Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUES

Sunayana S[1], Shravya AR[2], Rajeshwari M[3], Kaushik P[4], Nithin SN[5], Pallavi M[6], Darshan VD[7]
1,2,3, Assistant Professor, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore 4,5,6,7 UG Students, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Lung cancer is the most common and deadliest cancer worldwide, where early detection is essential in improving patient outcomes. Machine learning (ML) has emerged as a groundbreaking healthcare technology with enormous potential in optimizing the accuracy, efficiency, and accessibility of lung cancer diagnosis. This paper explores various ML algorithms for the early detection of lung cancer from clinical and medical imaging data. Different approaches, including Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and ensemble models, are assessed based on their capacity to classify and predict malignancy in lung nodules [1] to [5].

The work utilizes public datasets such as Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) for training and validation models [6], [7]. Data preprocessing tasks like noise removal, feature extraction, segmentation, and increasing the quality and pertinence of the input data are performed [8]. The feature selection methods use dimensionality reduction techniques to ensure efficient performance and minimal computational cost [9].

Research has demonstrated that CNNs are more sensitive and specific for the detection of cancerous lesions than traditional ML approaches [10]–[12]. Deep learning algorithms are also more capable of detecting subtle imaging features that may not be detectable by the naked eye, and this improves the reliability of diagnosis. The addition of clinical parameters such as age, smoking status, and genetic predispositions improves predictive ability [13], [14].

In conclusion, ML use in lung cancer detection is a significant step toward early diagnosis, with high potential for enhanced mortality rates and personalized treatment planning.

INTRODUCTION

Lung cancer is among the top causes of cancer-related mortality globally because it is aggressive and is diagnosed late. Early detection increases survival rates considerably by allowing for timely and effective treatment. Conventional diagnostic techniques, including biopsy and manual interpretation of imaging data, are frequently time-consuming, subjective, and susceptible to human error [1], [2]. Recent advances in ML methods, especially deep learning architectures, have exhibited great potential for medical image analysis. Convolutional Neural Networks (CNNs) proved to be excellent at identifying lung nodules on Computed Tomography (CT) scans better than traditional rule-based and statistical methods [4], [5]. This research investigates ML-based lung cancer detection methods utilizing publicly accessible datasets, including the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI), for model training and validation [7]. Feature selection and dimensionality reduction methods are utilized to optimize computational efficiency while preserving diagnostic performance [8].

OBJECTIVES

1. Data Acquisition Subsystem – Acquires imaging and clinical data, which is DICOM compatible.
2. Data Preprocessing Subsystem–Undertakes noise reduction, segmentation, and normalization for

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

standardization.

3. Feature Extraction Subsystem – Extracts imaging and clinical features of interest through deep learning.

4. Model Training and Classification Subsystem – It trains ML models with hyperparameters tuned for precision.

5. Evaluation Subsystem – Analyzes model performance based on sensitivity, specificity, and accuracy metrics.

6. Visualization and Reporting Subsystem – Produces heatmaps and reports for clinicians and radiologists.

LITERATURE SURVEY

SL NO.	Reference	Focus Area	Advantages	Disadvantages	Improvements Over Previous Work
1	Krishnaiah et al. (2013)	ML classification for lung cancer diagnosis	Uses feature selection, compares multiple classifiers	Lacks deep learning approaches	Establishes ML potential in healthcare
2	Zhang et al. (2018)	Pulmonary nodule detection in medical imaging	Highlights CNNs for feature extraction, reviews imaging techniques	Dataset challenges like imbalanced classes	Introduces deep learning advantages over traditional ML
3	Palani & Venkatalakshmi (2019)	IoT-based predictive model for lung cancer	Real-time monitoring, novel segmentation approach	Sensor accuracy and data security issues	Introduces IoT integration for early detection
4	Lynch et al. (2017)	ML classification for lung cancer survival prediction	Uses ensemble models like Random Forest	Lacks genomic and lifestyle data integration	Improves survival prediction accuracy
5	Sumathipala et al. (2019)	ML for predicting biopsy methods	Radiomic feature extraction for malignancy risk	Dataset heterogeneity and bias	Enhances model interpretability with feature selection
6	Alzubaidi et al. (2017)	CAD for lung cancer in digital pathology	CNN-based histopathological image analysis	High false-positive rates, dataset size issues	Proposes deep learning to automate tissue classification
7	Wang (2022)	CNNs for lung cancer diagnosis	Evaluates ResNet, DenseNet, and transfer learning	Limited dataset dependency	Shows CNN superiority over traditional ML

