The Scientific Temper (2025) Vol. 16 (4): 4089-4094
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.4.11
E-ISSN: 2231-6396, ISSN: 0976-8653
https://scientifictemper.com/

RESEARCH ARTICLE

Role of big data in enhancing lung cancer prediction and treatment

Ritu Nagila*, Abhishek Kumar Mishra, Ashish Nagila

Abstract

Lung cancer functions as one of the world's primary diseases, which causes death because patients receive their diagnoses too late and have limited choices regarding treatment. The research examines big data functionality for lung cancer prediction through extensive evaluation of patient healthcare records and image data along with hereditary information to establish predictive models. Evaluation of deep learning and ensemble techniques, as well as additional machine learning algorithms, takes place to measure their accuracy rates and operational efficiency.

Keywords: Lung Cancer, Big Data Analytics, Machine Learning, Early Detection, Predictive Modeling, Deep Learning.

Introduction

The worldwide occurrence of lung cancer stands as a leading cause of cancer fatalities since this form of cancer kills a major portion of patients who develop cancer. Lung cancer stands as the number one killer among cancer-related diseases based on World Health Organization (WHO) figures because patients only survive fewer than 20% when their symptoms emerge late (AlOsaimi *et al*., 2025). Survival rates increase substantially when medical intervention occurs early since it delivers better results to patients. The current diagnostic procedures, including biopsies and CT, MRI, and X-rays aside from molecular testing, prove to be both expert specialist-dependent and, time-consuming and costly (Marathe & Bhalekar, 2022). Doctor visits for diagnosis occur too late when the disease advances beyond the initial stages, which decreases the available treatment possibilities (Huang, Yang *et al*., 2023).

Recently integrated big data analytics and artificial intelligence enable medical research to establish more effective ways to detect lung cancer's first signs and predict its occurrence (Duranti *et al*., 2025). Through big data analytics institutions succeed in analyzing diverse medical data collections for improved fast diagnosis results(Parikh *et al*., 2019).

The literature presents multiple research which investigate the implementation of AI-based methodologies to diagnose lung cancer through studies on radiomics analysis and, computational pathology and predictive modeling (Mascalchi *et al*., 2025). Massive success emerges from the application of convolutional neural networks (CNNs) in deep learning models since they analyze lung images to produce cancer detection outcomes more precisely than classic approaches (Neal Joshua *et al*., 2021). Predictive analytics examines organized patient information about demographics together with smoking patterns and genetic indicators to detect people who will likely develop lung cancer (Parikh *et al*., 2019). The obstacles to progress include ensuring data privacy, making models more understandable and dealing with the integration of various information sources to generate complete prognostic insights (Nasrullah *et al*., 2019).

The research investigates big data analytics methods that solve these difficulties for the prediction of lung cancer. The potential of ML and DL algorithms for analyzing different medical datasets enables clinicians to get useful early diagnosis information from their data (Rabby *et al*., 2025). The study presents the benefits of uniting clinical data with genomic results and imaging scans for better prediction outcomes. The research performs a detailed computational evaluation that reviews predictive methods used in medical applications while examining both the positive and negative

Department of Computer Science and Engineering, IFTM University Moradabad,India.

***Corresponding Author:** Ritu Nagila, Department of Computer Science and Engineering, IFTM University Moradabad,India., E-Mail: ritu.upadhayay01@gmail.com

How to cite this article: Nagila, R., Mishra, A.K., Nagila, A. (2025). Role of big data in enhancing lung cancer prediction and treatment. The Scientific Temper, 16(4):4089-4094.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.4.11

Source of support: Nil

Conflict of interest: None.

Received: 16/03/2025 **Accepted:** 9/04/2025 **Published** : 25/04/2025

elements of big data processing models for cancer diagnosis (Parikh *et al*., 2019).

Medical staff should incorporate AI-based predictive models into their work processes for proactive healthcare interventions that lead to better patient results.

Novelty and Contribution

The analysis of this research departs from standard examination methods by using sophisticated algorithms to process multiple data sources, which generate a thorough understanding of lung cancer probability assessment. The main research findings consist of the following points:

A. Multi-Source Data Fusion

Predictive accuracy receives an enhancement through the integration of both structured clinical and demographic data and unstructured medical imaging and genomic data in the study. The method utilizes a new feature selection process to establish which attributes hold the most connection to lung cancer development progression.

B. Deep Learning-Based Image Analysis

CNNs serve as automated systems for analyzing lung images to decrease medical staff dependency on diagnostic reading.

