American Journal of Interdisciplinary Research and Innovation (AJIRI)

ISSN: 2833-2237 (ONLINE)

VOLUME 4 ISSUE 1 (2025)

PUBLISHED BY

E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Interdisciplinary Research and Innovation (AJIRI)

Volume 4 Issue 1, Year 2025
ISSN: 2833-2237 (Online)
DOI: https://doi.org/10.54536/ajiri.v4i1.3797
https://journals.e-palli.com/home/index.php/ajiri

Lung Cancer Detection and Classification Using Machine Learning: A Literature Review

Lenard Abiel D. Aure[1], Janice Angela V. Paco[1], Jessica Z. Panganiban[1], Cereneo S. Santiago Jr.[1*], Gersom S. Baradi[1]

Article Information

Received: September 18, 2024

Accepted: October 16, 2024

Published: March 01, 2025

Keywords

Detection and Classification, Lung Lesion, Machine Learning, Preprocessing, Sensitivity

ABSTRACT

Lung cancer is one of the pressing public health issues needing accurate and timely diagnosis. Machine learning (ML) is an effective method for analyzing medical images and supporting lung cancer diagnosis and it has significant potential to advance medical practice. This review explored the efficacy of current machine learning methods in detecting and classifying lung cancer. It analyzes the studies on preprocessing techniques, detection accuracy, and classification performance. Preprocessing techniques have significantly improved image quality through noise cancellation and feature enhancement, making it highly efficient. The sensitivity of the machine learning algorithms used for identification of lung cancer is also high, surpassing 90% of some research. This translates to a high probability of correctly identifying actual cancer cases. Support Vector Machines (SVM), Random Forest, and Convolutional Neural Networks (CNN) are among the most effective algorithms. Furthermore, machine learning accurately classifies lung nodules as benign or malignant, exceeding 85% in reported studies. SVM and K-Nearest Neighbor (KNN) are commonly used classification methods with promising results. Through continued research efforts to overcome existing challenges, machine learning could achieve heightened accuracy, seamless integration into clinical practice, and improved outcomes for patients with lung cancer.

INTRODUCTION

Artificial intelligence (AI) refers to the ability of machines to mimic human behavior especially when it comes to processing large amounts of data. Machine learning, a branch of AI, enables computer systems to learn from experience and improve over time without needing to be explicitly programmed (Xie *et al.*, 2021). Most studies regarding detecting lung cancer considered using machine learning to understand better which type of machine learning to use to produce the best result in detecting lung cancer.

One of the most common and deadliest types of cancer is lung cancer. Lung cancer happens when abnormal cells in lung tissue grow uncontrolled. Accurately identifying and classifying various lung cancer types remains challenging despite the progress made in the diagnostic method. In response to this challenge, researchers have shifted to machine learning techniques seeking more efficacy in cancer diagnosis. Several machine learning techniques can be utilized for cancer detection, including deep learning model Convolutional Neural Networks (CNN), which have become popular and effective in solving image classification problems. The accuracy rate has consistently exceeded that of a similar neural network model.

Different machine learning and hybrid approaches are consist of multiple techniques (Azevedo *et al.*, 2024). Machine learning technology has been used to classify and identify whether a person has lung cancer based on a given set of features (Thallam *et al.*, 2020). Researchers have applied preprocessing techniques to the datasets that have undergone these machine learning methods. Numerous studies have indicated the potential of machine learning in identifying and classifying lung nodules and predicting early-stage lung cancer. Researchers are utilizing this technology alongside machine learning techniques to diagnose lung cancer, which would be impossible with the conventional technique (Ingle *et al.*, 2021).

Medical images are crucial tools for diagnosing diseases, which enhance the ability to propose treatments in both early and advanced stages. They have also played a key role in identifying the structures and functions of organs within the body (Galeano *et al.*, 2021). This information aids in diagnosis and helps propose more effective treatments in the initial and advanced stages of the disease. Furthermore, virtual reality (VR) is becoming increasingly integral and widely used in various fields, especially in the medical field (Xu, 2021).

