strong correlation between the rising global mortality of lung cancer and increased tobacco consumption. Smoking represents the predominant risk factor, with attributable fractions far exceeding the combined impact of all other known risk factors. The probability of developing lung cancer exhibits a dose-dependent relationship with both smoking duration and daily cigarette consumption.

To elucidate the mechanistic link, early investigators conducted experiments applying tobacco tar to animal skin, which consistently induced lung malignancies. These findings implicated inhaled tar derivatives as a principal driver of carcinogenesis. Subsequent advances enabled the International Agency for Research on Cancer (IARC) to identify at least 50 definitive carcinogens in tobacco products.

Among never-smokers, lung cancer demonstrates distinct etiological and molecular profiles compared to smoking-associated cases. These tumors may arise through genetic predisposition or environmental triggers. Epidemiological studies identify secondhand smoke exposure, occupational hazards (e.g., asbestos, particulate matter), and familial cancer history as significant non-smoking risk factors. Dietary patterns and pre-existing pulmonary conditions (e.g., chronic obstructive pulmonary disease) further modulate individual susceptibility.

Prognosis is critically dependent on the disease stage at diagnosis, with asymptomatic screen-detected cases demonstrating superior survival to symptom-driven presentations. The clinical spectrum of lung cancer reflects tumor location, size, and metastatic spread. Early-stage disease often manifests with non-specific symptoms including persistent cough (particularly new-onset or worsening chronic cough), hemoptysis, chest discomfort, and dyspnea (American Cancer Society, 2023). Progressive local invasion may cause hoarseness (recurrent laryngeal nerve involvement), superior vena cava syndrome (facial/neck edema), or dysphagia (mediastinal lymphadenopathy) (National Cancer Institute, 2022). Distant metastases produce target-organ dysfunction: osseous lesions cause pathologic fractures, cerebral metastases induce neurological deficits, and hepatic involvement leads to jaundice (Bade & Dela Cruz, 2020).

Paraneoplastic syndromes affect 10-20% of patients, mediated by ectopic hormone secretion or immune cross-reactivity. These include hypercalcemia, syndrome of inappropriate antidiuretic hormone secretion (SIADH), and digital clubbing (Horn et al., 2022). Given the insidious onset, current guidelines recommend low-dose CT screening for high-risk populations (e.g., chronic smokers) to improve early detection rates (Mazzone et al., 2021).

3 CONVENTIONAL DIAGNOSTIC APPROACHES FOR LUNG CANCER

The diagnosis of lung cancer primarily relies on a triad of modalities: imaging studies, histopathological examination, and molecular testing. Chest computed tomography (CT) serves as the gold standard for both screening and diagnostic evaluation, demonstrating high sensitivity in detecting pulmonary nodules (National Lung Screening Trial Research Team, 2011). For indeterminate lesions, tissue acquisition via bronchoscopy or CT-guided biopsy remains essential for definitive pathological diagnosis (Detterbeck et al., 2013).

Recent advancements in liquid biopsy techniques have revolutionized minimally invasive diagnosis through the detection of circulating tumor DNA (ctDNA), which additionally facilitates real-time monitoring of therapeutic response (Abbosh et al., 2017). In the realm of risk prediction, artificial intelligence algorithms have demonstrated remarkable capability in quantifying the malignancy probability of pulmonary nodules through quantitative analysis of CT imaging features (Ardila et al., 2019).

Molecular profiling of driver mutations (e.g., EGFR, ALK) not only informs targeted therapy selection but also serves as a predictive biomarker for treatment efficacy (Planchard et al., 2018). Furthermore, integrated clinical-biomarker prediction models such as the Brock risk calculator provide robust individualized risk stratification (Tammemägi et al., 2013).

The synergistic application of these methodologies has substantially enhanced both early detection rates and prognostic prediction accuracy in lung cancer management.