C. Risk Stratification and Personalized Prediction

Machine learning technology creates a risk assessment model that divides patients into risk groups to enable prompt medical care. Through a patient-related input system, the model analyzes healthcare factors, including smoking behavior as well as inherited susceptibility to cancer and existing health issues for detailed lung cancer assessment.

D. Big Data Analytics for Enhanced Decision-Making

The study uses Hadoop and Spark big data processing frameworks to handle medical datasets on a large scale. Predictive analytics processing in real-time gives clinical personnel better tools to make prompt, informed medical choices about patient treatment.

E. Clinical Integration and Validation

A test of the proposed model takes place by applying it to genuine clinical data to prove its operational strength and data precision. The partnership between healthcare professionals helps establish practical applications that comply with current medical procedures.

The research presents significant advancement to AI-driven cancer diagnostics by focusing on these important aspects to establish precise and, time-sensitive and cost-efficient lung cancer prediction systems.

Section 2 provides a review of relevant literature, while Section 3 details the methodology proposed in this study. Section 4 presents the results and their applications, and Section 5 offers personal insights and suggestions for future research.

Related Works

There has been much exploration of lung cancer prediction and early detection in recent times through the analysis of machine learning together with deep learning and big data analytics. Medical imaging studies with CT scans and X-ray images to detect diseases early represent research areas that multiple investigations have examined regarding artificial intelligence involvement. CNNs serve the medical field through image classification and, segmentation, and anomaly detection while achieving high accuracy in distinguishing benign and malignant lung nodules. Predictive modeling receives enhanced capabilities because radiomics enables analysts to extract quantitative features from images that human eyes cannot detect.

A machine learning model was developed by (Gould *et al*., 2021) to predict non-small cell lung cancer (NSCLC) using clinical and lab data. It outperformed the mPLCOm2012 model in detecting cancer 9–12 months early. The model achieved an AUC of 0.86 and a sensitivity of 40.1% at 95% specificity. In comparison, mPLCOm2012 had an AUC of 0.79 and 27.9% sensitivity. It also surpassed standard screening criteria. Key features included blood cell and platelet counts. This approach improves early detection and could help reduce lung cancer deaths.

An ensemble Federated Learning-based method was proposed by (Subash Chandra Bose *et al*., 2023), enabling decentralized training across multiple datasets while maintaining data privacy. This approach achieved 89.63% accuracy on the Kaggle lung cancer dataset, showing improved performance over conventional models and highlighting the benefits of combining ensemble techniques with Federated Learning.

A hybrid deep learning model combining convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks to analyze patients' medical notes for accurate cancer prediction was developed by (Kesiku & Garcia-Zapirain, 2024). This model highlighted the potential of artificial intelligence in enhancing early lung cancer detection. Evaluated on the MIMIC IV dataset, the model achieved an accuracy of 98.1% and an MCC of 96.2%, outperforming traditional models like LSTM and BioBERT. Further comparisons on the Yelp Review Polarity dataset confirmed its superior performance. This approach demonstrates the effectiveness of deep learning in precision medicine and emphasizes the ongoing need for model refinement in clinical applications.

An Edge AI-based system called EASPLD for diagnosing both pneumonia and lung cancer was introduced by (Ponnada & Srinivasu, 2019). The system serves as a clinical decision support tool, utilizing a custom deep learning architecture (EASPLD-CNN) with seven convolution layers and a unique combination of 3×3 and 5×5 kernels—unlike other models whichypically use only one kernel size. It

processes lung X-ray and CT images sourced from the LIDC-IDRI and Mendeley datasets. EASPLD integrates modules for image input, enhancement, analysis, and result reporting, delivering outputs to clinicians or patients through visual, textual, and email notifications. This system highlights the growing role of deep learning in real-time, multi-disease diagnostic tools.

Proposed Methodology

The prediction system using big data analytics for lung cancer requires multiple operations, such as data acquisition and cleaning, followed by feature identification and model development and testing processes. Implementation of machine learning (ML) alongside deep learning (DL) with big data frameworks leads to precise cancer risk prediction through efficient evaluation processes (Huang, Arpaci *et al*., 2023). The system handles diverse medical data from clinical records and, imaging data and genetic information to deliver a complete lung cancer prediction model (Germer *et al*., 2024).