However, key limitations exist when utilizing machine learning for this purpose. First, most AI models need to incorporate clinical information such as previous images or patient history. This data is valuable because it can significantly improve diagnostic accuracy by providing more context for AI analysis. For example, Li *et al.* (2021) discovered that their ICLR model's performance could have improved due to the lack of clinical information like past images or patient history, highlighting a common challenge in the field. Additionally, inconsistencies in data generation can further limit the effectiveness of AI models. Silva *et al.* (2022) found that differences in scanner types and imaging techniques can cause problems when applying models to data from various sources. Despite progress in AI and machine learning, using this technology effectively for lung cancer detection still faces obstacles, particularly in processing diverse datasets. Data

[1] Department of Information Technology, Cavite State Universtiy, Silang Campus, Biga 1 Silang, Cavite, 4112, Philippines
[*] Corresponding author's e-mail: cssantiago@cvsu.edu.ph

from genomics, imaging, and clinical records often come in different formats, making it difficult to standardize them for AI/ML analysis.

Moreover, analyzing such large volumes of data requires significant computational resources. Despite these challenges, AI and machine learning offer a promising approach to detecting lung cancer. Research by Martínez-García and Hernández-Lemus (2022) explores the potential of AI algorithms in analyzing medical images to detect lung nodules, which could indicate cancer. This article provides insights into how machine learning is used in detecting and classifying lung cancer. Therefore, this literature review aims to evaluate and synthesize the accuracy of current machine learning methods in this field, providing a thorough assessment of the state of machine learning in detecting and classifying lung cancer.

MATERIALS AND METHODS

This paper reviewed the design of the comparative literature and explained the research objectives of the researcher before moving on to how relevant literature was obtained and selected. It also demonstrates the compilation of the gathered studies to establish a foundation for analyzing collected data.

The Design

The design was guided by the study of Durach *et al.* (2017), a systematic review that followed 6-step guidelines for performing a systematic literature review in management. First, the researchers start by identifying the research question. It was followed by conducting the research objectives for the study and continued by retrieving potentially related articles and selecting connected literature. We then synthesize relevant information from the literature and then proceed to the final step to report the review result.

Research Objectives

These three objectives guided the review: (1) to synthesize the effectiveness of preprocessing techniques in lung cancer detection, (2) to synthesize the effectiveness of machine learning in detecting lung cancer, and (3) to synthesize the effectiveness of machine learning in classifying lung lesions.

Retrieving and Synthesizing the Literature

The review utilized five credible online databases, IEEE, Science Direct, and Google Scholar (Springer Link and MDPI as the main reliable literature sources). A keyword-based string comprising machine learning, lung cancer, preprocess, detection, and classification was used to search for papers in the abovementioned databases. The initial stage of the selection process involved articles focused on machine learning-based detection of lung cancer. This was followed by thoroughly examining each abstract and reviewing the entire text for its image preprocessing techniques and classification approach. The screening, which involved reading and analyzing the full content of each article, resulted in 19 finalized articles.

From the current year, 2024, published literature within its 6-year range, which is from the year 2018 are collected. A total of nineteen articles are used as these match the purpose of this review. Relevant data from these are listed down, and after analyzing those, they are integrated into Table 1.

RESULTS AND DISCUSSION

Table 1: Summary of reviewed literature.

Author(s)	Preprocessing Effectiveness	Effectiveness of Detection	Effectiveness of Classification
Abdullah *et al.* (2021)	DICOM images converted to JPEG grayscale using MicroDom software. First, a median filter helps clean up any noise in the image, and then Gaussian filters are used to gently smooth everything out.	These results show how accurately different machine learning methods can detect lung cancer: SVM achieved 95.6%, KNN achieved 89.7%, and CNN achieved 92.1% accuracy.	This study produced a result in classifying lung cancer lesions with an accuracy of 95.56 % using SVM, 89.65% using KNN, and 92.11% using CNN.
Hussain *et al.* (2022)	Image enhancement techniques are applied first to improve quality, followed by extracting 22 gray-level co-occurrence (GLCM) features. Machine learning algorithms are then used to classify NSCLC from SCLC.	Lung cancer detection using thresholding and gamma correction, with SVM, decision tree, and Naive Bayes, achieved accuracies from 69.21% to 100%.	With minor image changes, SVM, DT, and Naive Bayes reached 96.49% to 100% accuracy in cancer detection. Upon enhancement, SVM achieved 100% accuracy.
Capizzi *et al.* (2020)	The local variance of each pixel in the X-ray image is calculated to detect and pinpoint potentially dangerous lung nodules, resulting in the creation of a variance matrix.	The Probabilistic Neural Network (PNN) resulted in the best performance with a sensitivity percentage of 95.56.	One hundred images were used, and 440 possible nodules were extracted. PNN used 320 nodules, which resulted in a correct classification of 92.56%.

