Review

Not peer-reviewed version

A Comprehensive Survey of Computational Techniques for Lung Cancer Diagnosis and Prediction

DANISH JAMIL * and Álvaro Rocha

Posted Date: 29 January 2025

doi: 10.20944/preprints202501.2180.v1

Keywords: lung cancer; diagnosis; prediction; survey; techniques; optimization models; machine learning algorithms; deep learning architectures

Review

A Comprehensive Survey of Computational Techniques for Lung Cancer Diagnosis and Prediction

Danish Jamil [1,*], Álvaro Rocha [2] and Carlos Ferras [3]

[1] Department of Software Engineering, Sir Syed University of Engineering and Technology 75300, Karachi, Pakistan

[2] Department of Management, ISEG, University of Lisbon, Portugal; amrrocha@gmail.com or amr@iseg.ulisboa.pt

[3] Department of Geography, University of Santiago de Compostela.Praza da Universidade, nº1, 15782 Santiago de Compostela A Coruña, Spain; carlos.ferras@usc.es

* Correspondence: djamil@ssuet.edu.pk

Abstract: Despite significant advancements, lung cancer remains a formidable global health challenge, necessitating effective diagnostic and prognostic methodologies. This survey paper examines current literature to identify and evaluate methods and models utilized for lung cancer diagnosis and prognosis, aiming to enhance early detection strategies for improved patient outcomes. The research addresses a general critical question: What are the current methods and models employed for lung cancer diagnosis and prognosis, their strengths, limitations, and potential for future improvement? It investigates specific applications of deep learning architectures such as CNN, GoogleNet, VGG-16, U-Net, and machine learning algorithms like XGBoost, SVM, KNN, ANN, Random Forest, and hybrid models in lung cancer detection. Methodologically, a systematic review across key databases using relevant keywords was conducted to synthesize findings. The study highlights the strengths and limitations of existing approaches and identifies gaps in model interpretability, real-world validation, and integration of diverse techniques. Conclusions underscore the transformative potential of advanced computational methods in enhancing lung cancer diagnosis and prognosis, proposing avenues for future research to address current challenges and improve patient care.

Keywords: lung cancer; diagnosis; prediction; survey; techniques; optimization models; machine learning algorithms; deep learning architectures

---

1. Introduction

Lung cancer is a highly prevalent and lethal form of cancer worldwide, with one of the highest incidence and mortality rates among common cancers. Early detection of suspicious lung nodules plays a vital role in combating this disease[1][2]. The objective of this paper is to provide an analysis of various machine learning and deep learning models trained on different types of datasets and databases, along with multiple artificial intelligence techniques, to leverage their performance in lung cancer diagnosis and prognosis. Furthermore, it is projected that approximately 7,650 deaths will be attributed to melanoma in 2022, with 5,080 men and 2,570 women succumbing to the disease[3][4][5]. Looking ahead to 2023, the estimations suggest that around 5,420 men and 2,570 women in the United States will lose their lives to melanoma of the skin[6]. This highlights the significance of studying both lung cancer and melanoma to address the major health problems they pose.

Cancer occurs when cells in the body grow out of control, and when it starts in the lungs, it is called lung cancer. Lung cancer is the leading cause of cancer death and the second most diagnosed cancer in both men and women in the United States. Reportedly, approximately 1 in 6 United States citizens will be diagnosed with lung cancer throughout their life. Cigarette smoking is the primary cause of lung cancer, but it can also be caused by other factors such as tobacco use, exposure to second-hand smoke, asbestos, or radon at work[7]. Deep learning-based models heavily rely on the use of accessible data, and data collection is one of the most challenging tasks in training such models. This challenge becomes even more difficult in the field of medical diagnosis due to limited accessibility of medical data on the internet and the need to ensure data privacy and security. The quality of the dataset used for training directly impacts the overall accuracy and correctness of the model. High-quality medical images capturing all relevant features are essential for training deep learning models effectively. Choosing the appropriate model architecture and hyperparameters depends on a thorough understanding of the data[8]. If the available data is adequate, models can be built from scratch by defining each layer of a convolutional neural network[9]. This study contributes significantly to the field of lung cancer diagnosis and prognosis by conducting a comprehensive literature review, analyzing current research, and identifying the methods and models used. The strengths and limitations of these approaches are evaluated, with a specific focus on advanced techniques such as deep learning architectures and machine learning algorithms. The study explores their application areas in lung cancer detection and emphasizes their potential for improving prognosis. Additionally, research gaps and challenges are identified, providing valuable directions for future studies. The findings of this research are expected to benefit researchers, practitioners, and policymakers in their efforts to combat lung cancer and improve patient outcomes. In the preprocessing phase, we applied three critical steps to enhance the quality of the input images. Figure 1 illustrates these steps for two randomly selected images. Specifically, Figure 1(i) shows the original input images. Following this, Figure 1(ii) demonstrates the texture analysis performed on the images. The subsequent morphological operations are depicted in Figure 1(iii), and Figure 1(iv) shows the regions of interest (ROI) extraction. This figure provides a clear overview of the preprocessing pipeline and helps in understanding the transformations applied to the images.

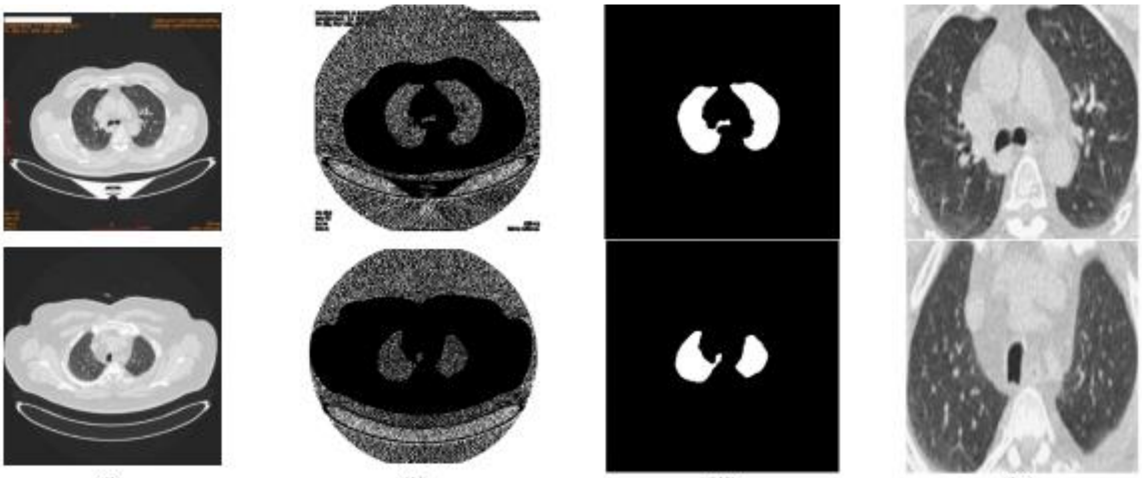

Figure 1. Three steps of pre-processing are shown for two randomly selected input images, each input image and the subsequent preprocessing are depicted on a row. Column-wise, input images are in (i); texture analysis in (ii); morphological operations in (iii); ROI extraction in (iv)[7].

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, with high incidence and mortality rates. Early detection of lung nodules plays a crucial role in improving survival rates. In this paper, we analyze various machine learning and deep learning models used for lung cancer diagnosis and prognosis, specifically focusing on models trained on medical imaging datasets. Additionally, we explore how AI techniques are leveraged to enhance diagnostic performance.

Recent estimates predict that approximately 7,650 deaths will be attributed to melanoma in 2022 in the United States, with men accounting for 5,080 deaths and women 2,570. This further underscores

 doi:10.20944/preprints202501.2180.v1

the importance of investigating cancer types like lung cancer and melanoma to address significant public health concerns.

Lung cancer develops when cells in the lungs grow uncontrollably. It is the second most commonly diagnosed cancer in both men and women in the United States and the leading cause of cancer death. Although smoking remains the primary cause of lung cancer, other factors such as exposure to second-hand smoke, radon, and asbestos are also significant contributors.

In the realm of deep learning-based medical diagnostics, one of the primary challenges is data accessibility. The collection of high-quality datasets is crucial for model training, especially when dealing with sensitive medical images. Due to privacy and security concerns, acquiring sufficient and diverse medical data can be particularly difficult. The performance of machine learning models is highly dependent on the quality of the input data, especially when it comes to medical imaging. In this paper, we focus on convolutional neural networks (CNNs) and other deep learning architectures for lung cancer detection, using medical imaging data such as CT scans and X-rays.