1. Data Collection and Preprocessing

Initiating the process requires gathering multiple medical data types that derive from medical facilities public health databases, and genomic laboratories. Heterogeneous medical datasets include organized demographic information with patient medical records alongside unorganized clinical images and pathologic materials. The predictive model requires high-quality medical data inputs, so preprocessing techniques such as data imputation normalization and augmentation address the missing values and inconsistencies found in medical data.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where X is the original feature value, $X_{\min}$ and $X_{\max}$ are the minimum and maximum values in the dataset, and X' is the normalized feature value.

2. Feature Extraction and Selection

The predictive model determines which data attributes hold the greatest value for diagnosing lung cancer. Imaging data undergoes radiomics feature extraction to obtain texture, shape and intensity characteristics, while clinical data uses Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) as machine learning-based or statistical feature selection approaches.

A relationship exists between the PCA transformation where:

$$Z = XW$$

Where X is the original data matrix, W is the eigenvector matrix, and Z represents the transformed feature space with reduced dimensions.

Model Training Collaborates with Deep Learning to Achieve the Aim

A predictive model uses Convolutional Neural Networks as its base deep learning technique for analyzing images, while it uses Random Forest or XGBoost algorithms for structured data classification. The CNN portion in the system extracts spatial patterns from lung images and machine learning methods evaluate patient information to produce better predictive models. The CNN model contains convolutional along with pooling and fully connected layers to identify hierarchical characteristics in medical images.

CNN operates through the convolution operation, which appears as follows:

$$F(i,j) = \sum_m \sum_n I(i-m, j-n) K(m,n)$$

Where $F(i,j)$ is the feature map, $I(i,j)$ is the input image, and $K(m,n)$ is the convolution kernel applied over the input data.

4. Model Evaluation and Validation

The metrics used for trained model evaluation include accuracy, sensitivity, specificity and the F1-score. The k-fold cross-validation process and other cross-validation techniques are used to stop overfitting and achieve generalization across various patient populations.

The model accuracy calculation method consists of the following:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP and TN. represent true positives and true negatives, while FP and FN denote false positives and false negatives, respectively.

5. Deployment and Real-Time Prediction

A big cloud-based data platform receives a trained model for real-time predictions to support lung cancer screening operations (Kasinathan & Jayakumar, 2022). Educational learning approaches enable secure data sharing between hospitals since they process aggregate information from multiple sites while safeguarding individual patient information (Bhatia *et al*., 2024). The proposed methodology is given in Figure 1.

Results and Discussions

The proposed big data model for lung cancer prediction underwent testing by using a variety of data consisting of clinical standards with medical images and genomic sequences. Before machine learning and deep learning models were applied, the dataset received preprocessing treatment to remove inconsistencies and normalize feature values (Kesiku & Garcia-Zapirain, 2024b). The predictive models achieved assessment through the calculation of four

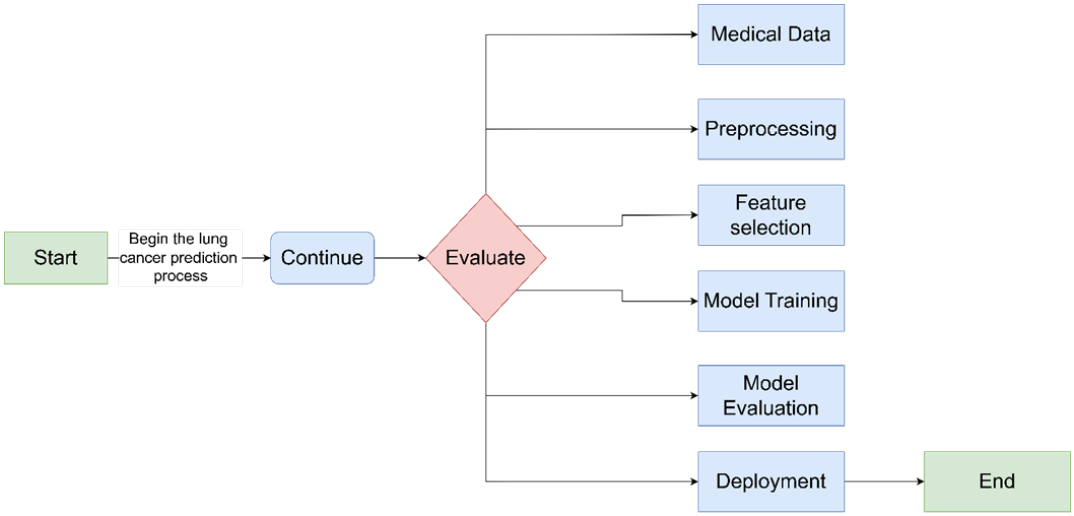

Figure 1: Proposed Methodology of Lung Cancer Prediction Using Big Data Analytics

evaluation metrics: accuracy, sensitivity and specificity and F1-score. CNNs turned out to provide superior performance than typical machine learning methods when used for image-based lung cancer predictions (Heidari *et al.*, 2023).

