The Application of Machine Learning in the Diagnosis of Lung Cancer

Shengkun Li

Department of
Electronic Engineering, Southwest
Jiaotong University, Chengdu, China
el22sl@leeds.ac.uk

Abstract:

Lung cancer remains the main cause of cancer-related incidence and mortality globally, with more than 2,220,000 newly diagnosed cases annually and a five-year survival rate of less than 25%. The complexity of diagnosis and treatment is exacerbated by Intra-tumoral Heterogeneity (ITH), which drives therapy resistance. Recent advances in the field of Machine Learning (ML) and Deep Learning (DL) offer promising solutions by enabling the analysis of high-dimensional medical data beyond human capability. This review explores the applications of ML in lung cancer diagnosis, focusing on Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Vision Transformer (ViT)-based models across radiomics, histopathology, and gene expression analysis. Innovative techniques such as semi-supervised learning, data augmentation, and optimization algorithms have enhanced model performance, achieving high accuracy in classifying lung cancer subtypes and predicting genetic mutations. Federated learning emerges as a privacy-preserving approach for collaborative training across institutions, addressing critical data security concerns. However, significant challenges remain, including limited model interpretability, generalizability across diverse populations, and integration into clinical workflows. Future research should prioritize interpretable Artificial Intelligence (AI) frameworks and privacy-preserving technologies to enable earlier diagnosis and tailored therapies for lung cancer patients.

Keywords: Lung cancer diagnosis; deep learning; convolutional neural networks.

1. Introduction

Lung cancer, defined as a primary bronchogenic carcinoma, is a malignant neoplasm arising from the bronchial epithelium, submucosal glands, or alveolar lining. It represents the predominant global malignancy in both incidence and cancer-related mortality. With over 2.20 million new cases diagnosed annually [1], approximately 75% of patients die within five years of diagnosis [1], highlighting its severe threat to human life. The complexity of treatment is significantly compounded by Intra-Tumoral Heterogeneity (ITH), wherein cellular diversity within tumors promotes resistance to therapy. Over recent decades, major collaborative efforts have driven advances in cancer research, resulting in extensive multimodal databases that integrate clinical records, radiological imaging, and genomic sequencing data. However, the rapid growth of such high-dimensional diagnostic and therapeutic data has rendered traditional expert-driven analysis insufficient. Faced with heterogeneous, large-scale datasets, researchers increasingly encounter limitations in manually identifying complex biomolecular patterns. This necessitates the integration of machine learning, which provides scalable computational frameworks to decipher intricate feature interdependencies, automate early cancer detection, and generate clinically actionable insights beyond human analytical capacity.

Machine Learning (ML) has been widely applied To engineer optimal resolutions for multifaceted conundrums, such as in the fields of healthcare, finance, environment, marketing, security and industry. The characteristics of ML methods are that they can examine a large amount of data and uncover their correlations, provide explanations, and detect regularities. ML can help improve the the dependability, efficiency, consistency, and precision of many disease diagnosis systems [2].

Currently, leveraging machine learning techniques in the diagnose field mainly focuses on computational biology and radiomic methods. In computational biology, deep learning has found use across five major domains: predicting protein structures and functions, advancing genome engineering, supporting systems biology and data integration, and improving phylogenetic analysis [2]. Another key area is radiomic methods. Radiomics involves four core stages: preparing images, delineating tumors, deriving features, and forecasting clinical outcomes [3]. At present, the main focus of research on lung cancer is concentrated in the field of imaging. Whether it is based on chest CT, PET-CT or head MRI, all are analyzed based on imaging data and then diagnosed. Recent advances in computational image feature extraction have enabled the development of features capable of capturing substantially more information than human visual perception, thereby accelerating the emergence of radiomics [3].

