



LUNG CANCER DETECTION USING MACHINE LEARNING-A DATA DRIVEN APPROACH

¹Usha Naik, ²Suresha D, ³H N Prakash

¹PG Student, ²Professor, ³Professor

¹²³Department of CSE,

¹²Srinivas Institute of Technology ,Mangaluru,Karnataka, ³Rajeev Institute of Technology ,Hassan,Karnataka India.

Abstract: Among the leading causes of fatalities from cancer is still lung cancer deaths globally, with its initial phases often presenting without noticeable symptoms, making timely diagnosis challenging. Traditional detection methods are time-consuming, require expert interpretation, and are prone to diagnostic delays. Here, we propose a machine learning-based approach in order to use structured patient data to identify lung cancer early, including symptoms, lifestyle factors, and demographic information. The model leverages renowned for its resilience, the Random Forest algorithm and high classification performance in medical diagnostics. By automating the the recognition of high-risk individuals, this method aims to enhance early diagnosis, reduce manual effort, and support clinical decision-making. Experimental findings show that the model attains a precision of 97%, demonstrating its potential as a reliable, cost-effective, and scalable tool for carcinoma of the lung prediction. The application of such intelligent systems can greatly enhance patient outcomes in the medical field and contribute to timely medical intervention.

I. INTRODUCTION

Cancer of the lung is among the most vital types of cancer globally, accounting for around 11.6% of all cancer cases and 18.4% of all cancer deaths in 2020. Cancer is significant global health issue, as the disease is often asymptomatic in the early stages and tends to progress rapidly. Lung cancer is most common form of cancer. The early identification is key to reducing lung cancer mortality rates. Machine learning algorithms have exhibited great potential in enhancing the precision method detecting lung cancer. Performance of methods such as support vector machines (SVMs), decision tree, multi-layer perceptron, neural networks, naïve Bayes, a random forest and other ensemble models and majority voting were utilized in connection with performance comparison. Structured dataset containing patients' symptoms and related attributes for predicting lung cancer. Each row representing patient and columns represent attributes such as patient demographics, symptoms and lifestyle factors.

BASIC WORKING PIPELINE

The procedure to use machine learning to detect lung cancer are

- Gathering and preparing data
- Selecting algorithm for machine learning
- Model learning and appraisal
- Model installation

Gathering and preparing data: An approach to lung cancer diagnosis utilizing machine learning starts with data collection from diverse sources. The data comprises symptoms like persistent or worsening chest pain and coughing, difficulty breathing, and coughing up blood

Selecting an algorithm for machine learning: Various lung cancer machine learning methods identification include KNN, deep learning, random forests, and logistic regression. The choice of the technique depends on variables such as the type of data., size, complexity and specific problem. Each algorithm has unique strengths and weaknesses.

Model learning and appraisal: The model training process involves data preparation and algorithm selection. The chosen algorithm is instructed on the initial dataset, analyzing and grouping the training data for forecasting. Model performance is assessed utilizing testing information like recall, accuracy, precision and score.

Model installation: The trained model can detect lung cancer, make predictions on new data, and integrate into clinical workflows. Continuous performance assessment and retraining using fresh information guarantee, accuracy and relevance in lung cancer detection. How well lung cancer works depends on accurate input data, suitable algorithms and appropriate evaluation metrics. Maintaining model accuracy is essential when integrating new data.

II. LITERATURE SURVEY

Several studies have examined the possibility of machine learning methods for early detection of lung cancer, with a focus on both structured symptom data and image-based diagnostics. 1. S. Wang et al. (2020) investigated the application of various ML techniques on clinical datasets to predict lung cancer outcomes. According to their report, random forests and other ensemble models significantly outperformed traditional classifiers regarding precision and robustness because of their capacity to handle imbalanced and noisy data. 2. K. Srinivas and B. K. Rani (2021) proposed a model for predicting lung cancer using patient symptoms and demographic data. Their study emphasized the importance of preprocessing techniques such as normalization and feature encoding, which were critical for improving model efficiency and interpretability. 3. L. Zhang et al. (2019) utilized Random Forest classifiers to predict cancer status from structured datasets containing symptoms, lifestyle factors, and demographics. According to their findings, random forest models achieved a precision of up to 95%, highlighting their effectiveness in real-world clinical applications. 4. A. Sharma et al. (2022) implemented a comparative analysis between logistic regression, SVM, and Random Forest. According to their findings, random forest yielded the highest precision and reduced overfitting, making it well-suited for early detection of diseases like lung cancer. 5. M. Patel and T. Roy (2023) examined machine learning frameworks in healthcare, focusing on their integration into clinical workflows. They highlighted the value of continuous model retraining and performance monitoring, ensuring real-time adaptability and accuracy in disease detection systems.