Chen *et al.* (2018)	An experienced radiation oncologist on non-enhanced CT images with contours conducted radiomic feature extraction	SVM with radiomics achieved 92.85% accuracy. A 1000-time permutation test using a 4-feature signature confirmed strong detection capability.	SVM with radiomics achieved 84% accuracy. Radiomics signatures varied in accuracy from 55.8% to 84%, depending on feature selection.
Faisal *et al.* (2018)	Data cleaning.	Recall/Sensitivity Auto Multi-Layer Perceptron (MLP) - 78.33%, Naïve Bayes - 85.00%, SVM - 79.17%, DT - 78.33%, Gradient Boosted Tree - 90.00%, Neural Network - 71.67% RF - 79.17%, Majority voting MLP+GBT+SVM - 88.57%	Accuracy of Auto MLP)-70.57%, Naïve Bayes - 79.71%, SVM - 60.28%, DT - 70.57%, GradientBoosted Tree - 83.71%, Neural Network - 66.57%, RF - 50%, Majority voting MLP+GBT+SVM - 76.57%
Günaydin *et al.* (2019)	2048x2048 images with 2-byte data were dimensionally reduced via PCA, boosting efficiency, aiding visualization, and enhancing machine learning performance.	Before processing: DT 96%, KNN 86.36%, SVM 85%. After processing: KNN dropped to 78.26%, SVM 72.41%, but ANN rose to 91.30%.	Before, DT led with 93.24%, KNN 74.32%, and SVM 50%. After: KNN rose to 75.68%, SVM fell to 55.41%, DT dropped to 79.97%, and ANN 82.43%.
Hoque *et al.* (2020)	Contour stretching enhances contrast. The median filter reduces noise, preserving edges. Segmentation by Otsu's method creates a binary image through thresholding.	Support Vector Machine (SVM) showed an effectiveness of 94.87%, indicating its high accuracy in detecting lung cancer.	SVM achieved 95% accuracy in classifying lung lesions, showing its strong ability to distinguish different types of abnormalities in lung images.
Hrizi *et al.* (2023)	The blue channel is extracted for nuclear staining to count cells. Morphological operations like opening and closure are applied to remove noise and fill holes in the image.	SVM achieved 97.33%, XGBoost 99.1%, Random Forest 98.79%, and Decision Tree 98.76% accuracy rates, showing strong performance across various classifiers.	SVM achieved 95.1%, XGBOOST 98.6%, Random Forest 98.1%, and Decision Tree 95.6% accuracy in classifying benign and malignant cases.
Ingle *et al.* (2021)	Preprocessing technique: Min-max normalization, which rescales data between zero and one. Ensuring standardized data samples for improved analysis and interpretation of data.	DT, Random Forest, and K-nearest Neighbor achieved average accuracies of around 73%, while AdaBoost demonstrated higher accuracy, averaging around 82%	DT, Random Forest, KNN, and AdaBoost detected lung cancer with accuracies ranging from 80.55% to 99.54%, averaging around 87% to 90.74%.
Islam *et al.* (2019)	Median filtering removed noise, the Gabor filter was enhanced, global thresholding was converted to binary, morphological opening was followed, and ROI selection was followed.	(GLCM): SVM 71%, KNN 57%, RF 43%, Gaussian Naïve Bayes 57%. Statistical Parametric: SVM 83%, KNN 68%, RF 63%, Gaussian Naïve Bayes 74%.	(GLCM): SVM 73.68%, KNN 68.42%, RF 57.89%, Gaussian Naïve Bayes 68.42%. Statistical Parametric: SVM 78.95%, KNN 68.42%, RF 63.15%, Gaussian Naïve Bayes 73.68%.
Jayaraj & Sathiamoorthy (2020)	The median filter removes noise from the grayscale CT scan image for better cancer identification. Then, a Gaussian filter smoothens the image and reduces noise speckles.	The proposed method (RF as a classifier) - 90.85% SVM - 79.87%, MLP - 76.59%, RBF Network - 90.35%, KNN - 87.5%.	The proposed method (RF as a classifier) - 89.9% SVM - 86.6%, MLP - 82.87%, RBF Network - 83.89%, KNN - 89%.