In the domain of detecting lung cancer, a substantial amount of have demonstrated the suitability of machine learning in this area. Some scholars have been able to make relatively accurate predictions by combining 12 different potential machine learning algorithms with 11 different symptoms of lung cancer [4]. A large number of models have been introduced into the diagnosis of lung cancer for experimental use, such as two-dimensional and three-dimensional CNNs, dual-stream architectures, NLP models, and visual transformer networks [5]. Riquelme et al. summarized advanced algorithms and machine learning architectures in CAD systems for diagnosis of lung cancer [5]. In addition, several studies have tested the application of DL in lung cancer diagnosis. A synthetical study by Chiu et al. underlined the use of ML in lung cancer screening via CXR and chest CT, emphasizing the FDA-approved AI system that are transforming fundamentally detection pattern, which demonstrates that AI detection for lung cancer has now entered the stage of practical application [6].

This article primarily examines the curation and analysis of medical imaging datasets for lung cancer. It explores artificial intelligence applications in diagnosis especially. The study also critically discussed some key implementation challenges to inform future advances.

2. Method

2.1 DNN-based Classification

In the area of lung cancer diagnosis, this DNN model is also considered a classic approach for classification tasks. Many researchers have expanded upon and conducted experiments with this model. A study provides a potential direction for accurate cancer diagnosis by utilizing a large Kaggle dataset combined with advanced deep learning methods [7]. Bhatia et al. employed a unique data segmentation method and innovative data processing techniques, and based on the DNN-based classification, established a lightweight lung cancer diagnosis model [8]. This article will elaborate on its usage methods and logic. The Fig. 1 below provides a simple description of their experimental procedure.

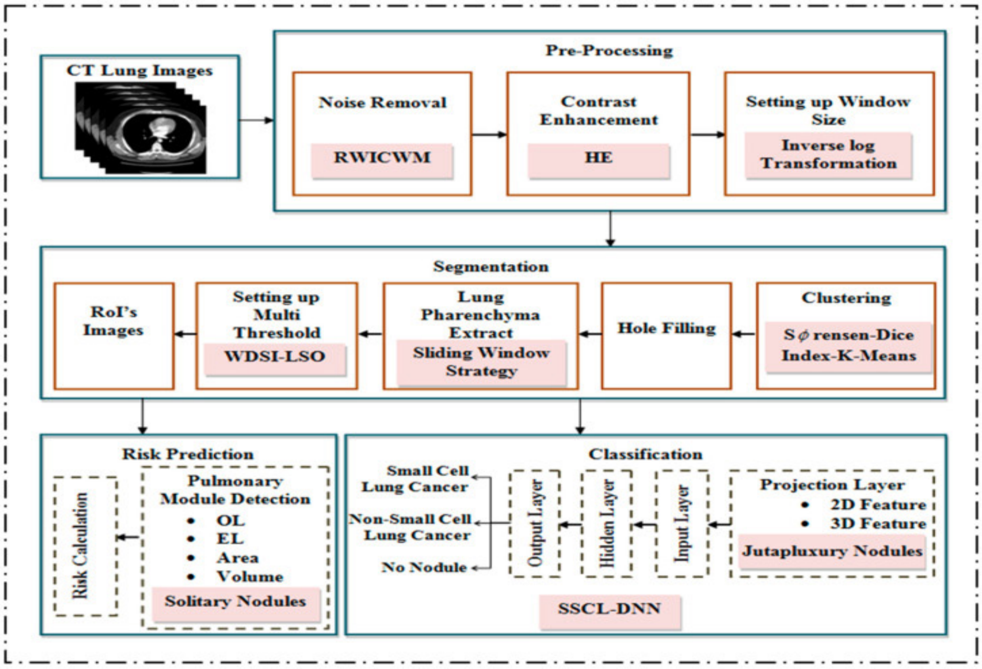

Fig. 1 The data processing structure used in the experiment [8].

In the pre-treatment stage of the experiment conducted by Bhatia et al., lung CT images in DICOM format from the LIDC dataset are used. A RWICWM filter is applied for denoising, which preserves edge details better than traditional filters such as Gaussian, Wiener, and Guided filters. Histogram equalization is then performed to enhance image contrast. Image quality is evaluated using metrics like PSNR, MSE, and SSIM.

