Jayendra S. Jadhav [1]

Jyoti Deshmukh

Research paper

DOI – 10.24874/QF.25.053

REVOLUTIONIZING EARLY LUNG CANCER DETECTION WITH MACHINE LEARNING: INSIGHTS FROM FEDERATED AND ENSEMBLE LEARNING

Abstract: *Timely detection of lung cancer leads to better patient results and helps doctors treat the condition more quickly. This research study combines Machine Learning, Federated Learning, and Ensemble Learning to create an improved way to identify lung cancer in its initial stages. Using machine learning for pattern detection, federated learning to protect data privacy, and explainable learning to improve diagnosis, this model brings a strong and flexible solution to healthcare. The proposed ensemble model combines multiple algorithms to improve prediction reliability and robustness against noisy and imbalanced datasets. Federated Learning ensures secure, decentralized data processing across institutions, maintaining patient confidentiality while enabling collaborative insights. Advanced preprocessing techniques and feature engineering optimize the dataset for meaningful analysis, while the ensemble voting mechanism ensures consistent and accurate predictions. Early detection and catching all cases through this approach has improved both how accurately doctors find cancer and how many cases they can identify. Our next research steps include adding more information sources alongside better prediction understanding features before implementing these tools in real hospital operations. By addressing both technical and ethical concerns, this research offers a solid basis for the creation of expandable AI-powered diagnostic instruments for lung cancer treatment*

Keywords: *ensemble learning, machine learning, federated learning*

1. Introduction

To increase survival chances, lung cancer must be detected early. Early illness detection enables prompt intervention, which lowers the risk of metastasis and enhances treatment results. This proactive approach also lessens the physical, emotional, and economic burdens on both patients and healthcare systems (Sivanagireddy et al., 2022).

In medical diagnostics, machine learning (ML) has become a game-changing instrument. ML systems examine data collections to detect unusual signs and patterns that normal medical tests struggle to catch. For lung cancer detection, ML processes imaging data, genetic profiles, and patient histories, enabling earlier and more accurate predictions compared to conventional techniques. This fosters faster decision-making and more tailored treatment strategies (Nimmagadda et al., 2024; Sivanagireddy et al., 2022).

Federated learning (FL) addresses privacy concerns in collaborative medical research by enabling decentralized data training. FL fa-

1 Corresponding author: Jayendra S. Jadhav
Email: jayendra071985@gmail.com

cilitates collective insights without requiring data sharing, ensuring patient confidentiality and adherence to data protection regulations. This approach is especially beneficial in healthcare, where multi-institutional collaborations are vital (Ruprah et al., 2024).

Ensemble learning (EL) enhances diagnostic accuracy by integrating multiple algorithms into a unified predictive model (Jadhav & Deshmukh, 2025). Techniques such as bagging, boosting, and stacking reduce overfitting and improve robustness. By combining EL with ML and FL, we create scalable, secure, and highly accurate systems for early lung cancer detection (Jehangir et al., 2022; Mamun et al., 2022).

The synergy of ML, FL, and EL offers transformative potential for lung cancer diagnostics. Despite challenges like data heterogeneity and computational demands, addressing these issues can revolutionize cancer detection and improve patient outcomes (Rao & Arshad, 2023).

2. LITERATURE SURVEY

This survey explores the evolving role of machine learning (ML) and federated learning (FL) in enhancing early detection strategies for lung cancer, focusing on their impact on diagnostic accuracy and data privacy.

2.1 Early identification of Lung Cancer

Identifying and treating lung cancer is pivotal for improving treatment outcomes and survival rates. Research indicates that early-stage lung cancer diagnosis can lead to significantly higher survival rates compared to late-stage discovery. Technologies and methodologies that facilitate early detection can therefore transform patient prognosis and reduce the healthcare burden.