Kaur *et al.* (2020)	The 2D image was converted into grayscale and enhanced to remove noise, redundancy, and blueness using Median and log Gabor filters.	Recall - 100% (calculated).	An accuracy of 93.3% using ANN to classify lung cancer stages has been achieved.
Makaju *et al.* (2018)	DICOM images were converted to grayscale JPEGs. Median filtering removes salt and pepper noise, while Gaussian filtering smooths speckle noise from CT scans.	Support Vector Machine (SVM) performed accurately and achieved 86.6% effectiveness in detecting lung cancer.	This study compares two models for lung cancer detection. The current model scored 88.4%, while the proposed model improved to 92%.
Maleki & Akhavan Niaki (2023)	Images were resized to 512 by 512 pixels, and then a median filter was applied to lessen the noise and quality of the images.	Classifiers (GB, RF, SVM) tested on PCA, LDA, and GA datasets, with accuracies: GB - 95%, RF - 78% (LDA), 85% (GA), SVM - 72% (PCA), 95% ((LDA), 72% (GA).	Three classifiers (GB, RF, SVM) were applied to 3 datasets. For PCA:GB 95%, RF 82%, SVM 73%. The LDA: GB/RF 78%, SVM 95%. The GA:GB 82%, RF 85%, SVM.
Nair *et al.* (2024)	CT scan images were converted to JPG for faster processing and filtered using an anisotropic nonlinear diffusion method to enhance nodule texture quality.	RW Segmentation with ANN achieved 0.35%; An enhanced method reached 9.78%. RW with Random Forest and improved RW both hit 99.98%.	RW segmentation with ANN: 96.82%, improved RW with ANN: 94.76%. RW with random forest: 99.03%, improved RW with ANN: 94.76%, with random forest: 99.65%.
Patra (2020)	RBF, KNN, Naive Bayes, and J48 classifiers were evaluated for accuracy and reliability through 10-fold cross-validation.	Machine learning accurately classified lung cancer between 75% and 81.25% using methods like KNN, Naive Bayes, RBF, and J48.	WEKA tool's results for lung cancer detection: KNN 0.75, Naive Bayes 0.775, RBF 0.813, and J48 7.68, measured on a scale from 0-1.
Rehman *et al.* (2021)	CT scan images were converted to grayscale and were put into the texture-based LBP technique to encode lung CT scans' global features.	Using Three Patch LBP (TPLBP) with discrete cosine transform (DCT), SVM achieved 86% accuracy, While KNN reached 82.4%	SVM attained 93% accuracy, and KNN reached 91% when utilizing TPLBP combined with DCT for image analysis.
Singh & Gupta (2018)	ROI extraction, grayscale conversion, Gaussian blur denoising, Outsu's adaptive thresholding, and morphological opening are applied	This study tested machine-learning techniques for lung cancer classification on a scale of 0-1. Accuracy varied from 0.6279 and 0.8916	KNN 89.95%, SVM 59.87%, DT 80.84%, Multinomial Naive Bayes 60.25%, SGD 64.01%, Random Forest 85.74%, and MLP 91.58%
Woźniak *et al.* (2018)	A 3x3 pixel window calculates local variance, then a boundary tracking algorithm identifies potential diseased tissues in X-ray images.	The Probabilistic Neural Network (PNN) detected 95% accuracy, demonstrating its effectiveness in detecting lung cancer.	Probabilistic Neural Network (PNN) achieved 92% accuracy in classifying lung nodules.

Effectiveness of pre-processing techniques.

Data preprocessing is the essential step of cleaning, organizing, and transforming raw data into a format that a machine learning model can understand and learn from effectively. The data preprocessing can be applied to the medical imaging data (x-ray, CT scans, and MRI scans) before experts and machine learning algorithms analyze them to search for the presence of lung cancer in the enhanced data. Ingle *et al.* (2021) discussed that using machine learning techniques to diagnose lung cancer can significantly improve accuracy. Its algorithm can spot signs of lung cancer that conventional processes can miss. This focus on potential cancer markers, facilitated by data preprocessing, reduces diagnostic errors.

Many researchers have used ML algorithms to predict lung cancer, which is impossible with the conventional technique. Pandiangan *et al.* (2019) proposed five stages of artificial neural networks: First, it reduces noise in

medical scans. Second, it enhances the image to make relevant features more prominent. Third, the network segments the lung region, separating it from the rest of the image. Fourth, object edge detection pinpoints potential areas of concern. Fifth, the network performs tumor boundary recognition, precisely outlining these areas. Once these five stages are completed, the neural network is able to determine whether the segmented lung regions are cancerous or not.