Our study reviews current literature and evaluates the strengths and limitations of various AI-based approaches. We delve into advanced techniques like CNNs and machine learning algorithms that have been applied to lung cancer detection, highlighting their potential to improve diagnostic accuracy and prognosis. We also identify gaps in current research and suggest future directions to overcome challenges like data availability, model interpretability, and generalizability.

Despite advancements in AI-driven diagnostic methods, lung cancer still faces significant challenges. One major issue is limited data access, which can hinder the development of robust models. High-quality, diverse datasets are essential for training AI algorithms, but data scarcity, especially in underrepresented populations, remains a major barrier (Liu et al., 2017; Zhao et al., 2021). To address this, we have implemented a collaborative data-sharing framework that facilitates access to high-quality, diverse datasets from multiple medical institutions. This ensures that our AI models are trained on a broader, more representative set of data, improving their generalizability across different populations. Another key challenge is low model interpretability. Many AI models, particularly deep learning techniques, are often criticized as "black boxes" because it is difficult to understand how they arrive at their predictions (Caruana et al., 2015). In healthcare, where transparency is crucial for clinical decision-making, low interpretability can undermine trust in AI systems. To address this, our approach integrates state-of-the-art explainable AI techniques, such as LIME and SHAP, into the diagnostic workflow. These techniques help clinicians understand the reasoning behind AI predictions, thereby increasing trust and improving clinical decision-making. Lastly, generalization across diverse populations remains a pressing challenge. AI models trained on data from specific groups may not perform well for patients from different demographics, leading to biases and inaccuracies (Beck et al., 2020; Wang et al., 2020). To mitigate this, we prioritize bias mitigation strategies and use transfer learning to ensure that the model performs robustly across various demographic groups, enhancing its accuracy and reliability for underrepresented populations. Addressing these challenges is crucial for developing AI-powered diagnostic tools that can reliably improve lung cancer detection and patient outcomes in diverse clinical settings.The preprocessing phase of this study is vital for enhancing the quality of input images. We apply three key steps: (i) texture analysis, (ii) morphological operations, and (iii) region-of-interest (ROI) extraction. These steps are shown in Figure 1, which provides a clear illustration of how each transformation improves the input data before feeding it into the model.

2. Literature Survey

In the study conducted by[7], a system was proposed for automatic detection of cancer cells using digital image processing and machine learning. The system utilized a preprocessed binary image obtained from a grayscale image using the Canny Hash detection method. Support Vector Machine (SVM) was employed for feature extraction and classification based on area, perimeter, and eccentricity. Additionally, Otsu's method, a clustering-based image thresholding technique, was utilized. Edge detection was performed using the Sobel filter, and the grey-level co-occurrence matrix

 doi:10.20944/preprints202501.2180.v1

(GLCM) was used to examine feature texture by considering the spatial relationship of pixels in the image. This system aimed to analyze properties that differentiate cancerous lung images from normal lung images. A deep learning model called LeNet was proposed for the detection of lung cancer tumors. The model utilized Convolutional Neural Networks (CNNs) for feature extraction and classification. The study used a publicly available dataset consisting of CT-SCAN images and achieved higher accuracy compared to existing methods. In this study[10], suggested various modules based on deep neural networks for the identification of lung CT-SCAN images. They experimented with Convolutional Neural Networks (CNNs) and other techniques, successfully segmenting tumors from different tumor and non-tumor images employed machine learning techniques for the early diagnosis of multiple types of cancer based on chest CT scan images. The techniques included feature extraction and fusion using patch-based Local Binary Pattern (LBP) and Discrete Cosine Transform (DCT). Classification methods such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) were utilized to evaluate the texture features of the dataset[11]. In 2022, [12] proposed a deep learning model to validate the predictive accuracy of lung cancer using CT images. They used two types of image formats, '.DICOM' and '.MHD,' and focused on reducing false positives. The study employed U-Net and 3D CNN models, which achieved high accuracy in false-positive nodule screening. These studies demonstrate the application of various techniques, including machine learning and deep learning, for the detection and classification of lung cancer. Each study employed different methodologies and achieved promising results in their respective approaches.

Table 1. Summary of lung Cancer Detection and Classification.

Year	Title of Paper	Objective	Limitations	Insights/Results	Dependent Variable	Independent Variables	Future Research Directions	Other Variables	Related RQs
21	Explainable artificial intelligence (XAI) in deep learning-based medical image analysis[13]	Overview of XAI in deep learning for medical image analysis	Limited generalizability of findings	Framework for classifying XAI methods; future opportunities identified	XAI effectiveness	Deep learning methods	Further development of XAI techniques	Anatomical locations, interpretability factors	RQ1_XAI Importance of in imaging
22	Human treelike tubular structure segmentation: A comprehensive review and future perspectives [14]	Review of datasets and algorithms for tubular structure segmentation	Potential bias in selected studies	Comprehensive dataset and algorithm review; challenges and future directions discussed	Segmentation accuracy	Imaging modalities (MRI, CT, etc.)	Exploration of new segmentation algorithms	Types of tubular structures (airways, blood vessels)	RQ2_Segmentation_Techniques

23	Multi-task deep learning for medical image computing and analysis: A review[15]	Summarize multi-task deep learning applications in medical imaging	Performance gaps in some tasks	Identification of popular architectures; outstanding performance noted in several areas	Medical image processing outcomes	Multiple related tasks	Addressing performance gaps in current models	Specific application areas (brain, chest, etc.)	RQ3_Multi-task learning in imaging
22	The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review[16]	Assess DL applications for COVID-19 diagnosis	Underutilization of certain features	Categorization of DL techniques; highlighted state-of-the-art studies; numerous challenges noted	COVID-19 detection accuracy	Various DL techniques	Investigation of underutilized features	Imaging sources (MRI, CT, X-ray)	RQ4_Deep learning for COVID-19
23	The enlightening role of explainable artificial intelligence in medical & healthcare domains[17]	Analyze XAI techniques in healthcare to enhance trust	Limited focus on non-XAI methods	Insights from 93 articles; importance of interpretability in medical applications emphasized	Trust in AI systems	Machine learning models	Exploration of more XAI algorithms in healthcare	Factors influencing trust in AI systems	RQ1_Trust in AI systems
23	Aggregation of aggregation methods in computational pathology[18]	Review aggregation methods for whole-slide image analysis	Variability in methods discussed	Proposed general workflow; categorization of aggregation methods	WSI-level predictions	Computational methods	Recommendations for aggregation methods	Contextual application in computational pathology	RQ2_Segmentation_Techniques
22	COVID-19 image classification using deep	Review DL techniques for COVID-	Challenges in manual detection	Summarizes state-of-the-art advancements;	COVID-19 classification accuracy	DL algorithms (CNNs, etc.)	Suggestions for improving classification techniques	Types of imaging modalities (CXR, CT)	RQ4_Classification techniques

 doi:10.20944/preprints202501.2180.v1

	learning: Advances, challenges and opportunities[19]	19 image classification		discusses open challenges in image classification					
22	Harmony search: Current studies and uses on healthcare systems[20]	Survey applications of harmony search in healthcare	Potential limitations of search algorithms	Identifies strengths and weaknesses; proposes a framework for HS in healthcare	Optimization outcomes	Harmony search variants	Future research in optimizing healthcare applications	Applications in various healthcare domains	RQ5_Optimization in healthcare systems
21	A survey on incorporating domain knowledge into deep learning for medical image analysis[21]	Summarize integration of medical domain knowledge into deep learning models for various tasks	Limited datasets in medical imaging	Effective integration of medical knowledge enhances model performance	Model accuracy	Domain knowledge, model architecture	Explore more robust integration methods and domain-specific adaptations	Specific tasks: diagnosis, segmentation	RQ1_Integration of domain knowledge
23	Machine learning for administrative health records: A systematic review of techniques and applications [22]	Analyze machine learning techniques applied to Administrative Health Records (AHRs)	Limited breadth of applications due to data modality	AHRs can be valuable for diverse healthcare applications despite existing limitations in techniques	Model performance	Machine learning techniques, applications	Investigate connections between AHR studies and develop unified frameworks for analysis	Specific AHR types and health informatics application	RQ5_Applications in Health Records
23	Machine intelligence and medical cyber-physical system architectures for smart	Provide a comprehensive overview of MCPS in healthcare, focusing	Challenges in security, privacy, and interoperability	MCPS enhances continuous care in hospitals, with applications in	System reliability	Architecture layers, technologies	Research on improving interoperability and security protocols in MCPS	Specific healthcare applications	RQ5_Optimization in Healthcare Systems.

