IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

# Lung cancer detection: A systematic literature study

Zaidan Mufaddhal *

*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, China.*

International Journal of Science and Research Archive, 2024, 13(02), 3184-3191

Publication history: Received on 13 November 2024; revised on 21 December 2024; accepted on 23 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2565

## Abstract

Lung cancer is the primary cause of cancer-related deaths. Lung cancer presents with symptoms only in its advanced stages. Machine Learning and Deep Learning can be used to detect lung cancer early. This study aims to find the state-of-the-art approach to detecting lung cancer. The topic at hand has both potential and challenges, which are highlighted by the diversity of datasets, model architectures, and methodological approaches. Interestingly, the incorporation of image data turns out to be crucial for practical uses, highlighting the shortcomings of models trained in the absence of such data. It becomes clear how important dataset size is, with larger datasets potentially providing benefits in terms of model robustness. While certain models such as CNN VGG-19, LCP-CNN, and FPSOCNN perform admirably, they also highlight subtle issues, like the requirement for refinement in order to properly categorize nodules and the expense of computing. Future research directions are informed by the identification of these strengths and limits, which highlight the necessity for customized optimizations and the taking into account of real-world constraints.

**Keywords:** Lung Cancer Detection; Machine Learning; Deep Learning; Systematic Review

## 1. Introduction

Globally, lung cancer is the primary cause of cancer-related deaths. Lung cancer consequently emerges as a serious public health issue. Lung cancer continues to be the most common cause of cancer-related death, accounting for 1.8 million deaths (18%) in 2020, according to the International Agency for Research on Cancer's (IARC) GLOBOCAN 2020 [1] estimates of cancer incidence and mortality. In their lifetime, 1 in 16 men and 1 in 17 women will receive a lung cancer diagnosis, according to the American Cancer Society 2023 [2].

Generally speaking, cancer cells often migrate to the middle of the chest as a result of normal lymph flow. The spread of cancer cells to other tissues is known as metastasis [3]. Since cancer tends to spread and becomes incurable in the event of a larger spread, early detection is crucial. Lung cancer presents with symptoms only in its advanced stages when survival is almost impossible and diagnosis is challenging.

Certain features must be identified and measured in order to identify malignant nodules. Cancer probability can be determined by combining the features that have been identified. Even for a skilled doctor, this task is exceedingly challenging because there is a difficult correlation between the presence of a nodule and a positive cancer diagnosis [3].

To address this, many works suggest utilizing Machine Learning (ML) and Deep Learning (DL) techniques in detecting lung cancer early, such as Support Vector Machine (SVM) [4], Logistic Regression (LR) [5], Convolutional Neural Network (CNN) [6], K-Nearest Neighbor (KNN) [7], Artificial Neural Network (ANN) [8], Random Forest (RF) [8], and many other algorithms. However, with the rapid development of ML, it is important to address the issue of finding which method has the highest considerable performance in detecting lung cancer.

* Corresponding author: Zaidan Mufaddhal
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, China.

This study aims to find the state-of-the-art approach to detecting lung cancer through reviewing the current related studies with a systematic review as suggested in [9]. This study evaluates and compares each method proposed in the existing related literatures. The comparison is based on its methods and performance. Also, this study considers the dataset used and the type of feature extraction performed in the reviewed papers, as it played an important role in defining the model's performance. Through comparison and evaluation, this study aims to define the strengths and weaknesses of each method. This study is limited to lung cancer detection using ML and DL approaches.

## 2. Research Method

In order to investigate studies pertaining to lung cancer detection, the systematic mapping study described in [9] is chosen. As shown in Figure 2.1, the systematic mapping study's methodology consists of five steps.

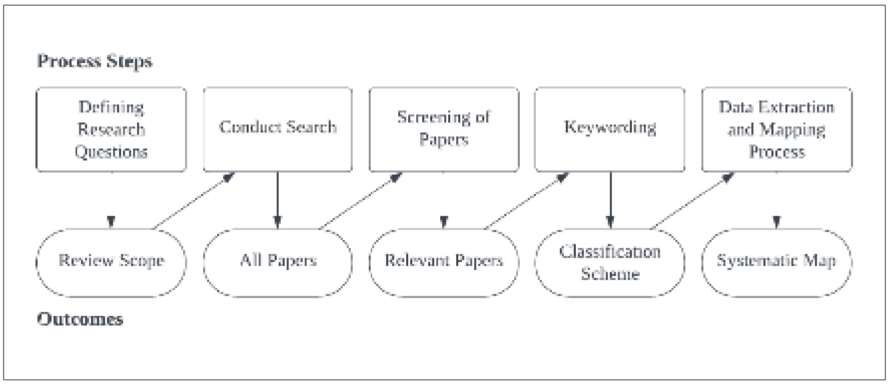

**Figure 1** Systematic Mapping Studies

A further explanation of the method is covered in the following subsections.