Learning about different preprocessing techniques creates many options for researchers to use for their studies. Most studies used noise reduction, image enhancement, grayscale method, and median filtering techniques. For instance, Hussain *et al.* (2022) studied different image filtering methods and levels, yielding a different result. The highest accuracy achieved without image enhancement methods was 99.89% using the machine learning algorithms SVM, polynomial, and RBF. On the other hand, by applying image enhancement techniques such as image adjustment, contrast stretching at a threshold of 0.02 to 0.98, and gamma correction with a gamma value of 0.9, 100% accuracy was achieved using RBF, SVM, and polynomial kernels. This demonstrates that these image enhancement methods significantly improve the accuracy of lung cancer detection.

Preprocessing may enhance the picture's quality by removing distracting features and enhancing future processes' results. Image analysis becomes more efficient by isolating foreground objects from their background. By extracting features such as intensity, texture, and color from each pixel, a unique mathematical representation of the image is created. After identifying objects, they can be labeled in the image for deeper analysis (Hussein *et al.*, 2022). Kumar *et al.* (2019) explored various evolutionary image segmentation techniques to assess whether CT scan images show signs of lung cancer. These methods likely aimed to isolate lung nodules from surrounding tissues.

Effectiveness in detecting lung cancer

Sensitivity, often referred to as recall, measures the proportion of actual positive cases that the algorithm correctly identifies. It indicates how well the algorithm classifies images with benign or malignant nodules, making it a crucial factor for any diagnostic tool to be considered practical and effective (Singh & Gupta, 2019). One of the two most important mechanisms is False Positive Reduction, the other being pulmonary nodule detection for the early diagnosis and precise management of lung cancer (Capizzi *et al.*, 2020). Detecting pulmonary nodules which are the small spots on the lungs that can be cancerous or benign is essential for early detection and can potentially save lives. According to Al Mohammad *et al.* (2017), there has been increasing convolutional neural networks (CNNs) research that applies deep learning (DL) models in computer-aided diagnosis (CAD) systems to enhance accuracy, reduce false positives, and improve execution time in detecting lung tumors, particularly. Like traditional feature-based CAD systems, these deep learning (DL) models generally follow three key steps: nodule detection and segmentation, feature extraction from the nodules, and clinical decision-making (Al Mohammad *et al.*, 2017). Various techniques are employed for lung cancer detection, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Random Forest (RF), and Convolutional Neural Network (CNN). SVM is a classification algorithm designed to create a decision boundary between two categories to predict labels from feature vectors (Buty *et al.*, 2016). KNN, a simple non-parametric method, is often used when there is limited prior knowledge of the data (Huang *et al.*, 2018). Artificial neural networks, using input features to assign weighted values, predict output (Shi *et al.*, 2011). Random Forest (RF) is a tree method in which trees are grown by binary recursive splitting of right-censored data (Rashidi *et al.*, 2019).

As a result, the researchers gathered sensitivity percentages from various studies to analyze. These studies revealed different levels of accuracy with machine learning methods. Hrizi *et al.* (2023) found that SVM achieved 97.33% and Random Forest achieved 98.79% accuracy rates in detecting lung cancer, demonstrating strong performance across different classifiers. Abdullah *et al.* (2023) reported that SVM achieved 95.6%, KNN 89.7%, and CNN 92.1% accuracy in their study on lung cancer detection. Kaur *et al.* (2018) and Makaju *et al.* (2017) achieved 100% sensitivity using SVM and ANN, respectively.

Although machine learning has been widely adopted for lung cancer detection in both clinical settings and research, there are still notable challenges to overcome. Convolutional Neural Networks (CNNs), in particular, have shown great success in identifying lung cancer from medical images, but further improvements are needed to address existing limitations. Deep learning, an advanced machine learning, enhances this capability with more complex algorithms. One critical challenge is balancing sensitivity—detecting as many actual cases as possible—while reducing false positives to avoid unnecessary tests. Machine learning methods like SVM and KNN have their strengths and weaknesses in detecting lung cancer. While some methods show promising results, issues such as overfitting and difficulty interpreting results remain. It is important to validate these methods rigorously across diverse populations and image types to ensure fairness and effectiveness.