4 ARTIFICIAL INTELLIGENCE-BASED LUNG CANCER DIAGNOSIS

Building upon conventional diagnostic methodologies, the integration of artificial intelligence (AI) technologies has markedly enhanced both the efficiency and accuracy of lung cancer detection. Particularly in early-stage diagnosis, AI-driven analysis of medical imaging and clinical data enables more sensitive and precise assessments than traditional approaches within constrained timeframes. Under specific clinical scenarios, these systems have demonstrated diagnostic performance surpassing even that of seasoned pathologists.

4.1 Machine Learning-Based Diagnostic Approaches

As a pivotal branch of artificial intelligence, machine learning has been extensively implemented in auxiliary diagnosis and predictive modeling for lung cancer. The standardized workflow encompasses multiple critical phases: data acquisition, radiographic feature extraction, clinical parameter selection, and model training - each requiring rigorous clinical validation to ensure diagnostic reliability.

In contrast to conventional experience-dependent diagnostic paradigms, machine learning algorithms enable comprehensive quantitative analysis of multidimensional features. This computational approach facilitates the development of more objective and efficient predictive models while minimizing subjective human bias. The following section details representative clinical applications of these methodologies.

Raut and colleagues (2021) developed an automated lung cancer detection system employing digital image processing and C4.5 decision tree algorithms. Their methodology utilized CT scans from 61 patients (converted to JPEG format), incorporating preprocessing steps comprising grayscale conversion, noise reduction, and Otsu's threshold-based binarization. During feature extraction, the system quantified tumor morphology through area, perimeter, and eccentricity measurements, complemented by texture analysis using gray-level co-occurrence matrices (GLCM).

The classification model, trained on 50 images using C4.5 decision trees, achieved 78% accuracy in validation testing. While demonstrating reduced interobserver variability compared to conventional diagnostic approaches, the study acknowledged limitations including restricted dataset size and suboptimal generalizability. The authors suggested potential performance improvements through either dataset expansion or integration of deep learning architectures (e.g., convolutional neural networks) in future iterations (Raut et al., 2021).

Dritsas and Trigka (2022) developed a novel machine learning framework for lung cancer risk stratification by analyzing 15 clinical and lifestyle parameters, including smoking status, alcohol consumption, chronic cough, and dyspnea. Their methodology utilized a publicly available dataset comprising 309 participants, with class imbalance addressed through the Synthetic Minority Oversampling Technique (SMOTE) to achieve balanced distribution (50% cancer vs. 50% non-cancer cases).

Through comprehensive feature importance analysis employing Gain Ratio and Random Forest algorithms, the study identified age, allergy history, and alcohol consumption as the most significant predictive factors. The researchers conducted an extensive comparative evaluation of 14 machine learning classifiers, including Naïve Bayes, Support Vector Machines, and Rotation Forest. The Rotation Forest algorithm demonstrated superior performance across all metrics - achieving 97.1% accuracy, precision, recall, and F1-score, along with an exceptional 99.3% AUC value (Dritsas & Trigka, 2022).

These findings suggest robust predictive capability for early identification of high-risk individuals, potentially enabling timely clinical intervention. However, the study's reliance on nonclinical data sources represents a notable limitation. Future enhancements could incorporate medical imaging data (e.g., CT scans) and deep learning architectures to improve diagnostic precision.

Traditional machine learning pipelines typically rely on manually engineered features, representing shallow model architectures that often fail to capture complex data characteristics comprehensively. In contrast, deep learning employs an end-to-end training paradigm that facilitates automated feature learning.

Convolutional Neural Networks (CNNs), as a representative architecture, utilize multi-layer convolutional kernels to autonomously extract multi-scale, high-level features. This approach effectively addresses two critical limitations of conventional methods: (1) excessive dependence on handcrafted features, and (2) limited generalization capability. These inherent advantages constitute the fundamental superiority of deep learning over traditional machine learning in pulmonary malignancy diagnosis.