Then in the segmentation phase, Anan improved K-means partitioning algorithm applying the Sørensen–Dice Index as the distance measure segments the images. This approach accurately isolates lung nodule regions and is compared with conventional clustering methods (e.g., standard K-means and density-based clustering) with respect to detection accuracy and error rates. This is the most remarkable innovation of this experiment. In the next stage, pulmonary nodules are further detected using WDSI-LSO, which integrates a scale parameter following a Weibull distribution scale factor with a light spectrum optimizer. Features such as area, volume, overlap tolerance, and elongation are extracted. If solid nodules are detected, risk assessment is performed using the PLCOm2012 model, incorporating environmental factors like smoking and family history to predict patient survival.

In the final classification stage, using the LUNA16 and LIDC-IDRI datasets, lung nodules are classified into categories: normal, benign, primary lung cancer, and metastatic lesions. Images, after enhancement and normalization, are fed into a semi-supervised and contrastive learning-based deep neural network (SSCL-DNN). This model consists of two subnetworks—one for classifier training and another for feature extraction through contrastive learning. The final output classifies nodules categorized as SCLC, NSCLC, or absence of nodules.

2.2 CNN-based Classification

Thangamani M et al. proposed a weighted CNN model with gene data expression [9]. This team highlights the critical challenge of early lung cancer prediction by developing a CNN-based model. The proposed architecture consists of two feature extraction blocks, each comprising a 5×5 convolutional tier attached by a 2×2 subsampling layer, then a 1×1 convolutional tier and a 1×1 subsampling tier. Finally, a softmax function is applied to classify and predict normal and abnormal cells effectively. In addition, the team employed Z-score normalization for data pre-processing and utilized the LFCS(Levy Flight Cuckoo Search) optimization algorithm for effective gene selection.

Ausawalaithong et al. applied CNNs to process a gaint scale dataset of chest X-ray images for anomaly detection [10]. They investigated the model's performance using three retrained variants evaluated across different datasets, focusing on accuracy, sensitivity, and specificity. Model A, which was trained using the ChestX-ray14 dataset, was effective in detecting lung nodules. Model B, based on the JSRT dataset, achieved higher specificity but showed relatively lower accuracy and sensitivity. In contrast, Model C, which was trained on both ChestX-ray14 and JSRT, not only demonstrated more consistent performance with lower standard deviation but also accurately localized lung cancer. The authors concluded that repeated model retraining tailored to specific diagnostic tasks can enhance performance, especially in scenarios with limited data, as shown by the superior overall results of Model C.

Da Silva et al. employed a CNN architecture optimized using the Particle Swarm Optimization (PSO) algorithm [11]. To ensure fair model comparison across particles, training and validation were conducted using consistent datasets. The experimental data was sourced from the LIDC-IDRI database, and evaluation was carried out across five distinct test subsets. Among them, Test-1 achieved an precision of 96.54%, The degree of sensitivity is 87.79%, selectivity is 98.215%.

The study mentioned a CNN model for early lung cancer diagnosis with CT scan images [12]. The dataset, obtained from Kaggle, consisted of 967 labeled CT images classified divided into four groups: adenocarcinoma, large-cell carcinoma, squamous-cell carcinoma, and healthy tissue Images were pre-processed by resizing to 64×64 pixels, noise removal, segmentation, and morphological smoothing to enhance feature extraction. The CNN architecture included three convolutional tier having 16, 32, and 64 filters respectively, each attached by max-pooling tiers to decrease spatial dimensions. The extracted features were reduced to a single dimension and passed through a totaly connected tier with 260 factors, utilizing softmax activation to generate probabilistic class assignments in multiclass tasks. While intermediate layers leveraged ReLU activations, the output layer adopted distinct nonlinear processing. Employing Adam optimization and multiclass cross-entropy loss, the model was trained across 50 epochs using 13-sample batches.The CNN achieved a testing accuracy of 92%, recall of 91.72%, AUC of 98.21%, and a loss of 0.328.

The team of Mamun et al. mentioned a CNN-based model using the AlexNet structure for lung cancer diagnosis from the images of CT scan result collected from hospitals [13]. The model processes CT scans categorized three classes: normal, benign, and malignant. The dataset consisted 110 lung cancer CT images, 70% for drill and 30% for testing. AlexNet's architecture includes several convolutional tiers, max-pooling tiers, and totaly connected tiers, with images resized to 227 × 227 × 3 pixels. The study highlights the role of CNN layers in feature extraction through convolution, activation functions such as ReLU, pooling for downsampling, and classification via fully connected layers with softmax activation.