Effectiveness in classifying lung lesions

Lung cancer can be classified into non-cancerous (benign) or cancerous (malignant) nodules, which form due to abnormal lung cell growth, leading to the development of lung nodules (Hrizi *et al.*, 2023). The two main types of lung cancer are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) (Abdullah *et al.*, 2023). SCLC, known for its aggressive nature and strong association with smoking, accounts for 15-20%

of all primary lung cancers. Monkam *et al.* (2019) noted that substantial progress has been made in the automated differentiation of pulmonary nodules from other lung lesions, like non-nodules. However, not all detected lung nodules are cancerous. Thus, the accuracy of the machine learning algorithms is crucial for validating their effectiveness on lung lesion classification.

Among the 19 studies presented in the table, the highest accuracy of 100% was achieved using SVM, RBF, and polynomial kernels, all with image enhancement methods (Woźniak *et al.*, 2018). This was followed by 98.6% by Hrizi *et al.* (2023) using XGBoost, 92.56% using a PNN with fuzzy logic Capizzi *et al.*, (2020), and 92% with the PNN (Woźniak *et al.*, 2018).

Hrizi *et al.* (2023) tested four machine learning algorithms: SVM, XGBoost, Random Forest, and Decision Tree Among these, XGBoost attained the highest accuracy of 98.6% for classifying lung lesions Their study involves two main processes The first process includes preprocessing, segmentation, feature extraction, and nodule detection Preprocessing included extracting the blue channel for cell counting and using morphological operations to improve image quality by removing noise and filling holes The second process involves classifying detected lung lesions as either benign or malignant This classification was performed on a dataset of 110 CT scans, categorized into 40 normal, 15 benign, and 55 cancerous nodules Hussain *et al.* (2022) achieved 100% accuracy using machine learning algorithms such as SVM, RBF, and polynomial kernels They tested various image filtering methods and levels to detect lung cancer, distinguishing between small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) Their findings showed that enhancing images using contrast stretching and gamma correction significantly improved prediction performance when analyzing 945 images from the Lung Cancer Alliance MRI dataset, which included 377 NSCLC and 568 SCLC cases.

The study of Woźniak *et al.* (2018) proposed a new classification method for more accurately determining the presence or absence of lung carcinomas in chest X-ray images. Their approach used a straightforward segmentation method with a probabilistic neural network, achieving an overall correct classification rate of 92%. However, they have found many false nodules after the segmentation stage. Therefore, they used PNN as their classifier to discriminate the true nodules.

For the neural network to function effectively, it is crucial to accurately describe the geometric properties of lung nodules. This was accomplished using feature extraction based on the momentum method. Their approach was tested on a dataset of 504 real two-exposure dual-energy subtraction chest radiographs, with 404 cases for training and 100 for testing. Despite some misclassification errors, where the network confused false nodules with true ones (6%) and vice versa (2%), the study yielded positive results. Their method, while relatively simple, proved effective, detecting low-contrast nodules. On the other hand, Capizzi *et al.* (2020) employed a fuzzy logic segmentation method with a PNN classifier, tested on images from clinical trials at Zagglębiowskie Oncology Centre, Poland, achieving 92.56% accuracy. They emphasized the clarity and simplicity of their methodology as a key advantage.

Upon reviewing various articles, machine learning algorithms have displayed impressive accuracy in classifying lung lesions, with some achieving 100% success rates when combined with specific algorithms and image enhancement techniques. This advancement holds significant potential in enhancing the early detection of lung cancer and reducing misdiagnoses. However, it is important to recognize that these outcomes are based on particular datasets and methodologies, underscoring the need for further validation in larger and more diverse patient records. Continuous improvement of machine learning algorithms is crucial to ensure their reliability and applicability in clinical scenarios.

CONCLUSION

This systematic literature review examines lung cancer detection and classification using machine learning, offering insights and recommendations for future research. Key contributions include a review of preprocessing techniques—such as noise reduction, image enhancement, grayscale conversion, and median filtering—that are critical for accurately analyzing lung cancer images. Preprocessing isolates and segments features like nodules and tumors, enabling machine learning algorithms to classify lung cancer types, including non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Commonly employed machine learning methods include CNN, SVM, KNN, ANN, and Random Forest, each varying in detection sensitivity, allowing future research to identify the most effective method.

Challenges like misclassification highlight the need for integrating clinical data (e.g., prior images, patient history) to improve system precision and reliability. Future research should explore compatible preprocessing and machine learning techniques to minimize detection errors. The review advocates for collaboration between researchers and clinicians to advance machine learning applications in lung cancer detection, with the goal of enhancing diagnostic accuracy and patient outcomes. By addressing current limitations and refining techniques, machine learning can become a crucial tool in clinical settings, potentially saving lives through early, accurate detection.