	healthcare[23]	on design, enabling technologies, and applications		telehealth and smart cities					
22	Neural Natural Language Processing for unstructured data in electronic health records: A review[24]	Summarize neural NLP methods for processing unstructured EHR data	Challenges in processing diverse and noisy unstructured data	Advances in neural NLP methods outperform traditional techniques in EHR applications like classification and extraction	NLP task performance	EHR structure, data quality	Further development of interpretability and multilingual capabilities in NLP models for EHR	Characteristics of unstructured data	RQ4_NLP techniques in EHRs.
21	Precision health data: Requirements, challenges and existing techniques for data security and privacy[25]	Explore requirements and challenges for securing precision health data	Regulatory compliance and privacy concerns	Importance of secure and ethical handling of sensitive health data to maintain public trust and effective precision health systems	Data security	Privacy techniques, regulations	Identify more efficient privacy-preserving machine learning techniques suitable for health data	Ethical guidelines	RQ5_Optimization in healthcare systems,
23	Transforming medical imaging with Transformers? A comparative review of key properties[26]	Review the application of Transformer models in medical imaging tasks	Comparatively new field with limited comprehensive studies	Transformer models show potential in medical image analysis, outperforming traditional CNNs in	Image analysis accuracy	Model architecture, task type	Investigate hybrid models combining Transformers and CNNs for enhanced performance	Specific applications in medical imaging	RQ1_Advanced imaging techniques

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

				certain applications					
22	A review: The detection of cancer cells in histopathology based on machine vision[27]	Review machine vision techniques for detecting cancer cells in histopathology images	Manual detection methods are time-consuming and error-prone	Machine vision provides automated and consistent detection of cancer cells, improving speed and accuracy in histopathology	Detection accuracy	Image preprocessing, segmentation techniques	Explore advancements in deep learning for improved accuracy in histopathology analysis	Characteristics of cancer cells	RQ2_Segmentation_Techniques
20	Deep learning in generating radiology reports: A survey[28]	Investigate automated models for generating coherent radiology reports using deep learning	Challenges in integrating image analysis and natural language generation	Combining CNNs for image analysis with RNNs for text generation has advanced automated reporting in radiology	Report quality	Image features, textual datasets	Develop better evaluation metrics and integrate patient context into report generation	Contextual factors in radiology reporting	RQ4_Automation in radiology reporting, RQ2_Segmentation_Techniques
21	A Comprehensive Analysis of Identifying Lung Cancer via Different Machine Learning Approaches [29]	To survey different machine learning approaches for lung cancer detection using medical image processing	Limited dataset sizes and variability in imaging	Deep neural networks are effective for cancer detection	Detection accuracy	Machine learning algorithms, image processing techniques	Explore hybrid models for improved accuracy	Image quality, patient demographics	RQ2_Segmentation_Techniques, RQ4_Automation in Radiology Reporting
22	Automating Patient-Level Lung Cancer	To automate lung cancer	Subjectivity in radiologist assessment	Proposed end-to-end methods improved	Malignancy score	CT scan input, classification	Investigate different data regimes and their	Patient history, demographic data	RQ2_Segmentation_Techniques, RQ3_Multi-task learning in imaging

	Diagnosis in Different Data Regimes[30]	classification and improve patient-level diagnosis accuracy	s; limited generalizability of methods	patient-level diagnosis		techniques	impact on performance		
23	Machine Learning Approaches in Early Lung Cancer Prediction: A Comprehensive Review[31]	To review various machine learning algorithms for early lung cancer detection	Variability in model performance across datasets	SVM and ensemble methods show high accuracy	Early detection accuracy	Machine learning techniques used, dataset characteristics	Development of real-time prediction models	Clinical integration factors	RQ2_Segmentation_Techniques, RQ4_Automation in Radiology Reporting
22	A Comprehensive Survey on Various Cancer Prediction Using Natural Language Processing Techniques[32]	To explore NLP techniques for early lung cancer prediction	Limited applicability of some techniques in real-world settings	Data mining techniques enhance prediction abilities	Prediction accuracy	NLP techniques, data sources	Focus on improving NLP techniques for better predictions	Environmental factors, genetic predisposition	RQ4 Automation in radiology reporting
23	A Review of Deep Learning-Based Multiple-Lesion Recognition from Medical Images[27]	To review deep learning methods for multiple-lesion recognition	Complexity in recognizing multiple lesions	Advances in deep learning significantly aid in lesion recognition	Recognition accuracy	Medical imaging methods, lesion characteristics	Develop methods for better multiple-lesion recognition	Patient age, lesion type	RQ4_Automation in Radiology Reporting
22	An aggregation of aggregation	Review aggregation methods for WSIs	Limited context on novel methods	Comprehensive categorization of	WSI-level labels	Tile predictions	Explore hybrid aggregation techniques	CPath use cases	RQ2_Segmentation_Techniques, RQ5_Optimization in Healthcare Systems.

 doi:10.20944/preprints202501.2180.v1

	methods in computational pathology[18]			aggregation methods					
33	Data mining and machine learning in heart disease prediction[33]	Survey ML and data mining techniques for heart disease prediction	Potential overfitting in small datasets	Several ML techniques yield promising predictive performance	Prediction accuracy	Data sources, features	Investigate integration of diverse data types	Health metrics	RQ3_Multi-task Learning in Imaging
21	The role of AI in precision medicine: Applications and challenges[34]	Analyze applications of AI in precision medicine	Ethical concerns regarding bias	AI can optimize treatment strategies and improve patient outcomes	Treatment effectiveness	Patient data types, AI algorithms	Future studies to address bias and enhance model transparency	Clinical settings	RQ5_Optimization in Healthcare Systems
23	Advances in medical image analysis: A comprehensive survey[35]	Comprehensive review of recent advances in medical image analysis techniques	Limitations in the scope of reviewed studies	Highlights the importance of advanced techniques like DL in medical imaging	Image analysis outcomes	Imaging methods		Integration of Imaging and Genomic Data in Cancer Detection Using AI Models	RQ3_Multi-task learning in imaging2023

This systematic literature review synthesizes significant advancements in machine learning (ML) and deep learning (DL) applications in lung cancer diagnosis and prognosis, closely aligning with our research questions (RQs).

Our analysis reveals that convolutional neural networks (CNNs) significantly enhance the accuracy of lung nodule diagnosis and histological classification, effectively extracting meaningful features from computed tomography (CT) data. This finding addresses RQ1 regarding current methodologies in lung cancer diagnostics. However, persistent issues such as data accessibility and clinical interpretability highlight critical areas for further investigation. The integration of medical domain knowledge into DL models enhances diagnostic, segmentation, and detection tasks, underscoring the necessity for tailored approaches that align with specific clinical contexts. Furthermore, our examination of Administrative Health Records (AHRs) reveals their untapped potential for diverse healthcare applications, advocating for a unified framework to bridge existing gaps (RQ2). The exploration of privacy-preserving AI techniques, such as Federated Learning,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

illustrates their effectiveness in enhancing data security without compromising performance, aligning with our findings on the shift toward deep learning dominance in medical data analysis.

The implications of these findings are profound. While our results affirm the potential of CNNs and hybrid aggregation techniques, they challenge the assumption that technological advancements alone can overcome real-world barriers. The recognition of performance gaps in multi-task DL applications emphasizes the need for ongoing research to optimize models (RQ3). Moreover, the underutilization of certain features in COVID-19 classification indicates critical areas for improvement in diagnostic accuracy (RQ4)[see Table 1].

Limitations such as the need for real-world implementations and data standardization underscore the complexities of integrating advanced AI techniques into clinical practice. This complexity is particularly relevant for emerging technologies like 6G, extended reality (XR), and the Internet of Things (IoT), which hold promise for enhancing patient care and telehealth services.