The different predictive models receive evaluation through various metrics, as shown in Table 1. The accuracy rate of 94.5% from CNN models surpassed what Random Forest alongside support vector machine (SVM) models could achieve. The detection power of lung cancer lesions through medical images increased noticeably using models based on CNN technology because of their high sensitivity and specificity levels.

The three models present their receiver operating characteristic (ROC) curves in Figure 2. The CNN model achieved the best performance result through its highest Area under the curve (AUC) metric. A ROC curve analysis exposes deep learning models as better than conventional machine learning methods in their ability to determine between malignant and benign lung nodules with elevated confidence rates.

The measurement of training time alongside inference time allowed researchers to review computational efficiency between the tested models. Table 2 exhibited a comparison of training and inferential time requirements. These networking models need more training time because of their intricate design yet they deliver faster inference testing suitable for essential clinical decision systems.

The predictive models presented their confusion matrix data in Figure 3. Clinical diagnosis depends on the

Table 1: Model performance comparison based on accuracy and sensitivity

Model	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*F1-Score (%)*
Random Forest	85.2	80.3	82.1	81.2
Support Vector Machine (SVM)	87.8	83.5	85.7	84.5
CNN (Deep Learning)	94.5	91.2	92.8	93.0

Table 2: Computational performance comparison of predictive models

Model	*Training Time (minutes)*	*Inference Time (milliseconds)*
Random Forest	12	45
Support Vector Machine (SVM)	18	30
CNN (Deep Learning)	120	12