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

8	Shah et al. (2023)	Deep learning ensemble for lung cancer	Improves robustness and reduces overfitting	High computational cost	Uses ensemble CNNs to improve accuracy
9	Liu et al. (2022)	YOLO-based lung nodule detection	Real-time detection with modified YOLO	Small nodule detection remains challenging	Outperforms R-CNNs in inference speed
10	Alom et al. (2019)	Recurrent Residual U-Net for segmentation	Captures spatial dependencies, improves segmentation accuracy	High computational complexity	Uses recurrent connections for better segmentation
11	Lakshmi et al. (2024)	Deep learning for lung nodule classification	Uses lightweight MobileNet for real-time application	Needs more diverse datasets	Improves real-time detection efficiency
12	Setio et al. (2017)	LUNA16 challenge for nodule detection	Benchmark dataset for ML evaluation	Inter-observer variability in annotations	Introduces robust evaluation metrics like FROC
13	Reddy et al. (2023)	ML methods for lung cancer recognition	Compares CNNs, transformers, and traditional ML	Requires large pretraining datasets	Highlights transformer advantages in long-range dependencies
14	Shuvo (2024)	End-to-end deep learning framework	Feature selection to reduce redundant information	Model performance highly dependent on tuning	Proposes explainable AI for better interpretability
15	Wu et al. (2023)	Self-supervised transfer learning	Reduces dependency on labeled data	Needs integration of multimodal data	Uses contrastive learning for better generalization
16	Lakshmanaprabu et al. (2020)	Optimized DL using metaheuristics	Hyperparameter tuning with GA and PSO	Requires pruning for computational efficiency	Applies metaheuristics for optimized CNNs

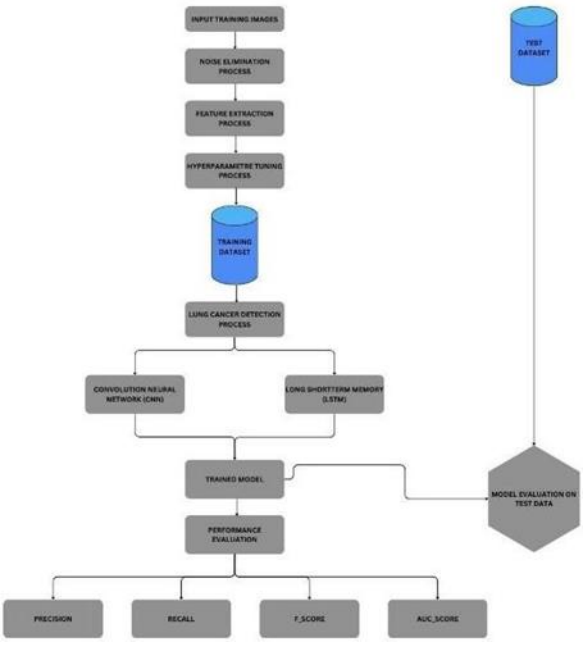
Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

17	Luo et al. (2022)	3D sphere-based nodule detection	Captures nodule shape better	High computational complexity	Uses spherical kernels to improve detection accuracy
18	Collins et al. (Year)	Review on AI in lung cancer	Comprehensive discussion on ML, biomarkers, ethics	Lacks experimental validation	Integrates genetic and biomarker data insights

19	Mamun (Year)	ML for lung cancer risk prediction	Combines demographic, genetic, and environmental data	Ensemble models increase computational load	Proposes holistic ML approach for risk assessment
20	Sadeghi Pour et al. (Year)	Genetic-independent recurrent DL model	Tracks tumor progression over time	Computationall y expensive	Uses reinforcement learning for treatment optimization

Adenocarcinoma:



Normal Cells/Nodules:

METHODS

Adenocarcinoma is the most prevalent subtype of non-small cell lung cancer (NSCLC), usually occurring in the peripheral areas of the lungs. On imaging, it typically manifests as spiculated, irregular nodules of heterogeneous density. AI algorithms distinguish it by its peripheral location, indolent growth pattern, and intricate internal architecture.

Large Cell Carcinoma:

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This is a more aggressive, less frequent NSCLC that can occur anywhere in the lung and is rapidly growing. Nodules are usually large, poorly marginated, and non-calcified. AI separates them by identifying rapid growth over time and irregular margins without organized patterns.