Han Li proposed FLE-CNN, an high-level CNN model designed to detect cancer in histopathology images [8]. This model incorporates a residual fusion unit to capture comprehensive contextual information and employs a dual-domain attention and information refinement mechanism. In a five-class cancer classification task, FLE-CNN outperformed other state of the art deep learning structure, achieving enhanced sensitivity, character, F1-score, and accuracy.

2.3 ViT-based Classification

Kumar, A. et al. proposed a VIT model for the classification of lung cancer diagnosis and colon disease using the LC25000 histopathological image dataset, which consist of 25,000 color images across 5 classes [14]. All the figures was split into training (80%) and testing (20%) subsets. The ViT model divides input images into fixed-size patches (16×16), which are linearly embedded and combined with positional embeddings before being processed through multiple transformer encoder layers incorporating multi-head self-attention, multilayer perceptrons, and layer normalization. The model used a frozen pre-trained transformer as a feature extractor with a new classification head added on top. Various hyperparameters, including batch size, patch size, number of epochs, and activation functions, were tuned, with experiments conducted using patch sizes of 4, 8, and 16, batch sizes of 16 and 32, and epochs ranging from 24 to 50. The best model configuration achieved a high accuracy, supported by strong quantitative evaluation metrics including precision (positive predictive value), recall (sensitivity), F1-score (harmonic mean), and ROC-AUC (area under receiver operating characteristic curve) [14]. strategies for expanding datasets including rotation, flipping, and zooming were also applied to enhance training. The study demonstrates that the ViT architecture, leveraging self-attention mechanisms instead of traditional convolutional layers, is highly effective for cancer diagnosis on images which collected from hospital.

Not only that, there are also studies presenting results after practical application, such as the retrospective analysis by Luoqi Wen et al. [15]. They published a retrospective study that analyzed lung adenocarcinoma patients patients who received CT-guided transthoracic biopsies or surgeries at the First Affiliated Hospital of Wenzhou Medical University from 2017 to 2022, focusing on predicting EGFR mutation status. A total of 525 patients meeting strict inclusion criteria—including pathological confirmation, a single malignant nodule per patient, and CT scans performed within one month of invasive procedures—were enrolled as the internal dataset. An external validation set comprising 30 patients was incorporated from the publicly available TCIA dataset, with slightly relaxed timing criteria due to limited sample size. Preprocessing of CT images involved scaling the annotated tumor ROIs to 224×224 pixels and normalizing their Hounsfield Unit (HU) values, followed by data augmentation (random rotation, flipping, zooming) to improve model robustness. A Vision Transformer model (ViT-B/16) pretrained on large-scale datasets was fine-tuned using transfer learning over

400 epochs with the Optimization was performed using Adam, with cross-entropy serving as the loss function. demonstrating strong generalization ability across datasets [15]. Grad-CAM provided visual attention maps highlighting tumor regions influential to the model's EGFR mutation predictions, potentially aiding clinical decision-making. Compared to traditional clinical and radiomics-based approaches, the deep learning model effectively captures complex spatial information without extensive feature engineering. The constraints of this study involve its single-center retrospective design, a limited number of cases, and the utilization of CT images with a slice thickness of 5 mm. Future work should explore larger multicenter cohorts and thinner-slice CT imaging to further enhance predictive accuracy. Overall, this study presents a promising non-invasive, CT image-fundimentaled ViT DL structure for accurate EGFR mutation status prediction in lung glandular carcinoma, with potential applications in personalized medicine.

3. Discussion

Although AI and machine learning have made great strides in lung cancer diagnosis, several critical challenges remain, mainly concentrated in three key areas: interpretability, applicability, and privacy. First, the lack of interpretability in many AI models hinders clinicians' trust and understanding, limiting their integration into routine clinical decision-making. Second, applicability issues arise from the variability in imaging protocols, limited generalization among heterogeneous patient groups and clinical contexts, and the scarcity of gaint, well-annotated datasets necessary for robust model training. Lastly, privacy concerns related to the sensitive nature of medical data pose significant barriers to data sharing and large-scale collaborative research, which are essential for improving AI performance and clinical adoption. Resolving these issues is critical to enabling AI to achieve its maximum impact in lung cancer diagnostics.

