Proceedings of the 3rd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/89/2025.21453

Application of Machine Learning in Lung Disease Detection: A Review

Taiyang Chen[1,a,*]

[1]*Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, United States*
a. chentaiyang1102@gmail.com
**corresponding author*

Abstract: Lung disease remains the leading cause of global mortality, which requires efficient and accurate methods for detection. Recent advancements in machine learning algorithms, artificial intelligence and image analysis technologies offer potential for early diagnosis and improved patient outcomes. This review summarizes different types of medical images that can be or have been frequently used in lung disease diagnosis, such as chest X-rays and CT scans. The paper then explores the role of key machine learning techniques in detecting various lung diseases, such as lung cancer, pneumonia and tuberculosis (TB), assessing the applications, strengths, and limitations of Support Vector Machines (SVM), Random Forest (RF), and Convolutional Neural Networks (CNN) in lung disease diagnosis. Despite these advancements, this paper also discusses potential challenges in using machine learning and artificial intelligence for lung disease detection, highlighting problems and limitations like interpretability, computational cost and ethical concerns that must be addressed in the future.

Keywords: Machine Learning, Lung Disease, Healthcare AI

1. Introduction

Lung diseases are one of the leading causes of death worldwide. About 1.5 million deaths are caused by tuberculosis every year, especially in resource-restricted countries [1]. In 1987, lung cancer ranks as one of the top causes of cancer-related deaths among women, and by 2000, it was estimated to make up 25% of all female cancer deaths [2]. In 2018, community-acquired pneumonia caused the deaths of 0.8 million children worldwide [3]. The SARS-CoV-2 virus can cause symptoms such as pneumonia and inflammation in the lung, and more than 6.8 million deaths were reported globally over three years into the COVID-19 pandemic [4].

Early and accurate detection of lung diseases is crucial in improving patient outcomes, as it allows for timely intervention and treatment. Traditional methods for lung disease detection include skin test, blood test and sputum sample test. Medical images, especially from high-resolution scans like computed tomography (CT), allow for precise visualization of lung structure, helping detect subtle abnormalities in tissues that may indicate disease [5, 6]. Recent advances in machine learning and deep learning play an important role in efficient detection of lung diseases. They leverage vast amounts of patient data to build predictive models which can assist clinicians in lung diseases diagnosis. For example, Support Vector Machines (SVM), Random Forest (RF), and Convolutional Neural Networks (CNN) have been used to analyze a variety of medical data including images, greatly

DOI: 10.54254/2753-8818/89/2025.21453

improving the efficiency of lung disease diagnosis. This review explores how different machine learning algorithms are applied to medical images for detecting lung conditions, highlighting their benefits. This article also discusses the potential computational and ethical drawbacks of machine learning in lung disease detection and provides some directions for future research.

This study reviews the application of machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), and Convolutional Neural Networks (CNN), in detecting lung diseases from medical images. It examines how these techniques improve diagnostic accuracy and efficiency with imaging modalities like chest X-rays, CT, and PET-CT scans. Additionally, it addresses the computational and ethical challenges associated with using AI in healthcare and suggests directions for future research to enhance performance, interpretability, and fairness in lung disease diagnostics.

2. Types of Medical Images in Lung Disease Detection

Chest X-rays are widely used as an initial imaging modality for lung disease due to their quick acquisition, cost-effectiveness, and broad availability. They contain features associated with diseases. For example, distinct findings can be observed in X-ray images of COVID-19 infections even in the early stages of infection, especially in the lower lobes and posterior segments, showing peripheral and subpleural distribution. Haritha et al. developed a COVID-19 detection model based on the ChestX-ray14 dataset which consists of more than 100k X-ray images labeled with 14 distinct respiratory diseases [7].

Computed Tomography (CT) scans provide detailed cross-sectional images of the lungs. CT images of lung cancer tumors reveal a high contrast that highlights variations in tumor intensity, internal texture, and shape [8]. CT scans generate cross-sectional images, allowing for high-resolution visualization of lung structures, which makes them highly effective in detecting subtle or small lesions that might be missed on X-rays. For instance, CT images of lung cancer tumors reveal a high contrast that highlights variations in tumor intensity, internal texture, and shape. This high contrast, coupled with the resolution of CT scans, makes it possible to observe even minor abnormalities, assess tumor margins, and evaluate characteristics like spiculation, calcification, and pleural involvement.