- *Reference (Kukreja et al., 2022): S. Kukreja et al. utilize machine learning algorithms to analyse early detection methodologies, highlighting the potential of ML in identifying lung cancer from less obvious symptoms and routine screenings.*
- *Reference (Sivanagireddy et al., 2022): K. Sivanagireddy and colleagues discuss the utilization of correlation and regression analyses to predict lung cancer, emphasizing the enhancement of early detection rates through sophisticated data analysis techniques.*

Table 1. Importance of Early Detection

Reference Number	Paper Description	Analysis
(Kukreja et al., 2022)	Machine learning algorithms for lung cancer detection	• **Pros**: Early symptom identification; • **Cons**: Requires large datasets; • **Technology**: Machine Learning
(Sivanagireddy et al., 2022)	Predicting Early Lung Cancer with Correlation and Regression Techniques	• **Pros**: Accurate early detection; • **Cons**: Limited to available data; • **Technology**: Statistical Analysis

2.2 Role of Machine Learning

Machine learning offers transformative potential in diagnosing lung cancer by analyzing complex datasets to detect patterns indicative of early-stage cancer. ML models can analyse data from genetic profiles, patient histories, and even real-time symptom reports to offer predictions.

- *Reference (Sivanagireddy et al., 2022)*: Demonstrates how regression models can predict lung cancer from clinical data, showing ML's potential to leverage existing healthcare data for early warning signs.
- *Reference (Kesiku & Garcia-Zapirain, 2024):* C. Y. Kesiku and B. Garcia-Zapirain explore the AI-enhanced predictions which provide a hybrid model's precision in early cancer detection, highlighting the technological sophistication of ML in healthcare.

Table 2. Role of Machine Learning

Reference Number	Paper Description	Analysis
(Sivanagireddy et al., 2022)	Early Lung Cancer Prediction using Correlation and Regression	• **Pros**: High predictive power; • **Cons**: Dependent on quality of data inputs; • **Technology**: Regression Analysis
(Kesiku & Garcia-Zapirain, 2024)	AI-Enhanced Lung Cancer Prediction	• **Pros**: High precision; • **Cons**: Complexity in model training; • **Technology**: AI, Machine Learning

2.3 Federated Learning in Healthcare

Federated learning represents a significant advancement in respecting patient privacy while allowing for the collective training of ML models. This approach is especially pertinent in healthcare settings, where data sensitivity is paramount.

- *Reference (Kesiku & Garcia-Zapirain, 2024):* The work of Kesiku and Garcia-Zapirain demonstrates the application of FL in enhancing data privacy while allowing comprehensive data analysis across multiple institutions.
- *Reference (Jehangir et al., 2022)*: Researchers evaluate machine learning methods for lung cancer detection while demonstrating how federated learning maintains patient data privacy across networks.

Table 3. Federated Learning in Healthcare

Reference Number	Paper Description	Analysis
(Kesiku & Garcia-Zapirain, 2024)	AI-Enhanced Lung Cancer Prediction	• **Pros**: Enhances privacy; • **Cons**: Complex implementation; • **Technology**: Federated Learning
(Jehangir et al., 2022)	Early Detection of Lung Cancer Using ML Technique	• **Pros**: Data security; • **Cons**: Requires high computational resources; • **Technology**: Federated Learning

2.4 Role of Ensemble Learning

Ensemble Learning improves diagnostic robustness by combining the strengths of multiple models, addressing issues like overfitting and bias:

- *Reference (Jehangir et al., 2022)*: Jehangir et al. demonstrate how ensemble ML models enhance performance using algorithms like Random Forest and Gradient Boosting.
- *Reference (Mamun et al., 2022)*: Mamun et al. highlight EL's accuracy and reliability in lung cancer predictions through systematic reviews.

Table 4. Ensemble Learning in Healthcare

Reference Number	Paper Description	Analysis
(Jehangir et al., 2022)	Lung Cancer Detection using Ensemble of ML Models	• **Pros:** High accuracy; • **Cons:** High setup cost; • **Technology:** Ensemble Learning
(Mamun et al., 2022)	Lung Cancer Prediction using Ensemble Techniques	• **Pros**: Improved robustness; • **Cons**: Computationally intensive; • **Technology**: Ensemble Learning

2.5 Integration of ML, FL and EL in Lung Cancer Detection

The synergy between ML, FL and EL is particularly potent for early lung cancer detection, combining the strengths of both to create robust, scalable, and secure diagnostic tools.