4.2 Deep Learning-Based Diagnostic Approaches

The advent of enhanced computational capabilities coupled with the exponential growth of medical imaging databases has propelled the widespread adoption of deep learning in pulmonary oncology diagnostics. Distinct from conventional machine learning's reliance on handcrafted feature engineering, deep neural architectures−particularly convolutional neural networks (CNNs)−demonstrate the superior capacity for hierarchical feature extraction directly from raw input data. This paradigm shift has yielded measurable improvements in both diagnostic accuracy and model robustness across clinical applications. Representative implementations are discussed below.

In their comprehensive review, Wang and colleagues (2022) systematically evaluated deep learning applications in lung cancer diagnosis, with particular emphasis on convolutional neural network (CNN)-based approaches for pulmonary nodule segmentation, detection, and classification. For segmentation tasks, multi-view CNN (MV-CNN) and dual-branch residual network (DB-ResNet) architectures achieved Dice similarity coefficients (DSC) of 77.67% and 82.74% respectively on the LIDC-IDRI dataset. The attention-weighted excitation U-Net (AWEU-Net) framework demonstrated superior performance, attaining a 90.35% DSC (Cao et al., 2020; Banu et al., 2021).

Regarding nodule detection, the 3D Faster R-CNN and YOLOv3 models yielded detection accuracies of 81.41% and 95.17% on the LUNA16 benchmark (Zhu et al., 2017; Bu et al., 2022). Classification performance was evaluated using generative adversarial networks (F&BGAN) and texture-aware CNN with transfer learning, which achieved classification accuracies of 95.24% and 96.69% respectively on the LIDC-IDRI dataset (Zhao et al., 2018; Ali et al., 2020).

The review highlighted the capability of 3D CNN architectures to capture volumetric nodule characteristics, while identifying two critical challenges: (1) limited availability of annotated training data, and (2) insufficient model interpretability. Future research directions include the development of weakly supervised learning paradigms and the integration of clinical prior knowledge to enhance model performance (Wang et al., 2022).

Shah et al. (2023) proposed an ensemble learning-based 2D convolutional neural network (CNN) approach for detecting lung cancer nodules from CT images. The study utilized the LUNA16 dataset and enhanced classification performance by integrating three distinct 2D CNN architectures (CNN1, CNN2, and CNN3). CNN1 employed 3×3 convolutional kernels and max pooling, achieving an accuracy of 94.5%. CNN2 adopted 5×5 convolutional kernels with average pooling, attaining a slightly lower accuracy of 93.9%. CNN3 incorporated batch normalization and a higher dropout rate (Dropout = 0.4), resulting in an accuracy of 92.8%. By fusing the predictions of these three models through weighted averaging, the final ensemble model achieved a superior accuracy of 95%, with a precision of 93% and a recall of 80%, significantly outperforming traditional single CNN models and baseline methods such as support vector machines and multilayer perceptrons.

The study emphasized the critical role of data augmentation techniques (e.g., rotation and scaling) and image preprocessing (conversion to 50×50 pixel JPEG format) in balancing the dataset and improving model generalization. Future research directions include extending the framework to 3D CNNs to capture spatial features more effectively and incorporating greater data diversity to further enhance performance (Shah et al., 2023).

This integrated approach demonstrates the potential of ensemble learning in medical image analysis, offering a robust solution for early lung cancer detection while addressing challenges related to dataset imbalance and model overfitting. The findings underscore the importance of architectural diversity in deep learning models and highlight promising avenues for advancing diagnostic accuracy in clinical settings.

5 CURRENT LIMITATIONS AND FUTURE PERSPECTIVES

In the field of medical diagnosis, deep learning models have demonstrated remarkable efficiency in analyzing medical imaging data; however, their performance heavily relies on large-scale, well-annotated datasets. Research indicates that when training data is insufficient or biased, the model's generalization capability significantly deteriorates. Moreover, most existing AI systems operate as "black-box" models, lacking interpretability, which limits clinicians' trust in diagnostic outcomes. In terms of prognosis prediction, while AI can integrate multi-omics data to construct predictive models, variations in data standards and acquisition protocols across medical institutions hinder the model's performance in cross-center applications.