Despite the impressive performance of DL method in lung cancer detection, piece of the major unresolved challenges remains their lack of interpretability. Current models, such as CNNs and their variants, often function as „black boxes," providing little insight into how specific predictions are made. Although visualization approaches including saliency maps, Grad-CAM, and feature activation analysis have been applied to highlight areas of interest within medical images, these methods are often post hoc and may not align with the actual diagnostic reasoning of radiologists. Furthermore, the features extracted by deep learning models typically lack clear medical definitions, making it difficult for clinicians to validate or rely on these outputs. This gap between model predictions and clinically interpretable biomarkers limits the integration of AI into routine practice and hinders trust among healthcare professionals. As a result, developing intrinsically interpretable models and incorporating domain-specific medical knowledge into network architectures have become key research priorities in the field of medical imaging [16].

Deep learning-based lung cancer detection faces a critical challenge in the form of constrained model generalizability. Many models are trained on specific datasets that do not fully represent the diversity of patient populations and clinical environments, leading to decreased performance when applied to new or different settings. Furthermore, much of the research emphasizes detection accuracy while under-addressing tumor characterization, which is vital for early and precise diagnosis. The opaque "black box" nature of deep learning models further complicates clinical adoption by limiting transparency in decision-making processes. In addition, concerns regarding data security and patient privacy present substantial obstacles to data sharing and collaborative research efforts, which are essential for model improvement and validation. Addressing these issues through improved generalization techniques, robust privacy protections, and mitigation of dataset biases is essential for advancing AI applications in lung cancer diagnosis [17].

Future work in machine learning for lung cancer diagnosis should focus on integrating expert systems and domain knowledge to enhance model interpretability and clinical relevance. Additionally, federated learning offers a promising approach to preserve patient privacy while enabling collaborative training on distributed datasets from multiple institutions. By advancing these areas, machine learning can achieve more accurate, reliable, and privacy-conscious lung cancer detection, ultimately supporting earlier diagnosis and personalized treatment strategies in clinical practice.

In lung cancer diagnosis, incorporating domain expertise is essential, particularly for medical image analysis. Conventional approaches depend on manually engineered features—such as morphological characteristics, texture patterns, and contextual cues—that serve as important prior information to enhance detection accuracy [18]. Although deep learning excels at automatic feature extraction, it often faces challenges like overfitting and limited interpretability when applied to small medical datasets. By fusing expert knowledge with deep learning models, these limitations can be effectively addressed, leading to improved detection accuracy and robustness [18]. For example, incorporating shape features (e.g., HOG), texture descriptors (e.g., LBP and Haralick), and environmental context helps better distinguish benign from malignant nodules and reduces false positives [18]. Looking ahead, the combination of domain expertise and advanced machine learning

ISSN 2959-6157

is likely to be a key direction in lung cancer diagnosis, enabling more precise, reliable, and personalized healthcare solutions.

Federated learning has emerged as a critical solution for the future of lung cancer diagnosis within the evolving landscape of smart healthcare systems. By allowing collaborative model development across different centers without transmitting sensitive patient information, federated learning effectively mitigates privacy risks associated with medical data sharing. For instance, Abbas et al. proposed a fused weighted federated profound extreme ML approach integrated with edge computing, achieving a lung cancer prediction accuracy of 97.2%, surpassing state-of-the-art methods [19]. This approach not only ensures data privacy but also facilitates fast and efficient data transmission, making it highly suitable for Healthcare 5.0 environments [19]. As medical AI continues to develop, federated learning will be indispensable for balancing robust model performance with stringent data privacy requirements.