- *Reference (Mamun et al., 2022)*: Highlights how combining ML and EL ensures accurate predictions and model robustness.
- *Reference (Rao & Arshad, 2023):* Discusses how FL enhances security and privacy in ensemble-based models for lung cancer detection.

Table 5. Integration of ML and FL

Reference Number	Paper Description	Analysis
(Mamun et al., 2022)	Relies on deep learning algorithms to spot lung cancer when it first appears.	• **Pros**: Improved Accuracy; • **Cons**: Extensive Data Requirements; • **Technology**: DL,FL
(Rao & Arshad, 2023)	This paper shows how combining several machine learning methods can help spot lung cancer at an early stage.	• **Pros**: Scalable and secure; • **Cons**: High Computation cost; • **Technology**: ML, FL, EL

The literature emphasizes the critical importance of early detection in improving lung cancer survival rates. Machine learning (ML) and ensemble learning (EL) play key roles by analysing diverse datasets and combining multiple models to enhance detection accuracy and reduce biases. Federated learning protects patient data privacy as healthcare providers collaborate to create models using data from multiple medical facilities. The combined use of ML, EL, and FL will enable secure systems for accurate early diagnosis that can boost healthcare treatment effectiveness.

3. System Methodology

Figure 1 illustrates our approach using machine learning and ensemble learning to screen for lung cancer early. Below is a step-by-step explanation of how the methodology aligns with the insights from the article.

- *Input: Patient Symptoms:* The dataset includes patient-reported symptoms, such as coughing, chest pain, fatigue, and wheezing, along with risk factors like smoking and alcohol consumption. These features serve as the input for predicting lung cancer. The focus is on capturing relevant information that correlates strongly with the disease. In lung cancer detection, these inputs are vital, as symptoms often overlap with other respiratory conditions. By collecting a broad range of symptom data, the model builds a foundation for accurate analysis.
- *Data Pre-processing:* Before applying machine learning models, raw data needs to be processed to ensure it is clean, consistent, and usable. Pre-processing is crucial to prevent biases or inaccuracies in the results.
  - *Data cleaning:* Missing categorical values were imputed using *mode*, while numerical values (e.g., age, smoking index) used *mean imputation*. Outliers were identified and removed using *Interquartile Range (IQR)* filtering to ensure data consistency.
  - *Feature Scaling:* Applied *Min-Max Scaling* to normalize numerical variables within *[0, 1]*, preventing dominant features from skewing model learning. For example, age (25–80) was rescaled proportionally to ensure uniform influence.
  - *Handling Class Imbalance:* Since *cancerous cases outnumber non-cancerous ones, SMOTE (Synthetic Minority Oversampling Technique)* was used to gener-

ate synthetic minority class samples, improving model generalization and reducing bias.

— *Feature Encoding: One-Hot Encoding* was applied to categorical features (e.g., gender), while *binary encoding* (0 = Absent, 1 = Present) was used for symptoms (e.g., coughing, wheezing) to ensure efficient model interpretation.