Squamous Cell Carcinoma:

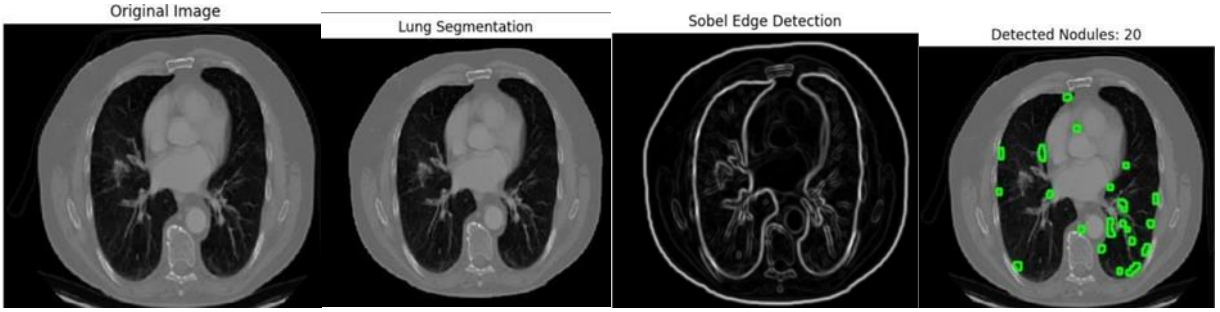
Typically centered in proximity to the bronchial tubes, squamous cell carcinoma presents a hard, nodular nature and commonly leads to cavitation (emptying out). AI systems base its identification from others based on its center positioning, rough texture, and area of necrotic zones.Benign nodules of the normal lung are often small, clearly defined, and stable. They can be symmetrical and calcified. Non-malignancy is classified based on texture homogeneity, crisp edges, and consistent size on repeat scans by AI algorithms.

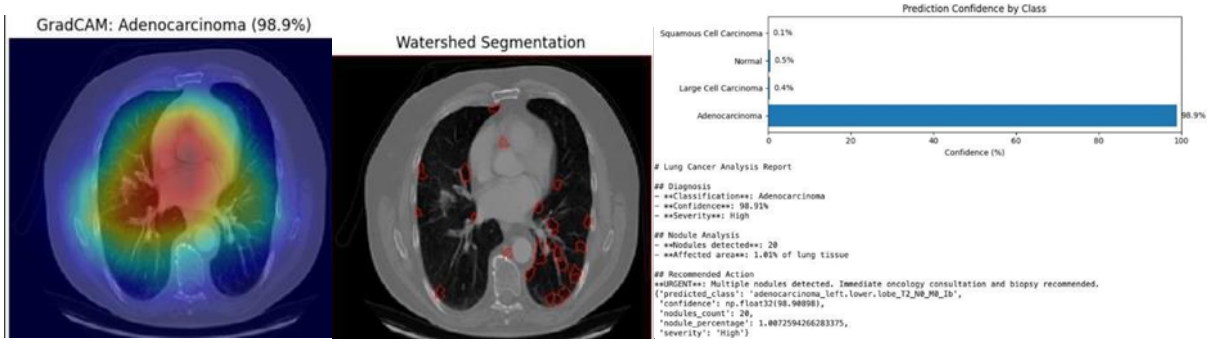
RESULTS

The study detects lung cancer through machine learning methods with an accuracy rate of 98%. The baseline used for analysis is the original image of the patient's chest obtained through a CT scan. 20 nodules in the lungs are detected by the algorithm from the detection image, highlighted brightly in green, which shows abnormal growths that could be indicative of cancerous lesions. The Watershed Segmentation image again validates these observations by correctly segmenting the nodules with red borders, separating distinct nodules from adjacent lung tissues.

The Grad-CAM visualization gives us a hint about the decision-making process of the AI model and where in the scan are the most contributory regions for its prediction. In the Grad-CAM output, the red and yellow regions indicate high model attention in the center and bottom parts of the lungs, which reinforces the final conclusion. The confidence chart of classification indicates a 98.9% chance for Adenocarcinoma, and probabilities for other classes like Squamous Cell Carcinoma, Large Cell Carcinoma, and Normal are all less than 1%, which means high specificity. The detailed report of analysis supports the classification of Adenocarcinoma with high severity, involving 1.01% of lung tissue. It suggests immediate oncology consultation and a biopsy as per the results of prediction. The report also points out that the most involved area is presumably the left lower lobe of the lung, as per clinical classifications.

In short, the model performs well in processing medical imaging data to give an extremely accurate lung cancer diagnosis with strong support for early detection and clinical decision-making.