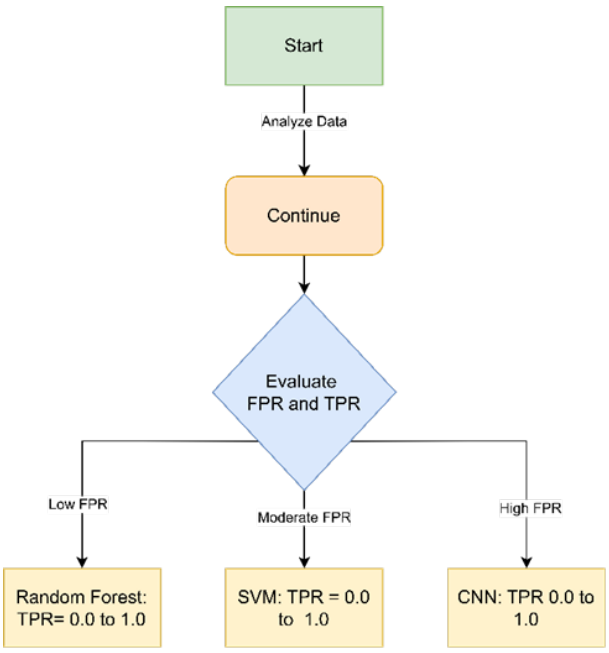

Figure 2: ROC curve data for predictive models

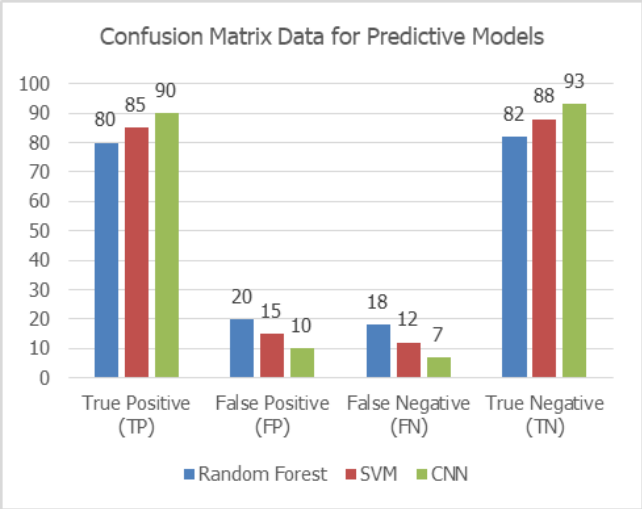

Figure 3: Confusion matrix data for predictive models

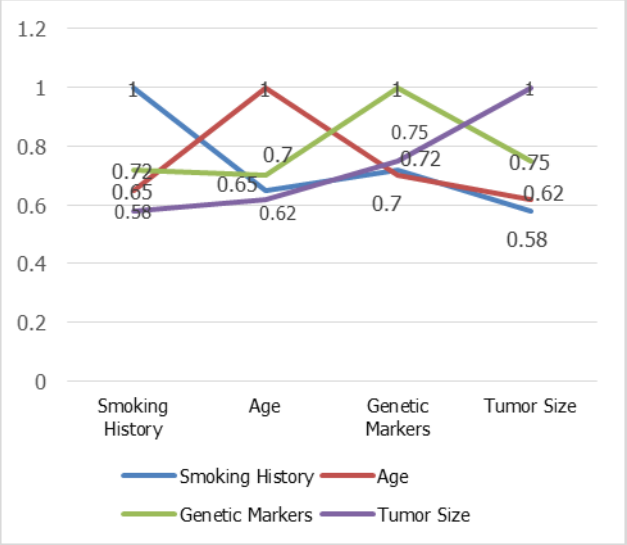

Figure 4: Correlation matrix of key clinical features

CNN model because it creates the lowest number of false positives and false negatives and thus offers enhanced reliability. The prediction accuracy of deep learning methods produces accurate lung cancer predictions by decreasing the number of incorrect diagnoses, which are verified through confusion matrix analysis.

As shown in Figure 4, tumor size exhibits a strong correlation with both genetic markers and smoking history, indicating their potential combined influence on cancer progression. Figure 4 contains the principal component analysis results displaying key clinical feature relationships that affect lung cancer risk. The most important elements which predict lung cancer emergence include smoking background alongside chronological age and genetic risk factors in combination with tumor measurements according to the correlation matrix results. The gained insights help dismantle the elements which affect lung cancer formation while enhancing custom-made risk measurement tools.

The study proves that big data analytics improve lung cancer prediction when implementing deep learning algorithms. A predictive diagnosis system becomes more effective through the incorporation of clinical data with imaging reports and genomic information. Additional investigation must be done to overcome issues related to data privacy as well as model interpretability and computational complexity. The research community must invest effort to develop both explainable AI methods together with federated learning approaches to achieve better clinical uptake for practical implementation.

Conclusion

Lung cancer prediction achieves its most potential through Big data analytics because it combines large datasets with state-of-the-art machine learning approaches. Predictive models display their capacity to detect lung cancer early, which results in better patient outcomes and decreased mortality statistics. The resolution of three primary challenges involving data protection, model transparency and combining diverse information sources needs immediate action. Future studies must develop AI algorithms for optimal performance while upholding ethics in data usage and, at the same time, introduce AI-supported clinical decision systems in healthcare practice.

Acknowledgments

I sincerely thank my supervisor for their valuable guidance and support throughout this research. I am also grateful to my institution, colleagues, and the research community for their contributions. Special thanks to my family and friends for their constant encouragement.

Conflict of interest

The author(s) declares no conflict-of-interest

References

AlOsaimi, H. M., Alshilash, A. M., Al-Saif, L. K., Bosbait, J. M., Albeladi, R. S., Almutairi, D. R., Alhazzaa, A. A., Alluqmani, T. A., Al Qahtani, S. M., Almohammadi, S. A., Alamri, R. A., Alkurdi, A. A., Aljohani, W. K., Alraddadi, R. H., & Alshammari, M. K. (2025). AI models for the identification of prognostic and predictive biomarkers in lung cancer: A systematic review and meta-analysis. *Frontiers in Oncology, 15*, 1424647. https://doi.org/10.3389/fonc.2025.1424647

Bhatia, I., Aarti, Ansarullah, S. I., Amin, F., & Alabrah, A. (2024). Lightweight advanced deep neural network (DNN) model for early-stage lung cancer detection. *Diagnostics (Basel, Switzerland), 14*(21), Article 12356. https://doi.org/10.3390/diagnostics14212356

Duranti, L., Tavecchio, L., Rolli, L., & Solli, P. (2025). New perspectives on lung cancer screening and artificial intelligence. *Life, 15*(3), 498. https://doi.org/10.3390/life15030498