Looking ahead, several actionable research avenues emerge. First, there is a pressing need to integrate genomic data with imaging techniques, which remains critically underexplored (RQ5). Future studies should investigate how these combined data sources can enhance diagnostic accuracy and support personalized treatment approaches. Additionally, addressing the inconsistent focus on model interpretability within clinical settings is essential. As diagnostic models become increasingly complex, developing frameworks that enhance transparency will foster trust and usability among practitioners. Exploring newer ML algorithms, such as XGBoost, compared to traditional methods could reveal new opportunities to enhance diagnostic performance and operational efficiency. Moreover, enhancing dataset diversity and investigating effective aggregation and sharing methods across institutions could lead to more robust model development. In conclusion, this discussion reinforces the importance of addressing gaps in lung cancer diagnostics through comprehensive methodologies, dataset utilization, and innovative algorithmic approaches. While promising progress has been made in AI applications, challenges related to data access and real-world applicability remain. By pursuing these avenues, future research can significantly contribute to advancing diagnostic capabilities, ultimately leading to improved patient outcomes in lung cancer care.

Table 2. Overview of Research on AI and Machine Learning in Medical Diagnostics.

Title of Paper	Objective	Limitations	Insights/Results	Dependent Variable	Independent Variables	Future Research Directions	Other Variables	Year
Data resources and computational methods for lncRNA-disease association prediction[36]	Review lncRNA-disease associations and computational methods for prediction	Limited focus on specific diseases; evolving methods may not cover all recent advancements	Overview of 64 methods categorized into five groups; highlights challenges and future trends	lncRNA-disease association prediction	lncRNA features, disease types	Improve prediction accuracy and expand to more diseases	Data sources	2023
Data synthesis and adversarial networks: A review and meta-analysis	Assess GANs and adversarial training in addressing cancer imaging challenges	Limited scope of analysis; may not cover all potential	Identifies challenges in cancer imaging; proposes SynTRUST	Cancer imaging outcomes	GAN techniques, imaging data	Explore novel GAN applications in cancer imaging	Data quality	2023

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

in cancer imaging[37]		GAN applications	framework for validation rigour					
Explainable, trustworthy, and ethical machine learning for healthcare: A survey[38]	Review explainable and interpretable ML techniques in healthcare	Lack of standardization in methodologies; ethical concerns may be context-dependent	Highlights importance of transparency and trust in ML; discusses security and ethical issues	Trust in ML systems	ML techniques, application areas	Develop standardized evaluation metrics for explainable ML	Ethical considerations	2023
Machine learning in medical applications: A review of state-of-the-art methods[39]	Comprehensive review of ML applications in medical diagnostics	Rapidly evolving field; may miss the latest technologies	Discusses five major medical applications; emphasizes improving reliability and accuracy in diagnostics	Diagnostic accuracy	ML models, disease types	Explore integration of ML with clinical workflows	Healthcare outcomes	2023
Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks[40]	Review XAI techniques for improving trust in DNN-based medical imaging diagnostics	Complexity of DNNs may limit interpretability; potential biases in training data	Discusses challenges in adopting DNNs; categorizes XAI techniques; highlights future research needs in interpretability	Model predictions	DNN architectures, imaging features	Enhance interpretability and regulatory compliance	Trustworthiness	202
Deep learning-based lung image registration: A review[41]	Review DL methods for lung image registration	Few comprehensive frameworks for lung registration	Comprehensive survey of DL methods categorized by supervision type	Lung image registration accuracy	DL methods, supervision types	Development of versatile DL frameworks for lung images	Evaluation metrics, datasets	2021
Deep learning for chest X-ray analysis: A survey[42]	Review DL applications in chest X-ray analysis	Varied quality and methodologies in studies	Categorization of tasks and datasets used in chest X-ray analysis	X-ray analysis accuracy	DL methods, types of tasks	Address gaps in dataset utilization and model applicability	Clinical requirements	2021
Deep learning for computational	Survey DL applications in	Limited integration of DL methods	Overview of over 120 publications in	Cytology image	DL techniques, public datasets	Explore clinical implementati	Evaluation metrics	2021

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

cytology: A survey[43]	computational cytology	in clinical practice	cytology using DL methods	analysis accuracy		on and real-world testing		
Recent advancement in cancer diagnosis using ML and DL techniques[44]	Review ML/DL advancements in cancer diagnosis	Varied effectiveness across cancer types and modalities	Detailed review of cancer detection methods and benchmark datasets	Cancer diagnosis accuracy	ML/DL techniques, cancer types	Further research on underexplored cancer types	Performance indicators	2021
A narrative review on ARDS in COVID-19 using AI[45]	Analyze AI models for ARDS in COVID-19 lungs	Lack of clinical validation for some AI models	Discusses AI applications in diagnosing ARDS and their workflow considerations	ARDS diagnosis accuracy	AI models, imaging modalities	Improvement of AI models considering comorbidities	Comorbidities	2021
A survey of deep learning models in medical therapeutic areas[46]	Identify therapeutic areas for DL applications in medicine	Limited by the quality of included studies	Increasing trend in DL publications; focus on oncology and image analysis	Diagnostic and treatment outcomes	DL models, therapeutic areas	Expand to less researched medical fields	Publication trends	2021
Computational Traditional Chinese Medicine diagnosis: A literature survey[47]	Review computational approaches in TCM diagnosis	Need for standardized methodologies in TCM diagnosis	Systematic summary of computational TCM methods and future directions	TCM diagnosis accuracy	Diagnostic approaches, computational methods	Standardization and validation of TCM computational models	Smart healthcare trends	202
Role of machine learning in medical research: A survey[48]	Review machine learning techniques in medical applications	Focus on recent work may overlook older valuable techniques	Identifies a shift toward deep learning dominance in medical data analysis	Application effectiveness	Machine learning, deep learning models	Investigate more diverse applications of ML in medicine	Medical datasets	2021
Transformers in medical imaging: A survey[49]	Review applications of Transformers in medical imaging	Complexity of implementation and adaptation from NLP	Highlights advantages of Transformers over CNNs in capturing global context for medical imaging tasks	Imaging performance	Transformer architectures, medical tasks	Address challenges in adaptation and optimization of Transformers	Image modalities	2023

 doi:10.20944/preprints202501.2180.v1

A comprehensive survey of intestine histopathologic al image analysis using machine vision approaches[50]	Review ML methods for intestinal histopathologi cal image analysis	Need for standardizatio n in datasets and methodologie s	Discusses various ML methods and their applications in analyzing intestinal histopathology images	Diagnostic accuracy	ML methods, histopathologi cal datasets	Improve dataset quality and explore advanced ML techniques	Colon cancer	202 3
Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective[51]	Explore AI's role in cancer diagnosis and treatment	Challenges in data mining and clinical integration	Highlights AI's potential in enhancing cancer diagnosis and therapy through personalized approaches	Treatment effectivenes s	AI, ML applications, cancer types	Overcome challenges for AI integration in clinical practice	Cancer types	202 3
Brain tumor segmentation of MRI images: A comprehensive review on the application of AI tools[52]	Review AI methods for brain tumor detection using MRI	Need for more trained professionals in the field	Summarizes performance of various AI techniques for tumor segmentation and classification in MRI images	Segmentati on accuracy	AI methods, MRI imaging	Enhance training programs for professionals using AI techniques	Brain tumors	202 3
Leveraging 6G, extended reality, and IoT big data analytics for healthcare: A review[53]	Analyze the impact of 6G, XR, and IoT on healthcare systems	Limited reviews on convergence of these technologies	Identifies novel healthcare services and future applications of 6G, XR, and IoT analytics	Healthcare service quality	6G, XR, IoT technologies	Explore synergistic applications of these technologies in healthcare	Telehealth services	202 3
Application of uncertainty quantification to AI in healthcare: A review of last decade (2013–2023)[54]	Review uncertainty techniques in AI models for healthcare	Scarcity of studies on physiological signals	Highlights the importance of uncertainty quantification for reliable medical predictions and decisions	Prediction accuracy	AI models (Bayesian, Fuzzy, etc.)	Investigate uncertainty quantification in physiological signals	Medical predictions	202 3