### 2.1. Defining Research Questions

The main exploration questions are based on the importance of investigating studies about lung cancer detection. The research questions are portrayed as follows:

- RQ1: What research approaches do the existing studies apply to detect lung cancer?
- RQ2: What kinds of factors should be taken into account while developing a model for detecting lung cancer?
- RQ3: Which strategies worked the best, and what strengths and weaknesses did those approaches have?

Thereupon, the outline of this study will provide the appropriate response to lung cancer detection, as well as point out the areas that need further investigation.

### 2.2. Conduct Search

This step involves searching for and locating all scientific papers pertaining to the research topic—lung cancer detection. We chose the term "lung cancer" as the primary search term for our study in order to find relevant papers. During the search, this term was associated with the term “detection”, or “classification”; and also, with “Machine Learning”, or “Deep Learning”. Google Scholar is selected as a searching tool for published papers. The selected papers is required to be published from high-quality publishers, such as IEEE, ScienceDirect, Springer, IOP Science, Nature, ResearchGate, Hindawi, and MDPI.

### 2.3. Screening of Papers

This step involves looking for papers that address the issues raised by our research questions. First, papers considered irrelevant by their title are disregarded. If there is any doubt, the abstract is examined as the next action to be taken. The following is a further explanation of the exclusion criterion: (1) papers written in a language other than English; (2) papers that use machine learning or deep learning for purposes other than detecting lung cancer; (3) duplicate papers; and (4) newsletters and grey literature (5) papers that are published more than 5 years ago.

### 2.4. Keywording using Abstracts

Using their keywords, all relevant papers are categorized in this step. First, each paper's abstract is examined to determine the most crucial terms and the primary contribution. The papers were then divided into different categories using those keywords.

### 2.5. Data Extraction and Mapping Process

The purpose of this method is to collect all the data needed to answer the study's research questions. The data extracted are the year of publication, authors, titles, methodologies applied, performance results, datasets used, and type of feature extractions used. The major goals and contributions of the studies are embraced by these data.

## 3. Results

### 3.1. Research Approaches

Table 1 is a carefully crafted table to answer RQ1, which is "What research approaches do the existing studies apply to detect lung cancer?", it includes 20 different research, making it easier to compare a selection of the medical image analysis literature. This table provides an overview of the many methods used to classify lung cancer and summarizes important information that is necessary for a more in-depth analysis of the methods' effectiveness. A multidimensional evaluation is made possible by the addition of columns like Reference, Method, Architecture, Feature Extraction, Accuracy, Specificity, Sensitivity, and AUC. Because the method, architecture, and feature extraction play such crucial roles in influencing the performance of the classification models [6], it is necessary to carefully examine them in the comparison study. The architectural design affects the model's ability to capture complex patterns, the methodological approach establishes the overall strategy, and the feature extraction technique determines the features' discriminative power [6]. These factors are crucial in determining the advantages and disadvantages of every approach.

A comprehensive evaluation also requires careful consideration of the performance indicators that are used, such as Area Under the Curve (AUC), Specificity, Sensitivity, and Accuracy [10], [11]. Specificity assesses the model's capacity to accurately identify situations that are not cancer, whereas accuracy indicates the model's overall robustness. Sensitivity measures the model's ability to correctly identify positive cases, which is important when making a medical diagnosis. Finally, AUC, or the Area Under the Curve, offers information about how well the model can differentiate between different classes. When taken as a whole, these performance indicators provide a thorough grasp of the advantages and disadvantages of every strategy, assisting scholars and professionals in reaching well-informed conclusions about the suitability and resilience of different lung cancer classification systems.

### 3.2. Dataset Used

Table 2 is a brief but thorough table that summarizes the comparison of several datasets used in the selected articles related to the detection of lung cancer. This table, which includes columns for Reference, Dataset, Type of Data, and Number of Samples, is essential for deciphering the complexities involved in selecting data sources for different types of research. Comprehending the nature of the data, particularly if it includes CT scan images or not, is essential to comprehend the variety and complexity of the methods of imaging that are being studied [21]. The inclusion of this data clarifies whether the suggested models can be applied to various forms of imaging and are generally applicable.