III. PROPOSED METHODOLOGY

There are numerous procedures for implementing machine learning methods for predicting lung cancer, including data collection, pre-processing, model training, and prediction. The suggested methodology's flow is intended to guarantee accuracy, interpretability, and real-world applicability.

The flowchart below illustrates the essential actions required in the process:

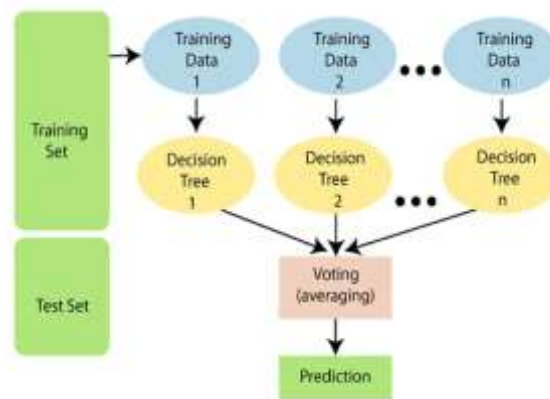


Figure 1: Flowchart of Random Forest Classifier

The recommended method of machine learning for predicting lung cancer is structured into a number of sequential steps, as shown in the flowchart. Data collecting is the initial phase of the process, where patient-related details like symptoms (e.g., persistent cough, chest pain), lifestyle factors (e.g., smoking habits), and demographic attributes (e.g., age, gender) are gathered. Data pre-processing is the next stage, which entails preparing the dataset to deal with incorrect or missing values, numerical encoding of categorical variables and normalizing continuous features to ensure uniformity. Following this, feature selection is conducted in order to determine, which features are most crucial for predicting lung cancer, thereby enhancing model efficiency and interpretability. The refined dataset is then divided into testing and training subsets to facilitate model training, where a random forest classifier is employed because of its robustness and high classification task accuracy. Once the model is trained, it proceeds to the prediction and evaluation stage, where the algorithm is tested on new patient data. The metrics including accuracy, precision, and recall are used to evaluate performance and confusion matrix outcomes. This flow guarantees that each stage—from input data to final prediction—contributes effectively to building a reliable and interpretable diagnostic instrument for the early identification of lung cancer.

IV. RESULT AND DISCUSSION

The suggested model for the identification of lung cancer was implemented utilizing a Random Forest classifier and evaluated on a structured dataset containing patient symptoms, lifestyle factors, and demographic information. Training and testing versions of the dataset were segregated in a 75:25 ratio to ensure effective model training and unbiased evaluation. The high predicted accuracy was shown by the model, which achieved a classification accuracy of approximately 97%, indicating its potential in identifying individuals at risk of lung cancer. Crucial performance metrics like as accuracy, recall, and the **confusion matrix** were analyzed to assess the effectiveness of the model. The confusion matrix highlighted a low rate of outcomes with false positives and false negatives, which are essential in a medical diagnostic application where incorrect predictions can have serious repercussions. Additionally, analysis of feature importance revealed that factors such as smoking, chronic coughing, chest pain, and age significantly contributed to the prediction outcome. These results confirm the model's reliability and validate the usefulness of methods for machine learning in medical diagnostics. However, continuous improvement through retraining with updated data and validation in clinical environments is recommended to enhance robustness and ensure real-world applicability.