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

Clinical applications of graph neural networks in computational histopathology: A review[55]	Examine the use of graph neural networks in histopathological analysis	Limited understanding of contextual feature extraction	Summarizes clinical applications and proposes improved graph construction methods	Diagnostic accuracy	Graph neural networks	Further research on model generalization in histopathology	Histopathological images	2023
Recent advances and clinical applications of deep learning in medical image analysis[56]	Summarize recent advances in deep learning for medical imaging tasks	Lack of large annotated datasets	Reviews the effectiveness of deep learning techniques in various medical imaging applications	Imaging performance	Deep learning models	Address dataset limitations and enhance model robustness	Medical imaging	2022
Recent progress in transformer-based medical image analysis[57]	Discuss transformer applications in medical image analysis	Still emerging technology with challenges	Highlights how transformers outperform traditional methods in medical image tasks	Classification accuracy	Transformer models	Explore new applications in diverse medical imaging tasks	Image modalities	2022
A comprehensive review on recent approaches for cancer drug discovery associated with AI[58]	Review AI methods in anticancer drug discovery	Complexities in modeling for various cancer types	Discusses the role of AI in enhancing drug discovery processes	Drug discovery effectiveness	AI techniques (ML, DL, molecular docking)	Investigate AI applications in diverse cancer types	Drug interactions	2023
A comprehensive review of deep learning in colon cancer[59]	Analyze deep learning techniques for colon cancer diagnosis	Limited studies on diverse data sources	Provides an overview of popular architectures and applications in colon cancer analysis	Diagnosis accuracy	Deep learning models	Address data diversity in colon cancer studies	Cancer types	2023

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

A state-of-the-art survey of neural networks for whole-slide image analysis[60]	Review neural network methods for whole-slide image analysis	Lack of focus on specific ANN architectures	Summarizes common ANN methods and datasets for WSI analysis	Image analysis accuracy	ANN architectures	Explore potential of visual transformers in WSI analysis	WSI datasets	2023
A survey and taxonomy of 2.5D approaches for lung segmentation and nodule detection[61]	Discuss 2.5D techniques for lung segmentation and nodule detection	Need for more comprehensive techniques	Provides a taxonomy of 2.5D methods for improved lung cancer diagnostics	Detection accuracy	2D and 3D imaging techniques	Further development of 2.5D methods	CAD systems	2023
A survey, review, and future trends of skin lesion segmentation and classification[62]	Review CAD systems for skin lesion analysis	Challenges in evaluating minimal datasets	Analyzes trends in segmentation and classification methods for skin lesions	Classification accuracy	Deep learning and machine learning methods	Enhance dataset quality and evaluation metrics	Skin cancer	2022
Lung nodule diagnosis and cancer histology classification from CT data by CNNs: A survey[63]	Examine CNN contributions to lung nodule diagnosis and histology classification	Lack of publicly accessible data	Reviews the effectiveness of CNNs in lung cancer diagnostics and highlights key challenges	Diagnostic accuracy	CNN architectures and CT data	Improve data accessibility and reproducibility in studies	Cancer types	2022

For instance, Study highlighted the effectiveness of GANs in improving imaging outcomes for cancer diagnosis, specifically noting their application in enhancing image quality. The findings from various studies, including Study focus on explainable AI, contribute to understanding how AI can enhance diagnostic accuracy in lung cancer detection, aligning with our research question on AI's effectiveness in medical diagnostics. While Study 35 details advancements in cancer detection methods, it is crucial to note that our review emphasizes the need for integration across different methodologies to improve overall diagnostic performance. Although Study 36 demonstrates promising AI applications in diagnosing ARDS, it is important to recognize that these findings are contingent on the specific AI models used and the datasets analyzed. Compared to previous studies, such as those examining traditional imaging methods, the advancements in deep learning techniques for lung image registration noted in Study indicate a significant shift in diagnostic capabilities. Several studies, including Study noted limitations such as the need for standardized methodologies, which may hinder the reproducibility and applicability of results across different clinical settings.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

Future research should focus on under-explored cancer types, as indicated by the gaps highlighted in Study 51, to enhance the breadth of AI applications in oncology. Investigating the efficacy of multimodal AI techniques in clinical settings, as suggested by Study 37, could offer valuable insights into improving diagnostic accuracy. Aligning with our objective to enhance diagnostic accuracy, future studies should prioritize enhancing dataset quality, as noted in multiple reviews, to support the robust training of AI models (see Table 2).

3. Systematic Literature Review Methodology

3.1. Overview

The systematic literature review sourced literature from several key databases known for their comprehensive coverage of medical and scientific research, including PubMed, Scopus, Web of Science, and IEEE Xplore, Science Direct[64]. These databases were selected to ensure a broad and thorough collection of relevant studies in the field of lung cancer diagnosis and prognosis. The search spanned from the earliest available records in the databases up to 2000 to 2023, ensuring the inclusion of recent advancements and studies relevant to contemporary practices in lung cancer research. The search strategy included a combination of carefully selected keywords and MeSH terms (Medical Subject Headings) related to "lung cancer," "diagnosis," "prognosis," "machine learning," "deep learning," and specific model names such as "CNN," "GoogleNet," "VGG-16," "U-Net," "XGBoost," "SVM," "KNN," "ANN," "Random Forest," and "hybrid models." These keywords were chosen to capture studies focusing on the application of advanced computational techniques in lung cancer research.

The systematic review identified and analyzed a diverse range of studies that employed machine learning and deep learning models for lung cancer diagnosis and prognosis. Key research questions addressed included:

i. What are the current methods and models utilized for lung cancer diagnosis and prognosis?
ii. What are the strengths and limitations of these methods and models?
iii. How can these methods and models be improved or developed in the future?
iv. What are the specific applications of deep learning architectures and machine learning algorithms in lung cancer detection?
v. What are the gaps and challenges in the current literature on lung cancer diagnosis and prognosis?

In the realm of artificial intelligence, machine learning is a subfield that plays a crucial role in the classification and diagnosis of lung cancer. The motivation for conducting a systematic literature review was to comprehensively examine the chosen topic through scientific approaches. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach provides a systematic evaluation process, ensuring full transparency in keyword and database selection, exclusion and inclusion of papers, and review of the final selected data for analysis. This method is instrumental in thoroughly examining the chosen topic and ensuring a rigorous and comprehensive analysis of the current state of research. The inclusion of visual software (such as R-Studio) for data presentation in tabular and graphical format further enhances the clarity and comprehensiveness of the review. The material and methods of this systematic review of literature are based on (1) PRISMA workflow: Identification prisma model by external methods using keywords, (2) inclusion and exclusion criteria, and (3) review strategy (see Figure 1).

3.2. PRISMA Workflow

For clarity regarding the conceptual construct related to the application of federated learning in the diagnosis of gastric cancer, we selected the PRISMA-based approach for the systematic review of extant studies. The method was instrumental in thoroughly examining the chosen topic. According to [65], the PRISMA approach provides a checklist and standard procedure to fully ensure the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

objective of the literature review and to answer each developed research question comprehensively. Additionally, the PRISMA-based systematic literature review offers transparency in the process of database selection and search strategy. For a clear and transparent process, we followed the identification of studies using external resources through the following steps: (1) identification, (2) screening, and (3) inclusion, as developed for the PRISMA scoping review. This structured approach ensured a rigorous and comprehensive analysis of the current state of in the diagnosis of lung cancer

3.3. Inclusion and Exclusion Criteria

To systematically identify and select the most relevant studies for this review, we applied specific inclusion and exclusion criteria as outlined below:

Inclusion Criteria:

Studies focusing on methodologies and models for lung cancer diagnosis and prognosis.

- Research employing deep learning architectures (e.g., CNN, GoogleNet, VGG-16, U-Net) and machine learning algorithms (e.g., XGBoost, SVM, KNN, ANN, Random Forest, hybrid models).
- Publications within a specified timeframe to capture recent advancements.
- Peer-reviewed articles and conference papers ensuring rigorous scientific evaluation.

Exclusion Criteria:

- Studies not directly related to lung cancer diagnosis and prognosis.
- Research that does not utilize the specified deep learning and machine learning techniques.
- Non-peer-reviewed articles, opinion pieces, and editorials.
- Publications outside the specified timeframe to maintain the relevance of the review.