Furthermore, a key factor in determining the statistical significance and robustness of the results is the quantity of samples included in each dataset [24]. An inadequate sample size in a dataset could lead to biased or overfitted results, which could reduce the suggested classification models' external validity. Thus, researchers can navigate the complex landscape of lung cancer classification studies by carefully examining the type of data and the number of samples in each dataset. This allows for a cautious evaluation of the methodologies used and makes it easier to identify trends and patterns across a variety of datasets.

**Table 1** Comparison of methods for lung cancer detection and their performances

| Reference | Method | Architecture | Feature Extraction | Accuracy | Specificity | Sensitivity | AUC |
|---|---|---|---|---|---|---|---|
| [12] | Fuzzy Particle Swarm Optimization Convolution Neural Network (FPSOCNN) | | Fuzzy Particle Swarm Optimization (FPSO) | 95.62% | 96.32% | 97.93% | |
| [5] | Logistic Regression (LR) | | Principal Component Analysis (PCA) 70% | | 74.3% | 69.9% | 72.5% |
| [5] | Elastic Net | | PCA 70% | | 71.5% | 61.1% | 73.3% |
| [5] | Linear Support Vector Machine (SVM) | | PCA 70% | | 68.5% | 61.5% | 72.7% |
| [4], [13] | SVM | | | | 97.1% | 31.1% | 64.2% |
| [14], [15] | SVM | | Gray Level Co-occurrence Matrix (GLCM) | 90.9% | 61.5% | 86% | |
| [7] | SVM | | Local Binary Pattern (LBP) and discrete cosine transform (DCT) | 93% | | | |
| [4], [8] | K-Nearest Neighbors (KNN) | | | 76.5% | 85.7% | 56.3% | 71% |
| [7] | KNN | | LBP and DCT | 91% | | | |
| [4], [16] | Convolutional Neural Network (CNN) | VGG-16 | | 68.6% | 82.9% | 37.5% | 70.1% |
| [6] | CNN | AlexNet | Deep Feature Extraction | 99.52% | | | |
| [10] | CNN | ResNet-101 | Radiomics & Conv2d | 56% | 58% | 53% | |
| [10], [11] | CNN | Inception-v3/GoogLeNet | | 94.53% | 99.06% | 65.67% | 86.84% |
| [10] | CNN | Inception-ResNet-v2 | Radiomics & Conv2d | 64% | 51% | 69% | |
| [17] | CNN | DenseNet | | 90.85% | | | |
| [18] | CNN | VGG-19 | | 98.05% | | | 99.66% |
| [8] | Random Forest (RF) | | Linear discriminant analysis (LDA) | | | | |
| [8] | Artificial Neural network (ANN) | | LDA | | | | |
| [19] | Rotation Forest | | | 97.1% | | | 99.3% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [20], [10] | Hybrid Convolutional Neural Network and Long-Short Term Memory (CNN-LSTM) | | Conv2d | 97% | | | |
| [21] | Modified Gravitational Search Algorithm (MGSA) | | LDA | 94.56% | 94.2% | 96.2% | |
| [22] | Kernel Attribute Selected Classifier (KASC) | | SURF and GA algorithm | 98.18% | | | |
| [23] | Medial Early Sign (MES) | | | | 95% | 40.1% | 85.6% |
| [24] | Lung Cancer Prediction Convolutional Neural Network (LCP-CNN) | | | | | 99% | 94.5% |

**Table 2** Comparison of dataset applied for the model development in lung cancer detection

| Reference | Dataset | Type of data | No. of Samples |
|---|---|---|---|
| [8], [11], [12], [20] | The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) | CT scan images and XML file | 1018 |
| [5], [10] | NSCLC-Radiogenomics | CT scan, PET/CT scan images | 211 |
| [4] | Massachusetts General Hospital (MGH) | CT scan images | 335 |
| [19] | Kaggle Lung Cancer Dataset | CSV file | 284 |
| [7] | Kaggle Chest CT Scan images | CT scan images | 1000 |
| [6], [13], [15], [16] | Amrita Institute of Medical<br>Sciences Hospital (AIMS) Kochi | CT scan and PET-CT scan images | 200 |
| [21], [22] | VIA/I-ELCAP | CT scan images | 100 |
| [23] | KPSC tumor registry & American Joint Commission | CSV file | 834 |
| [24] | U.S. National Lung Screening Trial (NLST) | CT scan images | 10,368 |

The literature on the detection of lung cancer provides a wide range of information and methods. The comparative tables highlight the significant keynote on the performance of each method. The comparative analysis of each method is explained in more detail in the section that follows.

## 4. Discussion

In this section, the answer for RQ2 “What kinds of factors should be taken into account while developing a model for detecting lung cancer?” and RQ3 “Which strategies worked the best, and what strengths and weaknesses did those approaches have?” are elaborated thoroughly with the consideration of the results from the systematic mapping studies.