Germer, S., Rudolph, C., Labohm, L., Katalinic, A., Rath, N., Rausch, K., Holleczek, B., & Handels, H. (2024). Survival analysis for lung cancer patients: A comparison of Cox regression and machine learning models. *International Journal of Medical Informatics, 191*, 105607. https://doi.org/10.1016/j.ijmedinf.2024.105607

Gould, M. K., Huang, B. Z., Tammemagi, M. C., Kinar, Y., & Shiff, R. (2021). Machine learning for early lung cancer identification using routine clinical and laboratory data. *American Journal of Respiratory and Critical Care Medicine, 204*(4), 445–453. https://doi.org/10.1164/rccm.202007-2791OC

Heidari, A., Javaheri, D., Toumaj, S., Navimipour, N. J., Rezaei, M., & Unal, M. (2023). A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems. *Artificial Intelligence in Medicine, 141*, 102572. https://doi.org/10.1016/j.artmed.2023.102572

Huang, S., Arpaci, I., Al-Emran, M., Kılıçarslan, S., & Al-Sharafi, M. A. (2023). A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability. *Multimedia Tools and Applications, 82*(22), 34183–34198. https://doi.org/10.1007/s11042-023-16349-y

Huang, S., Yang, J., Shen, N., Xu, Q., & Zhao, Q. (2023). Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. *Seminars in Cancer Biology, 89*, 30–37. https://doi.org/10.1016/j.semcancer.2023.01.006

Kasinathan, G., & Jayakumar, S. (2022). Cloud-based lung tumor detection and stage classification using deep learning techniques. *BioMed Research International, 2022*(1), Article 4185835. https://doi.org/10.1155/2022/4185835

Kesiku, C. Y., & Garcia-Zapirain, B. (2024a). AI-enhanced lung cancer prediction: A hybrid model's precision triumph. *IEEE Journal of Biomedical and Health Informatics*, 1–14. https://doi.org/10.1109/JBHI.2024.3447583

Kesiku, C. Y., & Garcia-Zapirain, B. (2024b). AI-enhanced lung cancer prediction: A hybrid model's precision triumph. *IEEE Journal of Biomedical and Health Informatics*, 1–14. https://doi.org/10.1109/JBHI.2024.3447583

Marathe, M., & Bhalekar, M. (2022). Detection of lung cancer using CT scans with deep learning approach. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1026–1031). IEEE. https://doi.org/10.1109/ICCES54183.2022.9835901

Mascalchi, M., Marzi, C., & Diciotti, S. (2025). Artificial intelligence propels lung cancer screening: Innovations and the challenges of explainability and reproducibility. *Signal Transduction and Targeted Therapy, 10*(1), Article 18. https://doi.org/10.1038/s41392-024-02111-9

Nasrullah, N., Sang, J., Alam, M. S., Mateen, M., Cai, B., & Hu, H. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors (Basel, Switzerland), 19*(17), 3722. https://doi.org/10.3390/s19173722

Neal Joshua, E. S., Bhattacharyya, D., Chakkravarthy, M., & Byun, Y.-C. (2021). 3D CNN with visual insights for early detection of lung cancer using gradient-weighted class activation. *Journal of Healthcare Engineering, 2021*, Article 6695518. https://doi.org/10.1155/2021/6695518

Parikh, R. B., Gdowski, A., Patt, D. A., Hertler, A., Mermel, C., & Bekelman, J. E. (2019). Using big data and predictive analytics to determine patient risk in oncology. *American Society of Clinical Oncology Educational Book, 39*, e53–e58. https://doi.org/10.1200/EDBK_238891

Ponnada, V. T., & Srinivasu, S. V. N. (2019). Edge AI system for pneumonia and lung cancer detection. *International Journal of Innovative Technology and Exploring Engineering, 8*(9), 1908–1915. https://doi.org/10.35940/ijitee.I8584.078919

Rabby, M. S., Islam, M. M., Kumar, S., Maniruzzaman, M., Hasan, M. A. M., Tomioka, Y., & Shin, J. (2025). Identification of potential biomarkers for lung cancer using integrated bioinformatics and machine learning approaches. *PLOS ONE, 20*(2), e0317296. https://doi.org/10.1371/journal.pone.0317296

Subashchandrabose, U., John, R., Anbazhagu, U. V., Venkatesan, V. K., & Thyluru Ramakrishna, M. (2023). Ensemble federated learning approach for diagnostics of multi-order lung cancer. *Diagnostics, 13*(19), 3053. https://doi.org/10.3390/diagnostics13193053