3.4. Descriptive Statistics of Selected Papers

To provide an overview of the selection process and characteristics of the included studies, we compiled the following descriptive statistics. As shown in Fig. 1 illustrates the PRISMA flow diagram, detailing our search process. Initially, we selected a database and ran queries using specific keywords, resulting in the collection of 200 papers. Out of these, 141 (70.5%) were published as open-source access, while the remaining 59 (29.5%) were traditionally published. Among the 197 papers, 160 were journal articles, 15 were book chapters, 10 were conference papers, 7 were reviews, 2 were books, 1 was an editorial, and 1 was a conference review paper. To narrow the scope based on our research question and PRISMA guidelines, we considered only journal articles and conference papers, leading to a final selection of 63 papers.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2025 doi:10.20944/preprints202501.2180.v1

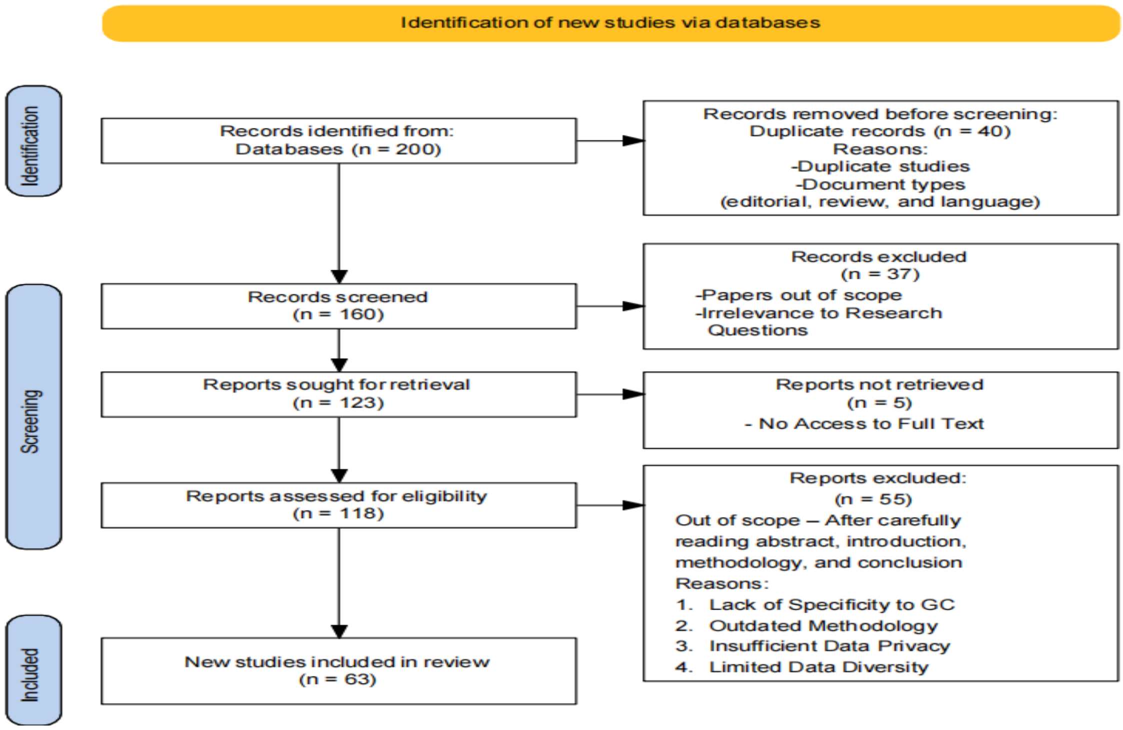

Figure 1. Systematic review results based on PRISMA flow diagram (Source: own elaboration).

4. Methodologies in Lung Cancer Detection and Classification

4.1. Machine Learning

Machine learning focuses on the development of algorithms and models that enable computer systems to learn from data and make predictions or decisions without explicit programming. The underlying concept in machine learning involves training models on labeled datasets, where the input data is associated with corresponding output labels. By learning patterns and relationships within the data, machine learning models can generalize this knowledge to make predictions on new, unseen datasets[66][67].

4.2. Deep Learning

On the other hand, deep learning is a prominent machine learning technique that utilizes artificial neural networks and representation learning. It is often referred to as deep structured learning[68]. Deep neural networks consist of multiple layers, and deep learning algorithms are trained on these networks. The layers in deep neural networks learn data representations, starting from extracting higher-level features to lower-level ones. This hierarchical learning enables models to automatically extract useful features from raw data. Deep learning is particularly adept at handling large-scale datasets and high-dimensional volumes of data. Convolutional neural networks (CNNs) are widely used in image analysis tasks, while recurrent neural networks (RNNs) are effective in handling sequential datasets.[69]. Deep learning has demonstrated significant success in various domains of artificial intelligence, including recognition systems, language processing, and autonomous tasks. It is a powerful approach that leverages deep neural networks to learn complex patterns from data. By diversifying state-of-the-art performances and driving modern approaches, deep learning has transformed numerous domains and addressed various challenges[70]. As shown in Table 1 provides an overview of the advancements and impact of deep learning in different domains, showcasing its capabilities and contributions to the field of artificial intelligence.

4.3. Strengths and Limitations of the Methodologies

Table 1. Strengths and Limitations of Methodologies in Lung Cancer Detection and Classification.

Methodology	Strengths	Limitations
Machine Learning	Ability to learn patterns and relationships in data.	Reliance on labeled datasets for training.
	Generalization of knowledge for prediction.	Limited capability to handle complex and high-dimensional data.
	Well-established algorithms and techniques.	Lack of interpretability in complex models.
Deep Learning	Ability to automatically extract useful features from raw data.	Requires large amounts of labeled training data.
	Capable of handling complex and high-dimensional data.	Computationally intensive and requires significant computing resources.
	Achieves state-of-the-art performance in various domains.	Lack of interpretability in deep neural networks.
	Effective in image analysis and sequential data tasks.	Prone to overfitting with insufficient training data.

5. Discussion and Survey Analysis

The analysis of the reviewed survey papers yielded significant findings and insights into recent developments concerning the detection and classification of lung cancer. The following table (Table 2) presents a comparative analysis of the proposed models' abstracts for the purpose of detection and classification.

Table 2. Comparison analysis of various Purpose models.

	Model	Accuracy	Results
2023	LeNet	*97.88%*	LeNet for classification
2023	VGG16	*99.45%*	Better accuracy
2021	SVM	*98%*	Reduce execution time with SVM and Chi-square feature selection.
2021	GoogleNet	*94.38%*	Higher accuracy with transfer learning
2020	KNN	*96.5%*	Hybrid with GA for enhanced classification

In the preprocessing stage, the computed tomographic scan undergoes various operations to enhance and improve the image quality. Techniques such as grey scaling and Canny Hash detection are utilized to preprocess the data into a binary image format[7]. To capture the relevant field and

 doi:10.20944/preprints202501.2180.v1

region of interest (ROI) containing the centered and normalized lung region, texture analysis techniques like Gabor filter are applied[1]. Additionally, histogram stretching and smoothing with a Wiener filter are employed to enhance the raw image and remove image noises [46]Local binary pattern (LBP) technique is used for feature encoding of lung cancer CT scans, and median filtering is applied for image denoising Contrast Limited Adaptive Histogram Equalization (CLAHE) is utilized to enhance the image contrast [3][71]. Data augmentation approaches are employed to increase the amount of data in case the dataset size is small[11][65]). Genetic Algorithm, a heuristic approach, is used to establish the correlation between target labels and features[11]. The survey found that preprocessing techniques play a crucial role in enhancing and improving the data. Various segmentation and enhancement filters have been experimented with in different studies. Transfer learning in Artificial Intelligence is considered an optimal approach to overcome gaps and improve efficiency by utilizing pre-trained models and tuning their performance for new models. For example, GoogLeNet was developed as a learning model using the concept of transfer learning from a pre-trained neural network[72]. The analysis of papers revealed diverse research objectives, including increased accuracy, texture classification, and decreased runtime. The strengths of the proposed models were highlighted. K Nearest Neighbor (KNN) has been widely used as a classifier for recognition and pattern learning in lung cancer detection, particularly for detecting specific types of lung cancer cells. Support Vector Machine (SVM) has shown high accuracy in texture classification and is effective in distinguishing characteristics of lung cancer. SVM is often used as a classifier along with K Nearest Neighbor to improve the classification of lung cancer[1]. Deep learning models, including CNN, VGG16, VGG19, LeNet, and Inception V3, have demonstrated high accuracy rates in tumor segmentation and lung cancer detection. However, CNN has limitations and requires a large dataset for analyzing visual imagery, and it often requires lesser preprocessing compared to other classification algorithms (as shown in Table 2).

The analysis of the reviewed survey papers revealed significant developments in the detection and classification of lung cancer. Various models and preprocessing techniques have been evaluated, showing notable performance differences. For instance, VGG16 achieved the highest accuracy (99.45%), highlighting the potential of deep learning models for accurate classification. These findings directly address our research questions about the effectiveness of various models and preprocessing techniques in lung cancer detection and classification. Preprocessing techniques like gray scaling, Canny edge detection, and CLAHE significantly enhance image quality, which is crucial for accurate model performance. Transfer learning, particularly with GoogLeNet, emerged as a valuable approach for improving accuracy and efficiency by leveraging pre-trained models. Our results contribute to the broader literature by validating the effectiveness of deep learning models and hybrid approaches in lung cancer detection. This aligns with previous studies but also highlights the need for larger, well-annotated datasets to improve generalizability. Acknowledging these limitations is vital for understanding the scope and implications of our findings.