Analyzing different datasets used in lung cancer detection models provides fascinating insights into the complexities and difficulties present in real-world applications. Specifically, the decision to include or exclude image data during model training places a crucial restriction on models developed without image data, making it more difficult for them to handle image inputs in real-world situations. This emphasizes how important it is to use datasets that contain picture data. This is an important consideration when using lung cancer detection models in the real world, even though it has minimal impact on performance differences in experimental settings.

One important component influencing the development of models that also comes into focus is the consideration of dataset instances. Table 2 shows that the NLST dataset is underutilized in the evaluated literature, even with its large 10,368 instances. But the single use of the NLST dataset, in the LCP-CNN model, shows superior performance in terms of sensitivity and AUC compared to other approaches. This result emphasizes how bigger datasets can enhance the performance and robustness of models.

The LIDC-IDRI and AIMS datasets are often cited, but there is a significant difference in sample sizes between them, according to a review of citation patterns. Even though the CNN-LSTM model uses LIDC-IDRI to obtain the best accuracy, a thorough examination is hampered by the model's inability to provide anything other than accuracy results. Furthermore, FPSOCNN exhibits noteworthy performance, regularly achieving above 95% accuracy, specificity, and sensitivity using the same dataset.

Additionally, the assessment of CNN algorithms indicates that architecture and feature extraction play crucial roles in defining performance. In particular, the VGG-19 architecture shows remarkable performance in lung cancer detection, illustrating the impact of these design decisions on model effectiveness.

The analysis of certain models, such as CNN VGG-19, LCP-CNN, and FPSOCNN, clarifies their advantages and disadvantages, enabling future study directions for lung cancer detection approaches to be decided upon with knowledge. However, although CNN VGG-19 has excellent accuracy and AUC, its high computing requirements prevent it from being used in real-world scenarios with limited resources. The accuracy of the size-based nodule classification in LCP-CNN, which is intended for lung cancer diagnosis, has to be improved. Furthermore, FPSOCNN consistently demonstrates the need for additional optimization to resolve any potential shortcomings in its categorization paradigm. The intricacy of developing models is highlighted by these model-specific subtleties, which also highlight the necessity of focused improvements to improve effectiveness and dependability.

In the final analysis, the discussion highlights the need to carefully choose datasets, highlighting the value of image data and the possible benefits of larger datasets for the development of models. The evaluation of CNN VGG-19, LCP-CNN, and FPSOCNN, reveals their exceptional capabilities, making it possible to make informed decisions about future research directions for lung cancer detection techniques. To connect this paper with future research for improved lung cancer detection techniques, factors including computing cost, nodule size detection, and optimization difficulties offer insightful information for improving current models and motivating the creation of more reliable and effective techniques.

## 5. Conclusion

In conclusion, this comprehensive review and analysis of lung cancer detection methods highlights the complex field of digital image processing in medical diagnostics. The topic at hand has both potential and challenges, which are highlighted by the diversity of datasets, model architectures, and methodological approaches. Interestingly, the incorporation of image data turns out to be crucial for practical uses, highlighting the shortcomings of models trained in the absence of such data. It becomes clear how important dataset size is, with larger datasets potentially providing benefits in terms of model robustness. While certain models—such as CNN VGG-19, LCP-CNN, and FPSOCNN—perform admirably, they also highlight subtle issues, like the requirement for refinement in order to properly categorize nodules and the expense of computing. Future research directions are informed by the identification of these strengths and limits, which highlight the necessity for customized optimizations and the taking into account real-world constraints.

## References

[1] International Agency for Research on Cancer, “GLOBOCAN Lung Cancer Facts Sheet 2020,” 2020.

[2] American Cancer Society, “Cancer Facts and Figures 2023,” Atlanta, 2023.

[3] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, “Global Epidemiology of Lung Cancer,” *Ann Glob Health*, vol. 85, no. 1, Jan. 2019, doi: 10.5334/aogh.2419.

[4] T. L. Chaunzwa *et al.*, “Deep learning classification of lung cancer histology using CT images,” *Sci Rep*, vol. 11, no. 1, p. 5471, Mar. 2021, doi: 10.1038/s41598-021-84630-x.

[5] J. Morgado *et al.*, “Machine Learning and Feature Selection Methods for EGFR Mutation Status Prediction in Lung Cancer,” *Applied Sciences*, vol. 11, no. 7, p. 3273, Apr. 2021, doi: 10.3390/app11073273.