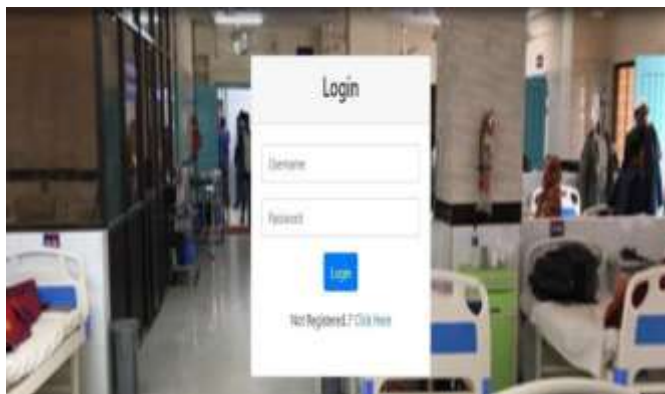


Figure 1: Lung cancer prediction interface: login interface

This page allows to login or register the user to the lung cancer prediction application.



Figure 2: Result representing lung cancer detected

The lung cancer detection model assessment was conducted utilizing a confusion matrix, which shed light on its classification performance. The data collected was divided into 75% for training and 25% for testing. The confusion matrix showed a high number very few false positives and false negatives, and a high proportion of real positives and true negatives, indicating excellent model reliability. The general precision of the model reached **97%**, demonstrating its strong capability in correctly identifying patients both with and without lung cancer. These findings imply that the random forest algorithm is effective for early-stage detection and can support timely medical intervention.

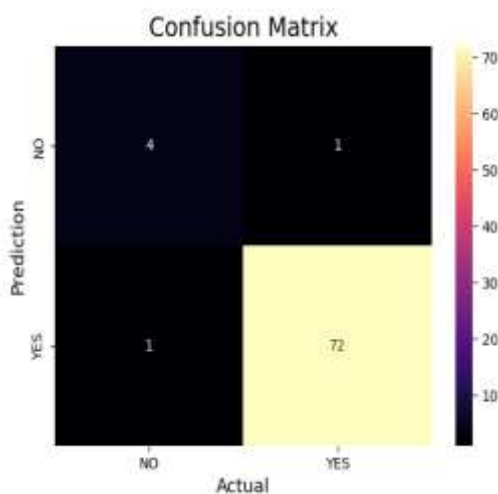


Figure 3: Model Performance Confusion Matrix

V. CONCLUSION AND FUTURE WORK

Finally, using random forest method, the proposed model for detecting lung cancer has shown high accuracy and reliability, attaining a classification precision of 97%. By analyzing structured information like patient symptoms, lifestyle habits, and demographic factors, the model effectively identifies individuals at risk of lung cancer, thus enabling early intervention and potentially reducing mortality rates. Using random forest improves classification performance in addition ensures model interpretability and robustness, making it a good fit to be included into clinical decision-support systems.

In the future, research will concentrate on expanding the model's capabilities by incorporating medical imaging data such as CT scans and chest X-rays to enhance diagnostic precision. Further improvements can include deploying the model in real-time healthcare applications for immediate risk assessment and feedback. Furthermore, investigating sophisticated deep learning techniques may yield even higher predictive accuracy. Validating the model on more extensive and varied datasets will also be essential to ensure generalization across different patient populations. Incorporating longitudinal patient data could allow the system to track the course of the illness over time, making it a complete lung cancer tool management.

VI. REFERENCES

- [1] S. Wang, Y. Zhou, Z. Zhang, et al., "Lung cancer detection using machine learning: A review," *IEEE Access*, vol. 8, pp. 89847–89863, 2020
- [2] K. Srinivas and B. K. Rani, "Prediction of lung cancer using machine learning techniques," *International Journal of Computer Applications*, vol. 179, no. 27, pp. 1–5, 2018
- [3] L. Zhang, X. Wang, and Y. Wang, "Random forest-based ensemble method for early lung cancer prediction," *Journal of Biomedical Informatics*, vol. 102, pp. 103356, 2020.
- [4] A. Sharma, R. Kumar, and A. Arora, "Performance analysis of machine learning algorithms for lung cancer detection," *Procedia Computer Science*, vol. 167, pp. 1810–1821, 2020.
- [5] M. Patel and T. Roy, "Machine learning in healthcare: Lung cancer detection and prediction," *Health Informatics Journal*, vol. 27, no. 4, pp. 1462–1475, 2021
- [6] J. Chen, D. Lin, and Y. Wang, "Application of ensemble learning for lung cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 107, 101909, 2020.
- [7] UCI Machine Learning Repository, "Lung Cancer Dataset," [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Lung+Cancer>
- [8] Breiman, L. "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.