REFERENCES

Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2023). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Quban Academy Journal, 1*(2), 141–149. https://doi.org/10.48161/qaj.v1n2a58

Al Mohammad, B., Brennan, P. C., & Mello-Thoms, C. (2017). A review of lung cancer screening and the role of computer-aided detection. *Clinical Radiology, 72*(6),

433-442. https://doi.org/10.1016/j.crad.2017.01.002

Azevedo, B. F. A., Rocha, A. M. A. C., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: A systematic literature review. *Machine Learning, 113*, 4055–4097. https://doi.org/10.1007/s10994-023-06467-x

Buty, M., Xu, Z., Gao, M., Bagci, U., Wu, A., & Mollura, D. J. (2016). Characterization of lung nodule malignancy using hybrid shape and appearance features. In S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2016* (Vol. 9900, pp. 662-670). Springer. https://doi.org/10.1007/978-3-319-46720-7_77

Capizzi, G., Mazzocchi, M., & Plesa, S. (2020). Small lung nodules detection based on fuzzy-logic and probabilistic neural network with bioinspired reinforcement learning. *IEEE Transactions on Fuzzy Systems, 28*(6), 1178–1189. https://doi.org/10.1109/TFUZZ.2019.2952831

Chen, C. H., Wu, H. T., Chang, H. C., Huang, Y. Y., & Chen, C. C. (2018). Radiomic features analysis in computed tomography images of lung nodule classification. *PLoS One, 13*(2). https://doi.org/10.1371/journal.pone.0192002

Durach, C. F., Kembro, J., & Wieland, A. (2017). A new paradigm for systematic literature reviews in supply chain management. *Journal of Supply Chain Management, 53*(4), 67–85.

Faisal, M. I., Bashir, S., Khan, Z. S., & Hassan Khan, F. (2018). An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICEEST.2018.8643311

Galeano Galeano, S. D., Esteban Mora Gonzalez, M., & Espinosa Medina, R. A. (2021). Alternative tool for the diagnosis of diseases through virtual reality. In *2021 IEEE 2nd International Congress of Biomedical Engineering and Bioengineering (CI-IB&BI)* (pp. 1–4). IEEE. https://doi.org/10.1109/CI-IBBI54220.2021.9626088

Günaydin, Ö., Günay, M., & Şengel, Ö. (2019). Comparison of lung cancer detection algorithms. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1–4). Istanbul, Turkey. https://doi.org/10.1109/EBBT.2019.8741826

Hoque, A., Farabi, A. K. M. A., Ahmed, F., & Islam, M. Z. (2020). Automated detection of lung cancer using CT scan images. In *2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 1030–1033). Dhaka, Bangladesh. https://doi.org/10.1109/TENSYMP50017.2020.9230861

Hrizi, D., Tbarki, K., Attia, M., & Elasmi, S. (2023). Lung cancer detection and nodule type classification using image processing and machine learning. In *2023 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 1154–1159). Marrakesh, Morocco. https://doi.org/10.1109/IWCMC58020.2023.10183237

Huang, S., Cai, N., Pacheco, P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics, 15*(1), 41-51. https://doi.org/10.21873/cgp.20063

Hussain, L., Alsolai, H., Hassine, S. B. H., Nour, M. K., Duhayyim, M. A., Hilal, A. M., Salama, A. S., Motwakel, A., Yaseen, I., & Rizwanullah, M. (2022). Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features. *Applied Sciences, 12*(13), 6517. https://doi.org/10.3390/app12136517

Ingle, K., Chaskar, U., & Rathod, S. (2021). Lung cancer types prediction using machine learning approach. In *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1–6). IEEE. https://doi.org/10.1109/CONECCT52877.2021.9622568

Islam, M., Mahamud, A. H., & Rab, R. (2019). Analysis of CT scan images to predict lung cancer stages using image processing techniques. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 961–967). Vancouver, BC, Canada. https://doi.org/10.1109/IEMCON.2019.8936175

Jayaraj, D., & Sathiamoorthy, S. (2019). Random forest based classification model for lung cancer prediction on computer tomography images. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 100–104). Tirunelveli, India. https://doi.org/10.1109/ICSSIT46314.2019.8987772

Kaur, L., Sharma, M., Dharwal, R., & Bakshi, A. (2018). Lung cancer detection using CT scan with artificial neural network. In *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)* (pp. 1624–1629). Bhubaneswar, India. https://doi.org/10.1109/ICRIEECE44171.2018.9009244

