



RADemics

Machine Learning Models for Early Detection of Lung Cancer Using Medical Imaging

Shailendra Kumar Diwakar, Kopila
Gurung

ERA UNIVERSITY, STNM MULTI SPECIALTY
HOSPITAL

Machine Learning Models for Early Detection of Lung Cancer Using Medical Imaging

¹Shailendra Kumar Diwakar, Assistant Professor, Department of Radiological and Imaging Techniques, Era Institute of Allied Health Science and Research, Era University, Lucknow. India. mrshailendra11071993@gmail.com

²Kopila Gurung, Medical Imaging Technologist, STNM Multi specialty Hospital, Health and Family Welfare Department Government of Sikkim, Gangtok Sikkim, India. k.ghonday@gmail.com

Abstract

Early detection of lung cancer remains a critical challenge in clinical oncology, as survival rates are highly dependent on timely diagnosis and intervention. Advances in medical imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), have enabled detailed visualization of pulmonary structures, yet manual interpretation of volumetric scans was labor-intensive and prone to inter-observer variability. Machine learning (ML) has emerged as a transformative solution for automated, accurate, and reproducible analysis of medical images, facilitating the identification of pulmonary nodules, assessment of malignancy risk, and prediction of disease progression. This chapter presents a comprehensive exploration of state-of-the-art ML techniques applied to lung cancer detection, encompassing supervised and deep learning models, hybrid feature selection approaches, reinforcement learning, and ensemble strategies for enhanced diagnostic performance. The discussion emphasizes data preparation, including annotation standards, multi-radiologist consensus, and preprocessing methods essential for building robust models. The chapter also addresses challenges related to multi-center dataset variability, class imbalance, and model interpretability, highlighting the importance of external validation and explainable AI frameworks to ensure clinical applicability and trustworthiness. Multi-modal imaging integration and predictive analytics are examined as advanced strategies for improving sensitivity and specificity in early-stage detection. By synthesizing current research, methodological advances, and practical considerations, this chapter provides a roadmap for developing accurate, generalizable, and clinically deployable ML-based diagnostic systems for lung cancer.

Keywords: Lung Cancer, Machine Learning, Medical Imaging, Multi-Modal Imaging, Feature Selection, Explainable AI

Introduction

Lung cancer continues to be one of the most lethal malignancies worldwide, accounting for a significant proportion of cancer-related mortality [1]. The asymptomatic progression of the disease in early stages often results in late diagnosis, when therapeutic interventions are less effective and survival rates decline drastically [2]. Conventional diagnostic methods, such as chest X-rays and tissue biopsies, are limited in sensitivity and specificity, particularly for small or atypical nodules [3]. While high-resolution imaging techniques such as computed tomography (CT), magnetic

resonance imaging (MRI), and positron emission tomography (PET) have significantly enhanced the visualization of pulmonary structures, the manual interpretation of these complex datasets was time-consuming and subject to considerable inter- and intra-observer variability [4]. Consequently, there was an increasing need for automated, precise, and reproducible diagnostic systems capable of identifying early-stage lung lesions, reducing false negatives, and supporting timely clinical decision-making [5].

Machine learning (ML) has emerged as a powerful computational tool for addressing the limitations of traditional diagnostic approaches [6]. By leveraging large volumes of medical imaging data, ML models can identify intricate patterns and subtle anomalies that may elude human observation [7]. Supervised learning algorithms, such as support vector machines, random forests, and gradient boosting, have demonstrated significant success in classifying pulmonary nodules, while deep learning architectures, particularly convolutional neural networks (CNNs), have excelled at feature extraction and hierarchical representation learning [8]. These models can be trained to differentiate between benign and malignant lesions, estimate malignancy risk, and predict disease progression [9]. The adaptability of ML systems to multi-dimensional imaging data and their ability to continuously improve with additional training data position them as essential tools for advancing early detection and precision diagnostics in lung oncology [10].

A crucial aspect of machine learning applications in lung cancer detection was the preparation and preprocessing of medical imaging data [11]. Annotation standards and multi-radiologist consensus are vital to ensure accurate labeling of nodules, reduce subjective variability, and maintain dataset quality [12]. Data preprocessing techniques, including noise reduction, intensity normalization, and image segmentation, enable models to focus on relevant anatomical features while minimizing irrelevant or redundant information [13]. Handling class imbalance, where malignant nodules are less frequent than benign ones, was essential to prevent bias in predictive outcomes [14]. Feature selection methods, particularly hybrid approaches that combine filter, wrapper, and embedded strategies, enhance model performance by identifying the most discriminative imaging characteristics, reducing dimensionality, and improving computational efficiency. The integration of these preprocessing and feature selection strategies lays the foundation for developing robust, interpretable, and clinically applicable machine learning models [15].