Positron Emission Tomography (PET) is another type of medical imaging used in lung disease detection, often combined with CT in a PET-CT scan. PET imaging is particularly valuable for assessing lung cancer, as it provides metabolic information that reveals how active a tissue is at the cellular level. In PET, a radioactive tracer, typically fluorodeoxyglucose (FDG), is injected into the patient. Cancer cells, which consume more glucose than normal cells due to their rapid growth, absorb more of this tracer, allowing PET imaging to highlight areas of high metabolic activity [9]. PET-CT scans are widely used in lung cancer staging and treatment planning [10, 11]. While CT scans provide detailed anatomical images, PET adds a layer of functional imaging by displaying “hot spots” of metabolic activity where cancer cells are more likely to be found. This dual capability is especially useful for detecting metastasis, as it can reveal cancer spread to lymph nodes or other organs that may not be anatomically enlarged but show increased metabolic activity. PET-CT has also proven useful in assessing the effectiveness of treatment. By comparing scans before and after treatment, oncologists can gauge changes in tumor metabolic activity, helping to determine whether the tumor is responding to therapy.

3. Machine Learning Algorithms and Applications

3.1. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm primarily used for classification tasks. It works by finding the optimal hyperplane that separates data points into different

DOI: 10.54254/2753-8818/89/2025.21453

classes, ensuring that the margin between the classes is maximized. SVM can handle complex datasets, making it a popular choice for a variety of medical data related classification problems, such as disease detection.

In previous studies, Mahdy et al. developed an automatic COVID-19 lung X-ray image classification system by combining multi-level thresholding and SVM. The authors achieved high performance, with an average accuracy of 97.48%, sensitivity of 95.76%, and specificity of 99.7% [12]. The study carried out by Widodo et al. focuses on several key areas: developing software for automatic lung segmentation using the Active Shape Model, segmenting diseased areas with Morphology Mathematics, and detecting lung diseases via SVM [13]. The software can identify abnormalities in CT scans with unclear lung edges and small abnormalities. Other studies combined SVM and feature extraction including texture and statistical features from CT images, and their models also achieved accuracy, sensitivity, and specificity greater than 90% [14, 15].

These studies show SVM has a great potential of classifying lung images. When combined with proper methods for data preprocessing and feature extraction, SVM can achieve almost perfect performance. Moreover, the mechanism of SVM is relatively simple, which can help prevent overfitting.

3.2. Random Forest

Random Forest is a machine learning algorithm that operates by constructing multiple decision trees during training and outputting the class that is the majority vote of the individual trees. It can be used to analyze features extracted from imaging data, such as CT scans or X-rays, to identify patterns indicative of diseases.

Liu et al. developed a fully automatic segmentation algorithm which used random forest to classify the superpixel features extracted from CT lung images. Random forest successfully classified lungs, thoracic cavity excluded lungs, and image background, resulting in high accuracy on the segmentation of lungs affected with diseases [16]. It can provide great assistance for lung disease detection.

Akgundogdu performed 2D discrete wavelet transform for feature extraction, which decomposed X-ray images into 24 statistical features in four categories. A random forest classifier was applied to distinguish normal and pneumonia images. The model has an accuracy of 97.11%, sensitivity of 91.79% and specificity of 99.09%. In 10-fold cross-validation, the area under the curve (AUC) is 0.99, which means the model has good performance [17].

3.3. Convolutional Neural Network

Neural Networks are a class of machine learning algorithms inspired by the structure and function of the human brain. They are particularly effective at identifying patterns and relationships in complex data. Among Neural Networks, Convolutional Neural Networks (CNNs) are a type of deep learning model specifically designed to process and analyze visual data, making them highly effective for image classification tasks. CNNs use a series of convolutional layers to automatically extract and learn features from input images, such as edges, textures, and shapes, which are crucial for accurate image analysis. By learning complex patterns directly from the image data, CNNs can achieve high accuracy in diagnosing lung conditions, often outperforming traditional machine learning models.

In past studies, Ahsan et al. proposed a hybrid MLP-CNN model. The CNN model contains three convolutional layers with ReLU activation, followed by three max-pooling layers, which is useful for extracting features from the X-ray images. The authors also included a Multilayer Perceptron (MLP) to process numerical/categorical data. Their method achieved accuracy around 95% in multiple experiments of distinguishing COVID-19 images from non-COVID-19 X-ray images [18].

DOI: 10.54254/2753-8818/89/2025.21453

Ashwini et al. used grid search optimization to create the optimal CNN model architecture and set its hyperparameters. Their model not only distinguished normal lung from lung disease, but also classified various diseases including COVID-19, lung cancer and pneumonia [19].