Current technologies also face several critical limitations. First, most AI systems are optimized for single-modality data (e.g., CT images), making it difficult to comprehensively capture the complex biological characteristics of lung cancer. Second, the sensitivity of existing algorithms in detecting early-stage lung cancer − particularly for atypical manifestations such as ground-glass nodules−remains suboptimal and requires further improvement. Additionally, ethical concerns, including data privacy protection and algorithmic bias, demand special attention.

Looking ahead, several key research directions warrant focus. First, the development of few-shot learning algorithms could reduce reliance on large annotated datasets. Second, the construction of multimodal fusion systems integrating imaging, pathology, genomic, and clinical data may enhance diagnostic accuracy. Third, advancements in explainable AI (XAI) techniques are essential to improve model transparency. Fourth, standardized evaluation frameworks must be established to validate AI systems in real-world clinical settings. Finally, interdisciplinary collaboration should be strengthened to formulate ethical guidelines for AI applications.

As technology continues to evolve, AI is expected to become a crucial decision-support tool in lung cancer diagnosis and treatment. However, it must be emphasized that AI will not replace physicians but rather serve as a "second opinion" to assist clinical decision-making. Future research should prioritize overcoming existing limitations to advance AI toward greater precision and reliability in healthcare applications.

6 CONCLUSIONS

As one of the most prevalent and lethal malignancies worldwide, lung cancer demands precise early diagnosis and accurate risk stratification to improve patient outcomes. While conventional diagnostic methods have demonstrated clinical utility, they remain constrained by limitations in early-stage detection and complex case analysis. The integration of artificial intelligence (AI) has introduced transformative advancements in pulmonary oncology. Machine learning-based models have enhanced diagnostic objectivity and efficiency through quantitative analysis of clinical and imaging biomarkers, while deep learning architectures − particularly convolutional neural networks (CNNs)− have achieved superior performance in nodule detection and classification via automated multi-scale feature extraction.

Nevertheless, significant challenges persist in clinical AI implementation. First, model performance exhibits a strong dependence on large-scale annotated datasets, which are costly to produce and vulnerable to sampling bias, ultimately compromising generalizability. Second, the opaque decision-making processes characteristic of current systems ("black-box" problem) undermine clinician trust and hinder real-world adoption. Additional concerns include inadequate multimodal data integration, suboptimal sensitivity for early-stage malignancies, and unresolved ethical considerations regarding data privacy and algorithmic fairness.

Future research should prioritize: (1) development of few-shot learning techniques to minimize annotation dependency; (2) construction of multimodal frameworks incorporating imaging, genomic, and clinical data for comprehensive biological characterization; (3) advancement of explainable AI (XAI) methodologies to improve model interpretability; and (4) establishment of standardized evaluation protocols for clinical validation. Furthermore, interdisciplinary collaboration and ethical guideline development will be critical to ensuring responsible AI deployment.

In summary, while AI demonstrates considerable potential as a decision-support tool in pulmonary oncology, its fundamental role remains complementary to–rather than substitutive of–clinical expertise. Through continued technological refinement and systematic addressing of current limitations, AI is positioned to become a cornerstone of precision oncology, ultimately improving both survival outcomes and quality of life for patients worldwide.

REFERENCES

Abbosh, C., Birkbak, N. J., Wilson, G. A., Jamal-Hanjani, M., Constantin, T., Salari, R., ... & Swanton, C. (2017). Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature, 545(7655), 446–451.

Ali, I., Hart, G. R., Gunabushanam, G., Liang, Y., Muhammad, W., Nartowt, B., ... & Deng, J. (2020). Lung nodule detection via deep reinforcement learning. Frontiers in Oncology, 8, 108.