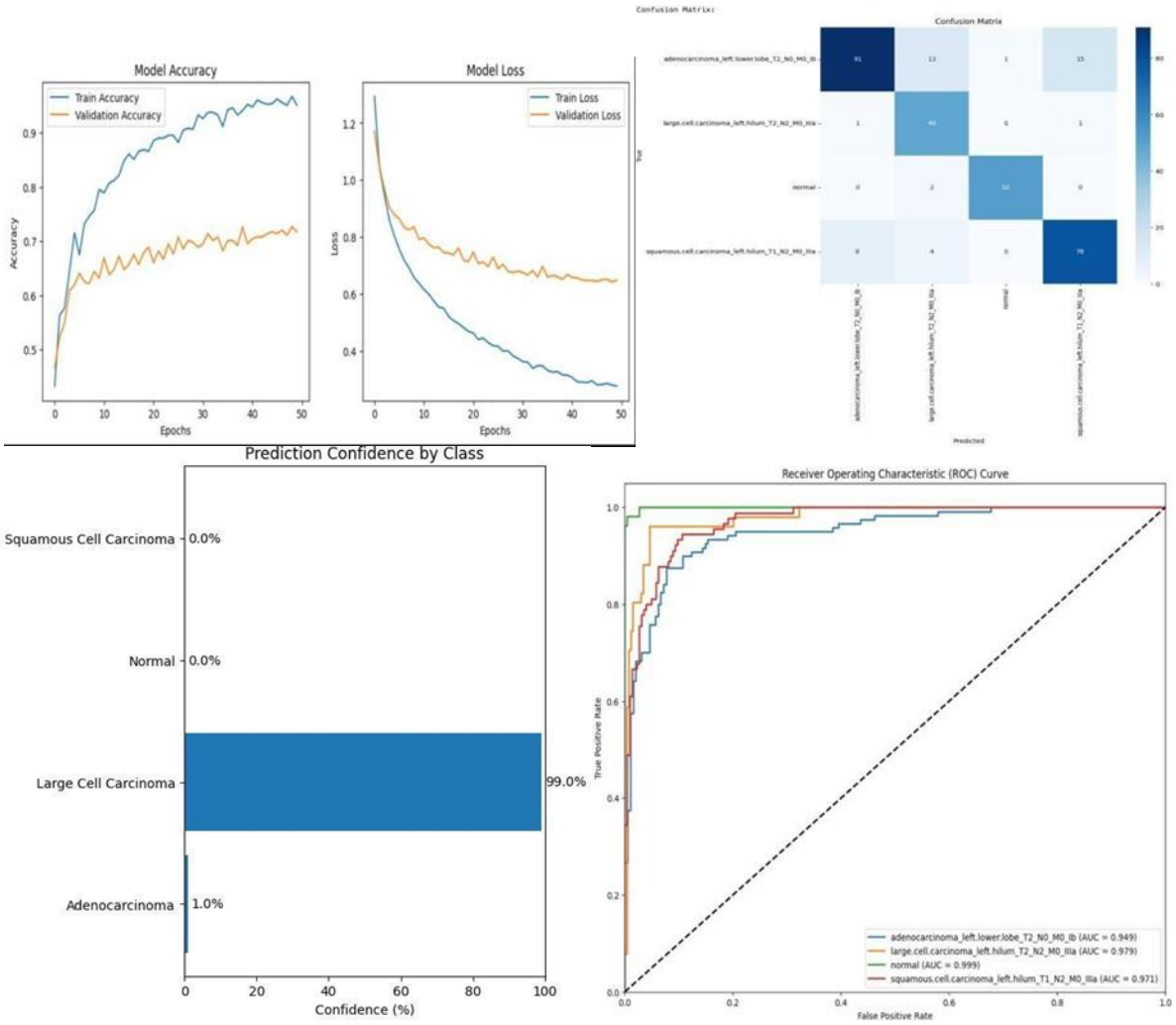





```
Final training accuracy = 0.975530207157135
Final validation accuracy = 0.888888955116272
```

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



REFRENCES

[1] Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. International Journal of Computer Science and Information Technologies, 4(1), 39–45. This study explores data mining classification techniques for lung cancer diagnosis, focusing on accuracy and reliability in prediction.

[2] Zhang, J., Song, Y., Hu, S., & Liu, X. (2018). Pulmonary Nodule Detection in Medical Images: A Survey. Biomedical Signal Processing and Control, 43, 138–147. This paper provides a comprehensive review of pulmonary nodule detection methods in medical imaging, assessing various machine learning and deep learning techniques.