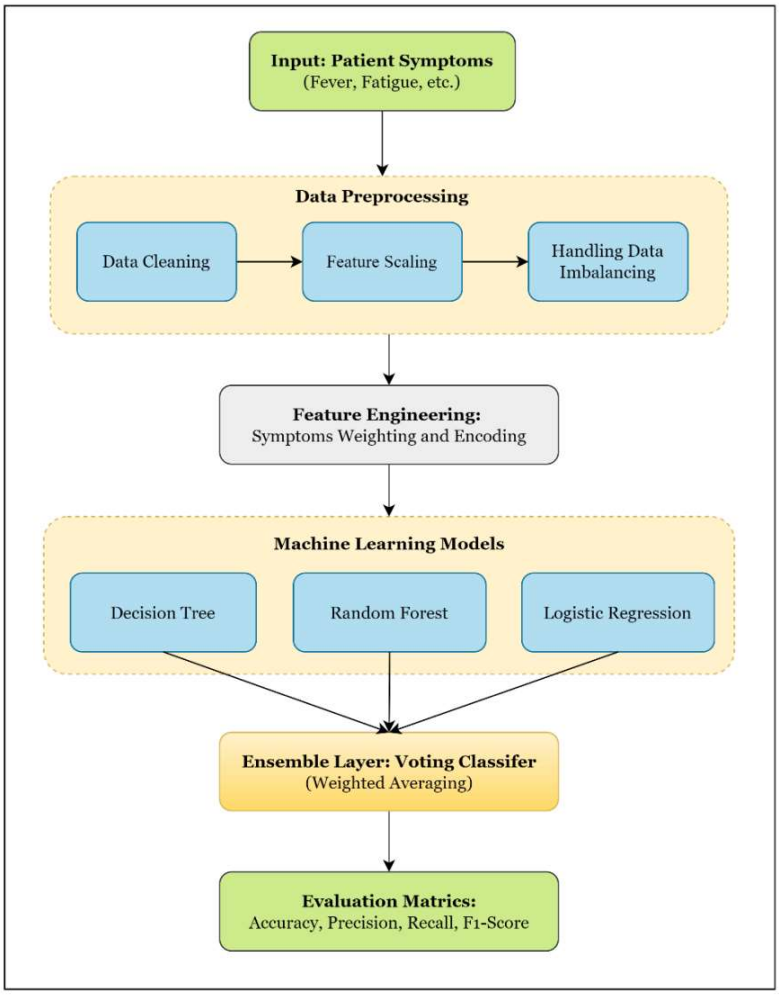

Figure 1. Methodology

- *Feature Engineering*: The methodology includes steps to enhance the dataset by refining the features, making it easier for the models to identify patterns. Symptom weighting assigns higher importance to critical features like "*Coughing of Blood*" and "Chest Pain", ensuring the model prioritizes key lung cancer indicators during training. Binary symptoms, such as "*Wheezing*", are numerically encoded as 0 (Absent) or 1 (Present) for seamless ML processing. These transformations enhance feature interpretability, decision boundary optimization, and overall predictive accuracy.
- *Machine Learning Models:* The research utilizes *ensemble learning* methods which integrates *Decision Tree (DT) with Random Forest (RF) and Logistic Regression (LR)* to boost lung cancer diagnostic preci-

sion. The following section provides technical details about each algorithm

- ***A Decision Tree***: The Decision Tree is used to create a tree-like structure where each internal node represents a decision based on a feature's value, and each leaf node represents a classification outcome (e.g., "Cancerous" or "Non-Cancerous").
  - *Splitting Criterion*: At each decision node, the decision tree *chooses the best feature* to split the data using the *Gini Impurity* metric:

    $$\boldsymbol{Gini(D) = 1 - \sum_{i=1}^{c} pi^2}$$

    Where pi represents the proportion of class i in dataset D.
  - To prevent overfitting, the tree depth is limited, and a pruning technique is applied to remove unnecessary branches. The Decision Tree assigns a class label to a given patient based on the majority class in the leaf node that the patient's features fall into. For instance, if a node has 80% cancerous cases and 20% non-cancerous, the classification at that node will be "Cancerous."
- ***Random Forest***: Random Forest is an ensemble method that builds multiple Decision Trees to make a more robust and generalized prediction. It aggregates the results from these trees to improve prediction accuracy and reduce overfitting.
  - *Bootstrap Aggregating (Bagging)*: RF trains each tree on a bootstrapped subset (random sampling with replacement), enhancing generalization and reducing variance.
  - *Feature Randomization*: Each tree considers a random subset of features at each split, ensuring diversity and reducing correlation.
  - *Voting Mechanism*: Final classification is determined through majority voting among all trees (e.g., if 7/10 predict "Cancerous," the outcome is "Cancerous").
  - *Final Prediction*: The ensemble aggregates individual tree predictions to classify cases as "Cancerous" or "Non-Cancerous" reliably
- ***Logistic Regression Model***: Logistic Regression calculates the probability that a given patient is "Cancerous" (class 1) or "Non-Cancerous" (class 0) based on a linear combination of input features, and applies the logistic function to map the result to a value between 0 and 1.
  - *Linear Model*: The model computes the weighted sum of the input features, $X_1,X_2,...,X_n$ (such as age, smoking history, coughing, etc.), using learned weights $b_1,b_2,...,b_n$, and an intercept b_o:
  - $Z = b_o+b_1X_1+b_2X_2+...+b_nX_n$
    Where X_1, X_2,…, X_n are the features (e.g., smoking index, coughing frequency, etc.)
  - *Sigmoid Function*: The output of the linear combination is passed through a sigmoid function to obtain a probability:

    $$P(Cancer = 1 \mid X) = \frac{1}{1+e-z}^{x}$$

  - This output represents the probability of the patient being classified as "Cancerous."
  - *Decision Threshold*: If $(P > 0.5)$, classify as cancerous. Else, classify as non-cancerous.

- *Ensemble Layer-Voting Classifier:* The outputs of the three models are combined using an ensemble method to improve accuracy and robustness. A weighted averaging technique is used, where models that perform better on validation data (e.g., Random Forest) are given greater influence in the final prediction. This layer improves prediction accuracy by capitalizing on each model's specific strengths.
- *Evaluation Metrics:* Standard indicators (as indicated in table 6) that guarantee the model's efficacy and dependability are used to assess its performance.

Table 6. Evaluation Metrics

Accuracy	We calculate overall prediction accuracy by finding the ratio between successful predictions and the total number of instances
Precision	The percentage of genuine positives among anticipated positives is determined by precision. By reducing false alarms, this statistic makes that the system is workable for clinical applications
Recall	Also referred to as sensitivity, and assesses the model's capacity to detect every instance of actual lung cancer. Since missing positive instances might result in delayed diagnosis, this is especially crucial for early detection.
F1-Score	Performs a complete analysis of diagnostic performance through its balanced combination of precision and recall metrics.

This approach combines machine learning, feature engineering, and thorough preprocessing methods to identify lung cancer. By emphasizing ensemble learning, the framework integrates the advantages of separate models to produce a trustworthy and precise diagnostic instrument. This systematic approach aligns with the article's emphasis on early disease detection through symptom-based analysis and advanced machine learning methods.

4. Model and Results Discussion

4.1 Model Overview

Through a voting classifier, the Ensemble Learning Model integrates the advantages of Random Forest, Gradient Boosting, and Logistic Regression. This model integrates predictions from multiple algorithms to improve diagnostic accuracy and robustness, particularly in complex and overlapping symptom scenarios (as shown in table 7).

Table 7. Ensemble Model Overview

Feature	**Ensemble Model**
Algorithm	LR + Random Forest + Gradient Boosting
Complexity	Moderate to High
Interpretability	Moderate
Key Focus	Ensemble Predictions

4.2 Efficiency of the Ensemble Model in Handling Real-World Challenges

The Ensemble Learning Model as shown in table 8 addresses common challenges in lung cancer detection:

- *Noise Handling* – Combines multiple algorithms to handle noisy and imbalanced datasets effectively.
- *Data Balance* – Naturally accommodates data imbalance through weighted contributions of individual models.
- *Scalability* – Performs well across diverse datasets, making it suitable for multi-institutional applications.

Table 8. Ensemble Model Efficiency

Feature	**Ensemble Model Efficiency**
Noise Handling	High
Data Balance	Well-Equipped
Scalability	Moderate to High

4.3 Model Classification Report

Table 9 shows the results of our Ensemble Learning Model to detect early lung cancer.