Training a model from scratch can be time-consuming. Transfer learning is a technique where a pretrained model, originally trained on a large dataset, is fine-tuned on a new, smaller dataset. By leveraging a model already trained on a general task, such as image classification, transfer learning enables the model to adapt to a more specific task like lung disease detection, without needing vast amounts of new data.

Kido at al. developed their model using a pre-trained CNN, specifically AlexNet, which was trained on the ImageNet dataset containing 1.2 million training images in 1,000 object categories. The CNN was first used to extract features from the training images, and these features were then input into a multiclass SVM for classification. Additionally, they employed the R-CNN, a framework using a CNN to classify image regions within an image, to detect abnormal regions in the images. The model successfully classified various kinds of lung abnormalities such as benign and malignant lung nodules [20].

Koravanavar and Tapaskar used Mobilenet V2, an image classification model trained using large images dataset. The MobilenetV2 model achieved 97% accuracy when classifying the normal and pneumonia chest X-ray data and achieved 100% accuracy for tuberculosis [21].

Lakhani and Sundaram evaluated the performance of AlexNet and GoogLeNet CNNs on chest radiographic images from pulmonary TB patients and healthy people. They found that the CNNs pretrained with the ImageNet dataset achieved better performance compared with untrained CNNs, concordant with previously published works, which is consistent with previously published works [22].

In summary, neural networks like CNN can process vast amounts of complex image data, greatly improving the efficiency of lung disease diagnosis. In addition, they can be adjusted to specific tasks by training them on different data, such as images of different lung diseases.

4. Challenges and Future Directions

Machine learning and deep learning approaches have some limitations. Random forest and support vector machine relies on feature selection from input data. Selecting wrong features or irrelevant information may negatively impact the model's performance. Some previous studies use complex pipelines to perform feature selection and classification, which requires much development work [23]. In addition, neural network algorithms can be computationally expensive, which causes challenges when they are applied to real-world clinical settings [24, 25].

Apart from the potential computational challenges above, using machine learning and artificial intelligence to diagnose lung diseases also has some ethical issues. Many neural network algorithms act as "black boxes", making it difficult to interpret their decision-making processes and thus reducing trust. If the model is trained on a dataset that lacks diversity, it may perform poorly or produce biased results for the underrepresented groups, leading to inequalities in healthcare [26]. Privacy and security related to data storage may also have problems [27]. Breach or misuse of image data along with their relevant medical records can lead to serious consequences for patients.

To address these challenges, Pasa et al. developed a simple and fast model for tuberculosis diagnosis consisting of 5 convolutional blocks followed by a global average pooling layer. Compared to previous publications, their model preserved high accuracy but reduced computational resources. Moreover, visualization results highlighted the areas where tuberculosis was present, which means the decision of the CNN aligns with how human experts tell whether a patient has tuberculosis or not [24].

Inspired by the studies above, some possible future directions are as follows. First, it is important to reduce the computational cost of model training and application, which can be done by adjusting model design and applying new model architectures. Additionally, to improve model performance and interpretability, further research is required to analyze the features used by the machine learning models for decision making. Finally, de-identified data with better diversity should be used in future research to reduce the ethical risks of using machine learning and AI to detect diseases based on images.

5. Conclusion

In conclusion, the application of machine learning techniques, particularly SVM, RF and CNN, has revolutionized lung disease detection using medical imaging. Different methods offer distinct advantages: SVM and RF can achieve good accuracy, sensitivity, and specificity, while CNNs demonstrate exceptional capability in complex tasks. Together, these tools enhance the accuracy and efficiency of detecting lung diseases such as pneumonia and TB, promising early and accurate diagnoses that could transform patient outcomes. However, interpretability, computational cost and ethical considerations remain major concerns, which require thoughtful resolution as machine learning becomes more embedded in healthcare.