4. Conclusion

In summary, machine learning and artificial intelligence have substantially advanced lung cancer diagnosis by enabling the analysis of complex, high-dimensional medical data beyond human capability. Despite notable progress with deep learning models such as CNNs and Vision Transformers, challenges persist regarding interpretability, generalizability, and data privacy. The integration of expert knowledge into AI frameworks enhances model robustness and clinical applicability, while federated learning presents promising solutions to privacy concerns by facilitating collaborative training without direct data sharing. Future research should prioritize the development of interpretable models, improve adaptability across diverse clinical environments, and leverage privacy-preserving methodologies. Collectively, these efforts will promote more accurate, reliable, and personalized lung cancer diagnosis, ultimately improve patient outcomes and advance precision medicine in oncology.

References

[1] Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS. Lung cancer. Lancet, 2021, 398: 535–554.

[2] Sapoval N, Aghazadeh A, Nute MG, et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun, 2022, 13: 1728.

[3] Li J, Li Z, Wei L, et al. Machine Learning in Lung Cancer Radiomics. Mach Intell Res, 2023, 20: 753–782.

[4] Maurya SP, Sisodia PS, Mishra R, et al. Performance of machine learning algorithms for lung cancer prediction: a comparative approach. Sci Rep, 2024, 14: 18562.

[5] Riquelme D, Akhloufi MA. Deep learning for lung cancer nodules detection and classification in CT scans. Ai, 2020, 1: 28–67.

[6] Chiu HY, Chao HS, Chen YM. Application of artificial intelligence in lung cancer. Cancers, 2022, 14: 1370.

[7] Shatnawi MQ, Abuein Q, Al-Quraan R. Deep learning-based approach to diagnose lung cancer using CT-scan images. Intelligence-Based Medicine, 2025, 11: 100188.

[8] Bhatia I, Aarti, Ansarullah SI, Amin F, Alabrah A. Lightweight advanced deep neural network (DNN) model for early-stage lung cancer detection. Diagnostics (Basel), 2024, 14(21): 2356.

[9] Thangamani M, Koti M, BA N, et al. Lung cancer diagnosis based on weighted convolutional neural network using gene data expression. Sci Rep, 2024, 14: 3656.

[10] Ausawalaithong W, Thirach A, Marukatat S, Wilaiprasitporn T. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. 2018 11th Biomedical Engineering International Conference (BMEiCON), 2018.

[11] Da Silva GLF, da Silva Neto OP, Silva AC, de Paiva AC, Gattass M. Lung nodules diagnosis based on evolutionary convolutional neural network. Multimed Tools Appl, 2017, 76: 19039–19055.

[12] Al-Yasriy HF, AL-Husieny MS, Mohsen FY, Khalil EA, Hassan ZS. Diagnosis of lung cancer based on CT scans using CNN. IOP Conf Ser Mater Sci Eng, 2020, 928: 022035.

[13] Mamun M, Mahmud MI, Meherin M, Abdelgawad A. LCDctCNN: lung cancer diagnosis of CT scan images using CNN based model. arXiv preprint arXiv:2304.04814, 2023.

[14] Kumar A, Mehta R, Reddy BR, et al. Vision Transformer based effective model for early detection and classification of lung cancer. SN Comput Sci, 2024, 5: 839.

[15] Weng L, Xu Y, Chen Y, et al. Using Vision Transformer for high robustness and generalization in predicting EGFR mutation status in lung adenocarcinoma. Clin Transl Oncol, 2024, 26: 1438–1445.

[16] Wang L. Deep learning techniques to diagnose lung cancer. Cancers (Basel), 2022, 14(22): 5569.

[17] Javed R, Abbas T, Khan AH, et al. Deep learning for lungs cancer detection: a review. Artif Intell Rev, 2024, 57: 197.

[18] Tan J, Huo Y, Liang Z, Li L. Expert knowledge-infused deep learning for automatic lung nodule detection. J Xray Sci Technol, 2019, 27(1): 17–35.

[19] Abbas S, Issa GF, Fatima A, Abbas T, Ghazal TM, Ahmad M, Yeun CY, Khan MA. Fused weighted federated deep extreme machine learning based on intelligent lung cancer disease prediction model for Healthcare 5.0. Int J Intell Syst, 2023: Article ID 2599161.