Table 9. Ensemble Learning Model Classification Details

	Precision	Recall	F1-Score	Support
Non-Cancerous	88%	81%	84%	7840
Cancerous	93%	96%	94%	15960
Model Accuracy			**91**%	

Insights from Quantitative Results:
The Ensemble Learning Model achieves a strong 91% overall accuracy, correctly classifying 91 out of 100 cases. It excels in detecting cancerous cases, achieving 93% precision and an impressive 96% recall, ensuring that almost all lung cancer cases are accurately identified. This high sensitivity significantly reduces the risk of missed diagnoses, which is critical for early intervention and improved patient outcomes.

However, the model faces challenges in distinguishing non-cancerous cases, with a recall of 81%, leading to 19% false positives. This equates to approximately 1,490 misclassified non-cancerous cases, potentially resulting in unnecessary diagnostic tests and patient anxiety. Despite this limitation, the model's high sensitivity for cancer detection reinforces its reliability in prioritizing early-stage diagnosis and improving clinical decision-making.

5. Deployment Challenges and Scalability Strategies

Deploying ensemble learning in clinical environments faces challenges related to computational demands, data variability, model interpretability, and regulatory compliance. Federated Learning (FL) requires high-performance GPUs/TPUs, limiting feasibility in resource-constrained settings. Model compression techniques (quantization, pruning) optimize efficiency for broader adoption. Data heterogeneity across hospitals affects generalization, which can be addressed through Domain Adaptation Techniques to standardize patient data distributions. Clinicians require transparent AI decision-making, making SHAP-based explain-ability crucial for trust and regulatory approval. Privacy laws (HIPAA, GDPR) restrict data sharing, but Secure Multi-Party Computation (SMPC) within FL ensures encrypted, privacy-preserving AI model training. Addressing these challenges through efficient computation, data standardization, explain-ability, and privacy measures enables scalable, real-world AI adoption in healthcare.

6. Clinical Validation Strategy

The clinical validation strategy *(as shown in figure 2)* will be implemented through a phased approach to ensure the model's accuracy, reliability, regulatory compliance, and seamless clinical integration. Initially, in Phase 1 (Retrospective Validation), the model will be trained and tested on historical hospital datasets, with predictions compared against expert radiologist diagnoses to assess baseline accuracy. Upon achieving satisfactory performance, Phase 2 (Prospective Testing) will involve deploying the model in real hospital settings, where live patient data will be analyzed. This phase will evaluate model accuracy, false positive/negative rates, and clinician feedback, ensuring its effectiveness in dynamic medical environments.

Following successful testing, Phase 3 (Regulatory Approval & Clinical Trials) will commence, where FDA and medical board approvals will be pursued. The model will be subjected to multi-institutional clinical trials,

where its predictive performance, interpretability, and impact on medical decision-making will be rigorously evaluated. Finally, Phase 4 (Full-Scale Implementation) will integrate the model into Electronic Health Record (EHR) systems, enabling real-time diagnostic support, continuous learning from new patient data, and automated model updates. This structured and iterative approach ensures that the model is clinically validated, ethically sound, and ready for large-scale deployment in real-world healthcare settings.