References

[1] *Nordholm, A. C., Andersen, A. B., Wejse, C., Norman, A., Ekstrøm, C. T., Andersen, P. H., ... & Koch, A. (2023). Mortality, risk factors, and causes of death among people with tuberculosis in Denmark, 1990-2018. International Journal of Infectious Diseases, 130, 76-82.*
[2] *Greenlee, R. T., Murray, T., Bolden, S., & Wingo, P. A. (2000). Cancer statistics, 2000. CA: a cancer journal for clinicians, 50(1), 7-33.*
[3] *Awasthi, S., Pandey, A. K., Mishra, S., & CAP Study Group. (2023). Identifying risk of death in children hospitalized with community-acquired pneumonia. Bulletin of the World Health Organization, 101(4), 281.*
[4] *Toussi, S. S., Hammond, J. L., Gerstenberger, B. S., & Anderson, A. S. (2023). Therapeutics for COVID-19. Nature Microbiology, 8(5), 771-786.*
[5] *Kieu, S. T. H., Bade, A., Hijazi, M. H. A., & Kolivand, H. (2020). A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions. Journal of imaging, 6(12), 131.*
[6] *Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., Van Den Bogaard, C., Cerello, P., ... & Jacobs, C. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical image analysis, 42, 1-13.*
[7] *Haritha, D., Pranathi, M. K., & Reethika, M. (2020, October). COVID detection from chest X-rays with DeepLearning: CheXNet. In 2020 5th international conference on computing, communication and security (ICCCS) (pp. 1-5). IEEE.*
[8] *Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., ... & Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications, 5(1), 4006.*
[9] *Saif, M. W., Tzannou, I., Makrilia, N., & Syrigos, K. (2010). Role and cost effectiveness of PET/CT in management of patients with cancer. The Yale journal of biology and medicine, 83(2), 53.*
[10] *Ambrosini, V., Nicolini, S., Caroli, P., Nanni, C., Massaro, A., Marzola, M. C., ... & Fanti, S. (2012). PET/CT imaging in different types of lung cancer: an overview. European journal of radiology, 81(5), 988-1001.*
[11] *Schwyzer, M., Ferraro, D. A., Muehlematter, U. J., Curioni-Fontecedro, A., Huellner, M. W., Von Schulthess, G. K., ... & Messerli, M. (2018). Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks–initial results. Lung Cancer, 126, 170-173.*
[12] *Mahdy, L. N., Ezzat, K. A., Elmousalami, H. H., Ella, H. A., & Hassanien, A. E. (2020). Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine. MedRxiv, 2020-03.*
[13] *Widodo, S., Rohmah, R. N., Handaga, B., & Arini, L. D. D. (2019). Lung diseases detection caused by smoking using support vector machine. TELKOMNIKA (Telecommunication computing electronics and control), 17(3), 1256-1266.*

[14] *Prabukumar, M., Agilandeeswari, L., & Ganesan, K. (2019). An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier. Journal of Ambient Intelligence and Humanized Computing, 10, 267-293.*
[15] *Vijh, S., Gaur, D., & Kumar, S. (2020). An intelligent lung tumor diagnosis system using whale optimization algorithm and support vector machine. International Journal of System Assurance Engineering and Management, 11(2), 374-384.*
[16] *Liu, C., Zhao, R., & Pang, M. (2020). A fully automatic segmentation algorithm for CT lung images based on random forest. Medical Physics, 47(2), 518-529.*
[17] *Akgundogdu, A. (2021). Detection of pneumonia in chest X-ray images by using 2D discrete wavelet feature extraction with random forest. International Journal of Imaging Systems and Technology, 31(1), 82-93.*
[18] *Ahsan, M. M., E. Alam, T., Trafalis, T., & Huebner, P. (2020). Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients. Symmetry, 12(9), 1526.*
[19] *Ashwini, S., Arunkumar, J. R., Prabu, R. T., Singh, N. H., & Singh, N. P. (2024). Diagnosis and multi-classification of lung diseases in CXR images using optimized deep convolutional neural network. Soft Computing, 28(7), 6219-6233.*
[20] *Kido, S., Hirano, Y., & Hashimoto, N. (2018, January). Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN). In 2018 International workshop on advanced image technology (IWAIT) (pp. 1-4). IEEE.*
[21] *Koravanavar, M. R., & Tapaskar, R. P. (2024, June). Chest X-ray Based Pulmonary Disease Classification Using Transfer Learning and CNN. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (pp. 1-6). IEEE.*
[22] *Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 284(2), 574-582.*
[23] *Vajda, S., Karargyris, A., Jaeger, S., Santosh, K. C., Candemir, S., Xue, Z., ... & Thoma, G. (2018). Feature selection for automatic tuberculosis screening in frontal chest radiographs. Journal of medical systems, 42, 1-11.*
[24] *Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., & Pfeiffer, D. (2019). Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. Scientific reports, 9(1), 6268.*
[25] *Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), 1285-1298.*
[26] *Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences, 117(23), 12592-12594.*
[27] *Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.*