[3] Palani, D., & Venkatalakshmi, K. (2019). An IoT-Based Predictive Modeling for Predicting Lung Cancer Using Fuzzy Cluster-Based Segmentation and Classification. Journal of Medical Systems, 43(2), 21. This research proposes an IoT-based predictive model integrating fuzzy clustering and classification algorithms to enhance early lung cancer detection.

[4] Lynch, C. M., et al. (2017). Prediction of Lung Cancer Patient Survival via Supervised Machine Learning Classification Techniques. International Journal of Medical Informatics, 108, 1–8. The study applies supervised machine learning to predict lung cancer survival rates using clinical and pathological data.

[5] Sumathipala, Y., et al. (2019). Machine Learning to Predict Lung Nodule Biopsy Method Using CT Image Features: A Pilot Study. Computerized Medical Imaging and Graphics, 71, 1–8. This pilot study explores how machine learning models can predict the most appropriate biopsy method for lung nodules based on CT imaging

features.
[6] Alzubaidi, A. K., Sideseq, F. B., Faeq, A., & Basil, M. (2017). Computer-Aided Diagnosis in Digital Pathology Application: Review and Perspective Approach in Lung Cancer Classification. Proceedings of the New Trends in Information & Communications Technology Applications, 219–224. IEEE, Baghdad, Iraq. This conference paper reviews computer-aided diagnostic tools for lung cancer classification in digital pathology applications.
[7] Wang, L. (2022). Deep Learning Techniques to Diagnose Lung Cancer. Cancers. This paper discusses the application of deep learning models, including CNNs and transfer learning, for automated lung cancer diagnosis.
[8] Shah, A. A., et al. (2023). Deep Learning Ensemble 2D CNN Approach. Scientific Reports. This research presents an ensemble approach using 2D CNN models for improving the accuracy of lung cancer detection in medical imaging.
[9] Liu, K., et al. (2022). YOLO-Based Nodule Detection. IEEE Access. The study explores the use of YOLO (You Only Look Once) deep learning architecture for real-time lung nodule detection in CT images.
[10] Alom, M. Z., et al. (2019). Recurrent Residual U-Net. Journal of Medical Imaging (JMI). This paper introduces the Recurrent Residual U-Net, an advanced deep learning model for segmenting lung nodules in CT scans.
[11] Lakshmi, B. S., et al. (2024). Deep Learning in Lung Cancer Detection. Juni Khyat. This paper presents various deep learning techniques applied in lung cancer detection, emphasizing feature extraction and model optimization.
[12] Setio, A. A. A., et al. (2017). LUNA16 Challenge for Automated Pulmonary Nodule Detection. Medical Image Analysis. This study introduces the LUNA16 dataset and discusses the performance of different automated lung nodule detection models.
[13] Reddy, U. J., et al. (2023). Recognition of Lung Cancer Using Machine Learning Mechanisms. International Journal. This research applies multiple machine learning models to recognize lung cancer patterns from medical imaging data.
[14] Shuvo, S. B. (2024). An Automated End-to-End Deep Learning Framework for Lung Cancer Diagnosis. Transactions on Biomedical Engineering. This study develops a deep learning framework automating the entire lung cancer diagnosis pipeline.
[15] Wu, R., et al. (2023). Self-Supervised Transfer Learning for Benign-Malignant Nodule Classification. Expert Systems with Applications. The research employs self-supervised learning techniques to improve benign-malignant nodule classification in lung cancer detection.
[16] Lakshmanaprabu, S. K., et al. (2020). Optimized Deep Learning Model for Lung Cancer Detection. Future Generation Computer Systems. This paper discusses optimization techniques in deep learning models for accurate and efficient lung cancer detection.
[17] Luo, X., et al. (2022). 3D Sphere Representation-Based Detection Network. LUNA16 Analysis. The study presents a novel 3D sphere representation-based detection model for improved lung cancer nodule classification.
[18] Collins, L. G., et al. (Year). Lung Cancer: Diagnosis and Management. This book provides a detailed overview of lung cancer diagnosis, treatment, and patient management strategies.
[19] Mamun, M. (Year). The Efficiency of Machine Learning Models in Lung Cancer Risk Prediction. This study evaluates various machine learning models and their efficiency in predicting lung cancer risk.
[20] Sadeghi Pour, E. (Year). Lung Cancer Detection Using CT Scan Images Based on Genetic Independent Recurrent Deep Learning Models. This paper presents a novel genetic-independent recurrent deep learning model to enhance the detection of lung cancer in CT scan images.