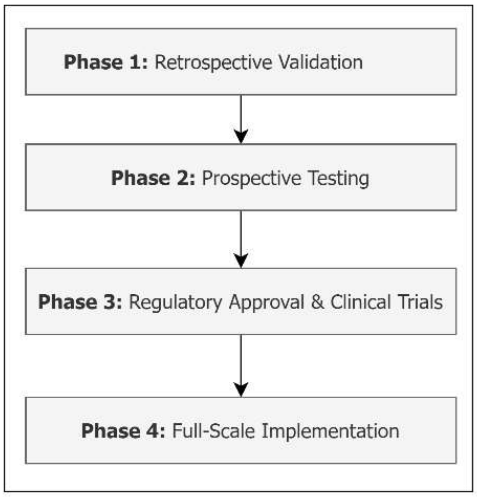

Figure 2. Clinical validation strategy

7. Conclusion and future work

Using ensemble learning models, researchers found that they could detect lung cancer with 91% accuracy while identifying 96% of all cancer patients in their testing data. The three algorithms - Logistic Regression, Random Forest, and Gradient Boosting - work together to provide stable performance when dealing with both inconsistent data and unbalanced datasets. Despite challenges like false positives and negatives, the model's ability to prioritize sensitivity for cancerous cases makes it a reliable tool for clinical diagnostics. This work underscores the importance of combining Machine Learning, Federated Learning, and Ensemble Learning to create secure, accurate, and scalable diagnostic solutions.

Future efforts should focus on integrating advanced data sources like imaging and genetic markers to enhance diagnostic accuracy. Developing hybrid models that combine Ensemble and Deep Learning could improve feature extraction and pattern recognition. Optimization for resource-constrained settings and testing on diverse datasets will enhance scalability and generalizability. Explainable AI techniques should be implemented to increase clinician trust and usability. Finally, real-world validation through clinical trials and collaborations with healthcare providers will ensure seamless adoption in clinical workflows, maximizing the model's impact on patient care.

References

Jadhav, J. S., & Deshmukh, J. (2025). Advancing Machine Learning in COVID-19 Diagnostics: Symptom-Based Classification and Ensemble Techniques. *South Eastern European Journal of Public Health*, 3044–3061. https://doi.org/10.70135/seejph.vi.3508

Jehangir, B., Nayak, S. R., & Shandilya, S. (2022, January). Lung cancer detection using ensemble of machine learning models. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 411-415). IEEE. doi: 10.1109/Confluence52989.2022.9734212.

Kesiku, C. Y., & Garcia-Zapirain, B. (2024). AI-Enhanced Lung Cancer Prediction: A Hybrid Model's Precision Triumph. *IEEE Journal of Biomedical and Health Informatics*. doi: 10.1109/JBHI.2024.3447583.

Kukreja, S., Sabharwal, M., & Gill, D. S. (2022, December). A Survey of Machine learning algorithms for Lung cancer detection. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 338-342). IEEE. doi: 10.1109/ICAC3N56670.2022.10074272.

Mamun, M., Farjana, A., Al Mamun, M., & Ahammed, M. S. (2022, June). Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. In *2022 IEEE World AI IoT Congress (AIIoT)* (pp. 187-193). IEEE. doi: 10.1109/AIIoT54504.2022.9817326

Nimmagadda, S. M., Likhitha, K., Srilatha, G., & Sree, S. M. (2024, April). Lung Cancer Prediction and Classification Using Machine Learning Algorithms. In *2024 International Conference on Expert Clouds and Applications (ICOECA)* (pp. 1012-1015). IEEE. doi: 10.1109/ICOECA62351.2024.00176.

Rao, B. D., & Arshad, M. (2023, January). Early detection of lung cancer using machine learning technique. In *2023 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-5). IEEE. doi: 10.1109/ICCCI56745.2023.10128389.

Ruprah, T. S., Regmi, B., Jadhav, S. B., & Singh, S. (2024, April). Early Stage Lung Cancer Detection Using Deep Learning. In *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)* (pp. 1-6). IEEE. doi: 10.1109/MITADTSoCiCon60330.2024.10575345.

Sivanagireddy, K., Yerram, S., Kowsalya, S. S. N., Sivasankari, S. S., Surendiran, J., & Vidhya, R. G. (2022, December). Early lung cancer prediction using correlation and regression. In *2022 International Conference on Computer, Power and Communications (ICCPC)* (pp. 24-28). IEEE. doi: 10.1109/ICCPC55978.2022.10072059.

Jayendra S. Jadhav
Department of Computer Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, Mumbai, India
jayendra071985@gmail.com
ORCID 0000-0001-6767-6580

Jyoti Deshmukh
2Department of Computer Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, Mumbai, India
Jyoti.Deshmukh@mctrgit.ac.in
ORCID 0000-0002-6671-3041