Kumar, K. S., Venkatalakshmi, K., & Karthikeyan, K. (2019). Lung cancer detection using image segmentation by means of various evolutionary algorithms. *Computational and Mathematical Methods in Medicine, 2019*, 4909846. https://doi.org/10.1155/2019/4909846

Li, K., Huang, Y., Cheng, H., Zhang, Y., & Liu, J. (2021). Assessing the predictive accuracy of lung cancer, metastases, and benign lesions using an artificial intelligence-driven computer-aided diagnosis system. *Quantitative Imaging in Medicine and Surgery, 11*(8), 3629–3642. https://doi.org/10.21037/qims-20-1314

Maleki, N., & Akhavan Niaki, S. T. (2023). An intelligent algorithm for lung cancer diagnosis using extracted features from computerized tomography images. *Healthcare Analytics, 3*, 100150. https://doi.org/10.1016/j.health.2023.100150

Makaju, S., Prasad, P. W., Alsadoon, A., Singh, A. K., & Elchouemi, A. (2018). Lung cancer detection using CT scan images. *Procedia Computer Science, 125*, 107–114. https://doi.org/10.1016/j.procs.2017.12.016

Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in Medicine, 8,* 784455. https://doi.org/10.3389/fmed.2021.784455

Monkam, P., Qi, S., Ma, H., Gao, W., Yao, Y., & Qian, W. (2019). Detection and classification of pulmonary nodules using convolutional neural networks: A survey. *IEEE Access, 7*, 78075-78091. https://doi.org/10.1109/ACCESS.2019.2920980

Nair, S. S., Meena Devi, V. N., & Bhasi, S. (2024). Enhanced lung cancer detection: Integrating improved random walker segmentation with artificial neural network and random forest classifier. *Heliyon, 10*. https://doi.org/10.1016/j.heliyon.2024.e29032

Pandiangan, T., Bali, I., & Silalahi, A. R. J. (2019). Early lung cancer detection using artificial neural network. *Atom Indonesia, 45*(1), 9–15. https://doi.org/10.17146/aij.2019.860

Patra, R. (2020). Prediction of lung cancer using machine learning classifier. In N. Chaubey, S. Parikh, & K. Amin (Eds.), *Computing Science, Communication and Security* (Vol. 1235, pp. 101–112). Springer. https://doi.org/10.1007/978-981-15-6648-6_11

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. *Academy Pathology, 6*, Article 2374289519873088. https://doi.org/10.1177/2374289519873088

Rehman, A., Kashif, M., Abunadi, I., & Ayesha, N. (2021). 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)* (pp. 101–104). Riyadh, Saudi Arabia. https://doi.org/10.1109/CAIDA51941.2021.9425269

Shi, P., Ray, S., Zhu, Q., & Chen, X. (2011). Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics, 12,* Article 375. https://doi.org/10.1186/1471-2105-12-375

Silva, F., Santos, J. A., & Oliveira, A. R. (2022). Towards machine learning-aided lung cancer clinical routines: Approaches and open challenges. *Journal of Personalized Medicine, 12*(3), 480. https://doi.org/10.3390/jpm12030480

Singh, G. A. P., & Gupta, P. K. (2019). Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. *Neural Computing and Applications, 31*, 6863–6877. https://doi.org/10.1007/s00521-018-3518-x

Thallam, C., Peruboyina, A., Raju, S. S. T., & Sampath, N. (2020). Early stage lung cancer prediction using various machine learning techniques. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1285–1292). IEEE. https://doi.org/10.1109/ICECA49313.2020.9297576

Woźniak, M., Połap, D., Capizzi, G., Lo Sciuto, G., Kośmider, L., & Frankiewicz, K. (2018). Small lung nodules detection based on local variance analysis and probabilistic neural network. *Computational Methods and Programs in Biomedicine, 161*, 173–180. https://doi.org/10.1016/j.cmpb.2018.04.025

Xie, Y., Wang, Y., Gu, Q., Zhao, Y., Chen, Y., & Liu, L. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational Oncology, 14*(1), 100907. https://doi.org/10.1016/j.tranon.2020.100907

Xu, Z. (2021). Design of cancer detection system based on CNN model and virtual reality with NLP voice output. In *2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)* (pp. 277–284). IEEE. https://doi.org/10.1109/AINIT54228.2021.00062