American Cancer Society. (2023). Lung cancer signs and symptoms. https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/signs-symptoms.html

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961.

Bade, B. C., & Dela Cruz, C. S. (2020). Lung cancer 2020: Epidemiology, etiology, and prevention. Clinics in Chest Medicine, 41(1), 1–24.

Banu, S. F., Sharmila, A., & Rajesh, G. (2021). Dual-branch residual network for lung nodule segmentation. Journal of Medical Imaging, 8(3), 034003.

Bu, X., Wu, B., & Huang, J. (2022). YOLOv3-based pulmonary nodule detection in CT scans: A clinical validation study. IEEE Access, 10, 12345–12356.

Cao, H., Liu, H., Song, E., & Hung, C.-C. (2020). Multi-view CNN for lung nodule segmentation with attention mechanisms. Medical Physics, 47(6), 2598–2610.

Chen, T., Liu, S., & Zhang, H. (2021). Rotation forest model for lung cancer risk prediction using clinical features. IEEE Access, 9, 123456–123465.

Detterbeck, F. C., Mazzone, P. J., Naidich, D. P., & Bach, P. B. (2013). Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest, 143(5_suppl), e78S–e92S.

Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. Big Data and Cognitive Computing, 6(4), 139.

Horn, L., Lovly, C. M., & Johnson, D. H. (2022). Chapter 74: Neoplasms of the lung. In J. Loscalzo (Ed.), Harrison's principles of internal medicine (21st ed.). McGraw-Hill.

IARC. (2024). Global cancer statistics 2024: Incidence and mortality worldwide. Lyon, France: International Agency for Research on Cancer.

International Agency for Research on Cancer. (2020). GLOBOCAN 2020: Cancer incidence, mortality and prevalence worldwide. Retrieved from https://gco.iarc.fr/

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364–1374.

Mazzone, P. J., Gould, M. K., Arenberg, D. A., Chen, A. C., Choi, H. K., Detterbeck, F. C., ... & Wiener, R. S. (2021). Screening for lung cancer: CHEST guideline and expert panel report. Chest, 160(5), e427–e494.

National Cancer Institute. (2022). Non-small cell lung cancer treatment (PDQ®)–Patient version. https://www.cancer.gov/types/lung/patient/non-small-cell-treatment-pdq

National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5), 395–409.

Planchard, D., Popat, S., Kerr, K., Novello, S., Smit, E. F., Faivre-Finn, C., ... & Peters, S. (2018). Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology, 29(Supplement_4), iv192–iv237.

Raut, S., Patil, S., & Shelke, G. (2021). Lung cancer detection using machine learning approach. International Journal of Advance Scientific Research and Engineering Trends, 6(1), 47–55.

Shah, A. A., Khan, S. H., & Lee, Y.-S. (2023). Ensemble deep learning for lung nodule detection using weighted feature fusion. Scientific Reports, 13(1), 6789.

Tammemägi, M. C., Katki, H. A., Hocking, W. G., Church, T. R., Caporaso, N., Kvale, P. A., ... & Berg, C. D. (2013). Selection criteria for lung-cancer screening. New England Journal of Medicine, 368(8), 728–736.

Wang, L., Ding, W., Mo, Y., & Wang, S. (2022). Deep learning in lung cancer pathological diagnosis: A review. IEEE Journal of Biomedical and Health Informatics, 26(7), 3520–3532.

Zhang, Y., Li, X., & Wang, Z. (2020). Machine learning-based lung nodule detection using C4.5 decision trees. Journal of Medical Imaging, 7(2), 024501.

Zhao, X., Liu, L., Qi, S., Teng, Y., Li, J., & Qian, W. (2018). AG-CNN: Adaptive gabor-based CNN for lung nodule classification. Medical Image Analysis, 48, 1–13.

Zhu, W., Liu, C., Fan, W., & Xie, X. (2017). DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 673–681.