[6] R. R. Subramanian, R. Mourya, V. Prudhvi, T. Reddy, B. Reddy, and S. Amara, “Lung Cancer Prediction Using Deep Learning Framework,” *International Journal of Control and Automation*, vol. 13, pp. 154–160, 2020.

[7] A. Rehman, M. Kashif, I. Abunadi, and N. Ayesha, “Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques,” in *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, IEEE, Apr. 2021, pp. 101–104. doi: 10.1109/CAIDA51941.2021.9425269.

[8] S. Nageswaran *et al.*, “Lung Cancer Classification and Prediction Using Machine Learning and Image Processing,” *Biomed Res Int*, vol. 2022, pp. 1–8, Aug. 2022, doi: 10.1155/2022/1755460.

[9] K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” *Inf Softw Technol*, vol. 64, pp. 1–18, Aug. 2015, doi: 10.1016/j.infsof.2015.03.007.

[10] P. Marentakis *et al.*, “Lung cancer histology classification from CT images based on radiomics and deep learning models,” *Med Biol Eng Comput*, vol. 59, no. 1, pp. 215–226, Jan. 2021, doi: 10.1007/s11517-020-02302-w.

[11] S. M. Ashhar *et al.*, “Comparison of deep learning convolutional neural network (CNN) architectures for CT lung cancer classification,” *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, no. 74, pp. 126–134, Jan. 2021, doi: 10.19101/IJATEE.2020.S1762126.

[12] A. Asuntha and A. Srinivasan, “Deep learning for lung Cancer detection and classification,” *Multimed Tools Appl*, vol. 79, no. 11–12, pp. 7731–7762, Mar. 2020, doi: 10.1007/s11042-019-08394-3.

[13] C. Dev, K. Kumar, A. Palathil, T. Anjali, and V. Panicker, “Machine Learning Based Approach for Detection of Lung Cancer in DICOM CT Image,” 2019, pp. 161–173. doi: 10.1007/978-981-13-5934-7_15.

[14] S. Baskar, P. M. Shakeel, K. P. Sridhar, and R. Kanimozhi, “Classification System for Lung Cancer Nodule Using Machine Learning Technique and CT Images,” in *2019 International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Jul. 2019, pp. 1957–1962. doi: 10.1109/ICCES45898.2019.9002529.

[15] S. Raut, S. Patil, and G. Shelke, “LUNG CANCER DETECTION USING MACHINE LEARNING APPROACH,” *INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS*, vol. 6, no. 1, Jan. 2021.

[16] S. Mukherjee and S. U. Bohra, “Lung Cancer Disease Diagnosis Using Machine Learning Approach,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, Dec. 2020, pp. 207–211. doi: 10.1109/ICISS49785.2020.9315909.

[17] N. Kalaivani, N. Manimaran, Dr. S. Sophia, and D. D Devi, “Deep Learning Based Lung Cancer Detection and Classification,” *IOP Conf Ser Mater Sci Eng*, vol. 994, no. 1, p. 012026, Dec. 2020, doi: 10.1088/1757-899X/994/1/012026.

[18] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, “Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases,” *Comput Biol Med*, vol. 132, p. 104348, May 2021, doi: 10.1016/j.compbiomed.2021.104348.

[19] E. Dritsas and M. Trigka, “Lung Cancer Risk Prediction with Machine Learning Models,” *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 139, Nov. 2022, doi: 10.3390/bdcc6040139.

[20] D. Mhaske, K. Rajeswari, and R. Tekade, “Deep Learning Algorithm for Classification and Prediction of Lung Cancer using CT Scan Images,” in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/ICCUBEA47591.2019.9128479.

[21] L. S.K., S. N. Mohanty, S. K., A. N., and G. Ramirez, “Optimal deep learning model for classification of lung cancer on CT images,” *Future Generation Computer Systems*, vol. 92, pp. 374–382, Mar. 2019, doi: 10.1016/j.future.2018.10.009.

[22] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, “A hybrid algorithm for lung cancer classification using SVM and Neural Networks,” *ICT Express*, vol. 7, no. 3, pp. 335–341, Sep. 2021, doi: 10.1016/j.icte.2020.06.007.

[23] M. K. Gould, B. Z. Huang, M. C. Tammemagi, Y. Kinar, and R. Shiff, “Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data,” *Am J Respir Crit Care Med*, vol. 204, no. 4, pp. 445–453, Aug. 2021, doi: 10.1164/rccm.202007-2791OC.

[24] M. A. Heuvelmans *et al.*, “Lung cancer prediction by Deep Learning to identify benign lung nodules,” *Lung Cancer*, vol. 154, pp. 1–4, Apr. 2021, doi: 10.1016/j.lungcan.2021.01.027.