6. Research Challenges and Opportunities

The surveyed paper highlights research gaps, challenges, and future opportunities in the field of lung cancer detection and classification. The proposed models and algorithms aim to address current limitations in accuracy and performance metrics, while considering training time and resource requirements. There is a need for robust algorithms that can scale efficiently. Hybrid models, which combine multiple algorithms and models, have shown promise in enhancing lung cancer detection. The integration of deep learning approaches, feature selection techniques, and conventional machine learning algorithms within hybrid architectures is a viable approach. However, the limited availability of datasets and the absence of annotated semantic labels pose challenges for model generalization and predictive diagnosis. It is crucial to acquire larger datasets with proper semantic labeling techniques to improve the generalizability of models. Future research should focus on developing robust algorithms that can scale efficiently without compromising accuracy. Additionally, efforts should be made to obtain comprehensive datasets with annotated

 doi:10.20944/preprints202501.2180.v1

semantic labels, as this will significantly contribute to the improvement of model generalization. This survey paper identifies research gaps, challenges, and future opportunities in lung cancer detection and classification, emphasizing the importance of robust algorithms. The utilization of hybrid models and the availability of larger annotated datasets hold potential as solutions to these challenges. By addressing these aspects, significant advancements can be made in the field of lung cancer detection and classification.

The integration of Decision Support Systems and Computer Aided Diagnosis (CAD) systems in clinical settings, alongside the validation of results in the presence of medical experts, requires focused development. The incorporation of validated and verified systems can assist healthcare professionals in clinical environments, leading to improved effectiveness and accuracy of the models. It is essential to address the validation and assessment of proposed models trained and tested on diverse and large-scale datasets, while promoting the adoption of validated interpretability for these models. Despite notable efforts and advancements in recent developments, there is a need to address the limitations and opportunities in the field. By doing so, the accuracy of early-stage lung cancer detection and treatment can be significantly enhanced, while optimizing resource utilization in healthcare organizations. This is summarized in Table 3, which provides an overview of the key aspects to be considered in order to improve the effectiveness of these systems as shown in Table 3.

Table 3. Research Challenges and Opportunities.

Challenges	Opportunities
Limited dataset size and lack of annotated labels	Acquire larger datasets with annotated semantic labels for improved generalizability
Scalability and efficiency of algorithms	Explore hybrid models combining deep learning and conventional ML algorithms
Validation in clinical settings with medical experts	Integrate Decision Support Systems and CAD systems in operational clinical environments
Limited interpretability of models	Validate and assess models on diverse and high-volume datasets for extensive interpretability adoption
Resource utilization in healthcare organizations	Develop robust and efficient algorithms to optimize resource utilization

7. Conclusion

In a nutshell, this systematic literature review comprehensive analysis of various models and preprocessing techniques has highlighted the significant potential of deep learning approaches in lung cancer detection and classification. With VGG16 achieving the highest accuracy and the effective use of transfer learning models like GoogLeNet, our research underscores the importance of preprocessing techniques and the need for larger, well-annotated datasets. By integrating deep learning models with conventional machine learning algorithms, this study address current limitations and enhance classification accuracy. This research not only contributes to the existing body of knowledge but also provides actionable insights for future studies. Future research should focus on developing robust algorithms that can scale efficiently without compromising accuracy, obtaining comprehensive datasets with annotated semantic labels, and integrating decision support systems in clinical settings. Reflecting on this research journey, this study encountered challenges such as limited dataset availability and the need for annotated labels, which were addressed through methods like data augmentation and transfer learning. By addressing these aspects, significant advancements can be made in the field of lung cancer detection and classification, ultimately leading to improved early-stage detection and treatment, and optimizing resource utilization in healthcare organizations. Ending on a high note, this research opens new avenues for further exploration and innovation in lung cancer detection, inspiring future studies to build upon these findings and drive

 doi:10.20944/preprints202501.2180.v1

positive change in the field. The potential for significant improvements in patient outcomes and healthcare efficiency underscores the importance and impact of continued research in this area.

References:

1. S. Nageswaran *et al.*, "Lung cancer classification and prediction using machine learning and image processing," *Biomed Res. Int.*, vol. 2022, 2022.
2. D. Jamil, S. Palaniappan, S. S. Zia, A. Lokman, and M. Naseem, "Reducing the Risk of Gastric Cancer Through Proper Nutrition-A Meta-Analysis.," *Int. J. Online \& Biomed. Eng.*, vol. 18, no. 7, 2022.
3. V. K. Raghu *et al.*, "Validation of a Deep Learning--Based Model to Predict Lung Cancer Risk Using Chest Radiographs and Electronic Medical Record Data," *JAMA Netw. Open*, vol. 5, no. 12, pp. e2248793--e2248793, 2022.
4. B. S. Chhikara and K. Parang, "Global Cancer Statistics 2022: the trends projection analysis," *Chem. Biol. Lett.*, vol. 10, no. 1, p. 451, 2023.
5. M. N. D. J. S. P. Sanjoy Kumar Debnath and S. B. and A. Lokman, "Prediction Model for Gastric Cancer via Class Balancing Techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. 23, no. 01, pp. p53-63, 2023, doi: http://paper.ijcsns.org/07_book/202301/20230108.pdf.
6. H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.
7. M. S. AL-Huseiny and A. S. Sajit, "Transfer learning with GoogLeNet for detection of lung cancer," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, pp. 1078–1086, 2021.
8. D. Chauhan and V. Jaiswal, "An efficient data mining classification approach for detecting lung cancer disease," in *2016 International Conference on Communication and Electronics Systems (ICCES)*, 2016, pp. 1–8.
9. D. Jamil, "Diagnosis of Gastric Cancer Using Machine Learning Techniques in Healthcare Sector: A Survey," *Informatica*, vol. 45, 2022, doi: 10.31449/inf.v45i7.3633.
10. W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, "Lung cancer detection using image processing and machine learning healthcare," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 2018, pp. 1–5.
11. S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung cancer detection using CT scan images," *Procedia Comput. Sci.*, vol. 125, pp. 107–114, 2018.
12. P. K. Vikas and P. Kaur, "Lung cancer detection using chi-square feature selection and support vector machine algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, 2021.
13. B. H. M. der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, p. 102470, 2022.
14. H. Li, Z. Tang, Y. Nan, and G. Yang, "Human treelike tubular structure segmentation: A comprehensive review and future perspectives," *Comput. Biol. Med.*, vol. 151, p. 106241, 2022.
15. Y. Zhao, X. Wang, T. Che, G. Bao, and S. Li, "Multi-task deep learning for medical image computing and analysis: A review," *Comput. Biol. Med.*, vol. 153, p. 106496, 2023.
16. A. Heidari, N. J. Navimipour, M. Unal, and S. Toumaj, "The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions," *Comput. Biol. Med.*, vol. 141, p. 105141, 2022.
17. S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical \& healthcare domains: A systematic literature review," *Comput. Biol. Med.*, p. 107555, 2023.
18. M. Bilal *et al.*, "An aggregation of aggregation methods in computational pathology," *Med. Image Anal.*, p. 102885, 2023.
19. P. Aggarwal, N. K. Mishra, B. Fatimah, P. Singh, A. Gupta, and S. D. Joshi, "COVID-19 image classification using deep learning: Advances, challenges and opportunities," *Comput. Biol. Med.*, vol. 144, p. 105350, 2022.
20. M. T. Abdulkhaleq *et al.*, "Harmony search: Current studies and uses on healthcare systems," *Artif. Intell. Med.*, vol. 131, p. 102348, 2022.

21. X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Med. Image Anal.*, vol. 69, p. 101985, 2021.
22. A. Caruana, M. Bandara, K. Musial, D. Catchpoole, and P. J. Kennedy, "Machine learning for administrative health records: A systematic review of techniques and applications," *Artif. Intell. Med.*, p. 102642, 2023.
23. T. A. Shaikh, T. Rasool, and P. Verma, "Machine intelligence and medical cyber-physical system architectures for smart healthcare: Taxonomy, challenges, opportunities, and possible solutions," *Artif. Intell. Med.*, p. 102692, 2023.
24. I. Li *et al.*, "Neural natural language processing for unstructured data in electronic health records: a review," *Comput. Sci. Rev.*, vol. 46, p. 100511, 2022.
25. C. Thapa and S. Camtepe, "Precision health data: Requirements, challenges and existing techniques for data security and privacy," *Comput. Biol. Med.*, vol. 129, p. 104130, 2021.
26. J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives," *Med. Image Anal.*, vol. 85, p. 102762, 2023.
27. W. He *et al.*, "A review: The detection of cancer cells in histopathology based on machine vision," *Comput. Biol. Med.*, vol. 146, p. 105636, 2022.
28. M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artif. Intell. Med.*, vol. 106, p. 101878, 2020.
29. G. Paliwal and U. Kurmi, "A Comprehensive Analysis of Identifying Lung Cancer via Different Machine Learning Approach," in *2021 10th International Conference on System Modeling \& Advancement in Research Trends (SMART)*, 2021, pp. 691–696.
30. A. Pardyl, D. Rymarczyk, Z. Tabor, and B. Zieliński, "Automating patient-level lung cancer diagnosis in different data regimes," in *International Conference on Neural Information Processing*, 2022, pp. 13–24.
31. S. N. A. Shah and R. Parveen, "An extensive review on lung cancer diagnosis using machine learning techniques on radiological data: state-of-the-art and perspectives," *Arch. Comput. Methods Eng.*, vol. 30, no. 8, pp. 4917–4930, 2023.
32. K. Jabir and A. T. Raja, "A Comprehensive Survey on Various Cancer Prediction Using Natural Language Processing Techniques," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2022, vol. 1, pp. 1880–1884.
33. B. Zhao, W. Wu, L. Liang, X. Cai, Y. Chen, and W. Tang, "Prediction model of clinical prognosis and immunotherapy efficacy of gastric cancer based on level of expression of cuproptosis-related genes," *Heliyon*, vol. 9, no. 8, 2023.
34. J. Lorkowski, O. Kolaszyńska, and M. Pokorski, "Artificial intelligence and precision medicine: A perspective," in *Integrative Clinical Research*, Springer, 2021, pp. 1–11.
35. A. Kazerouni *et al.*, "Diffusion models in medical imaging: A comprehensive survey," *Med. Image Anal.*, vol. 88, p. 102846, 2023.
36. N. Sheng *et al.*, "Data resources and computational methods for lncRNA-disease association prediction," *Comput. Biol. Med.*, vol. 153, p. 106527, 2023.
37. R. Osuala *et al.*, "Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging," *Med. Image Anal.*, vol. 84, p. 102704, 2023.
38. K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," *Comput. Biol. Med.*, vol. 149, p. 106043, 2022.
39. L. Benning, A. Peintner, and L. Peintner, "Advances in and the Applicability of Machine Learning-Based Screening and Early Detection Approaches for Cancer: A Primer," *Cancers (Basel).*, vol. 14, no. 3, p. 623, 2022.
40. S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Comput. Biol. Med.*, vol. 156, p. 106668, 2023.
41. H. Xiao *et al.*, "Deep learning-based lung image registration: A review," *Comput. Biol. Med.*, p. 107434, 2023.
42. E. Çall\i, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Med. Image Anal.*, vol. 72, p. 102125, 2021.

43. H. Jiang, Y. Zhou, Y. Lin, R. C. K. Chan, J. Liu, and H. Chen, "Deep learning for computational cytology: A survey," *Med. Image Anal.*, vol. 84, p. 102691, 2023.
44. D. Painuli, S. Bhardwaj, and others, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Comput. Biol. Med.*, vol. 146, p. 105580, 2022.
45. H. Ramdani, N. Allali, L. Chat, and S. El Haddad, "Covid-19 imaging: A narrative review," *Ann. Med. Surg.*, vol. 69, p. 102489, 2021.
46. Y. Kumar, S. Gupta, R. Singla, and Y. C. Hu, "A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis," *Arch. Comput. Methods Eng.*, vol. 29, no. 4, pp. 2043–2070, 2022, doi: 10.1007/s11831-021-09648-w.
47. Q. Zhang, J. Zhou, and B. Zhang, "Computational traditional Chinese medicine diagnosis: a literature survey," *Comput. Biol. Med.*, vol. 133, p. 104358, 2021.
48. A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Comput. Sci. Rev.*, vol. 40, p. 100370, 2021.
49. F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, p. 102802, 2023.
50. Y. Jing *et al.*, "A comprehensive survey of intestine histopathological image analysis using machine vision approaches," *Comput. Biol. Med.*, p. 107388, 2023.
51. M. Sufyan, Z. Shokat, and U. A. Ashfaq, "Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective," *Comput. Biol. Med.*, p. 107356, 2023.
52. R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. J. Ghoushchi, and M. Bendechache, "Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools," *Comput. Biol. Med.*, vol. 152, p. 106405, 2023.
53. H. F. Ahmad, W. Rafique, R. U. Rasool, A. Alhumam, Z. Anwar, and J. Qadir, "Leveraging 6G, extended reality, and IoT big data analytics for healthcare: A review," *Comput. Sci. Rev.*, vol. 48, p. 100558, 2023.
54. S. Seoni, V. Jahmunah, M. Salvi, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013--2023)," *Comput. Biol. Med.*, p. 107441, 2023.
55. X. Meng and T. Zou, "Clinical applications of graph neural networks in computational histopathology: A review," *Comput. Biol. Med.*, vol. 164, p. 107201, 2023.
56. X. Chen *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, p. 102444, 2022.
57. Z. Liu, Q. Lv, Z. Yang, Y. Li, C. H. Lee, and L. Shen, "Recent progress in transformer-based medical image analysis," *Comput. Biol. Med.*, p. 107268, 2023.
58. S. Pandiyan and L. Wang, "A comprehensive review on recent approaches for cancer drug discovery associated with artificial intelligence," *Comput. Biol. Med.*, vol. 150, p. 106140, 2022.
59. I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Comput. Biol. Med.*, vol. 126, p. 104003, 2020.
60. W. Hu *et al.*, "A state-of-the-art survey of artificial neural networks for whole-slide image analysis: from popular convolutional neural networks to potential visual transformers," *Comput. Biol. Med.*, vol. 161, p. 107034, 2023.
61. R. J. Suji, S. S. Bhadauria, and W. W. Godfrey, "A survey and taxonomy of 2.5 D approaches for lung segmentation and nodule detection in CT images," *Comput. Biol. Med.*, p. 107437, 2023.
62. M. K. Hasan, M. A. Ahamad, C. H. Yap, and G. Yang, "A survey, review, and future trends of skin lesion segmentation and classification," *Comput. Biol. Med.*, vol. 155, p. 106624, 2023.
63. S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, and A. F. Dragoni, "Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey," *Comput. Biol. Med.*, vol. 146, p. 105691, 2022.
64. S. B. Lunge *et al.*, "Therapeutic application of machine learning in psoriasis: A Prisma systematic review," *J. Cosmet. Dermatol.*, vol. 22, no. 2, pp. 378–382, 2023.
65. J. Li, S. : Supervisor, X. Wang, and M. Graeber, "Interpretable Radiomics Analysis of Imbalanced Multi-modality Medical Data for Disease Prediction," no. March, 2022, [Online]. Available: https://ses.library.usyd.edu.au/handle/2123/28187.

 doi:10.20944/preprints202501.2180.v1

66. I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Practical Machine Learning Tools and Techniques Third Edition*. Morgan Kaufmann, 2017.
67. T. J. Saleem and M. A. Chishti, "Exploring the applications of Machine Learning in Healthcare," *Int. J. Sensors Wirel. Commun. Control*, vol. 10, no. 4, pp. 458–472, 2020.
68. U. Kose and J. Alzubi, *Deep Learning for Cancer Diagnosis*. Springer Singapore, 2020.
69. A. Ameri, "A deep learning approach to skin cancer detection in dermoscopy images," *J. Biomed. Phys. Eng.*, vol. 10, no. 6, pp. 801–806, 2020.
70. Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification With Deep Learning: A Systematic Review," *Front. Oncol.*, vol. 12, 2022.
71. A. Shimazaki *et al.*, "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," *Sci. Rep.*, vol. 12, no. 1, p. 727, 2022.
72. M. Nishio *et al.*, "Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization," *PLoS One*, vol. 13, no. 4, p. e0195875, 2018.

Acknowledgements: We would like to thank all the people who prepared and revised previous versions of this document.