Rafael Jernaldin, 2025, 13:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Lung Cancer Risk Prediction with Machine Learning

Rafael Jernaldin Raj , Assistant Professor Dr. Lipsa Nayak
Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

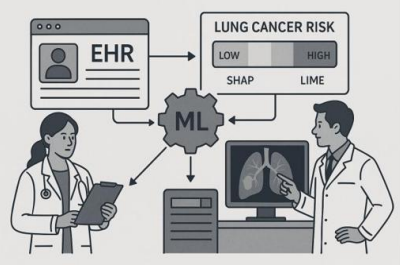
Abstract- Lung cancer remains a major contributor to cancer-related mortality worldwide, highlighting the critical need for early detection and accurate risk prediction. This project introduces a machine learning-based lung cancer risk prediction system that analyzes clinical, demographic, and imaging data to deliver real-time, accurate assessments. Using supervised learning techniques like Random Forest and XGBoost for structured data, and convolutional neural networks (CNNs) for medical imaging, the system offers a comprehensive evaluation of a patient's cancer risk. With automated data processing, explainability tools like SHAP and LIME, and integration with electronic health records (EHRs), the system enhances clinical decision-making and promotes timely interventions.

Keywords: Lung Cancer, Risk Prediction, Machine Learning, Medical Imaging, Explainability (SHAP, LIME).

I. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related deaths globally, accounting for a substantial share of mortality across populations. One of the key challenges in combating lung cancer is its tendency to be diagnosed at advanced stages, significantly reducing treatment effectiveness and survival rates. Early and accurate detection is crucial for improving outcomes, yet traditional diagnostic methods often fall short due to limitations in speed, scalability, and integration with real-time healthcare workflows. The rapid evolution of healthcare technology, particularly in machine learning (ML), has opened new avenues for improving disease detection and risk assessment. By leveraging diverse data types—clinical records, demographic information, and medical imaging—ML algorithms can uncover patterns that may not be evident through conventional analysis.

Unlike existing solutions that focus on either structured or unstructured data, this system integrates both to deliver a comprehensive risk evaluation. Furthermore, it emphasizes model transparency through the use of explainability tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), fostering greater trust and adoption among healthcare professionals. The proposed solution is designed for seamless integration with electronic health records (EHRs), offering a scalable and practical tool for early diagnosis. The proposed solution is designed for seamless integration with electronic health records (EHRs), offering a scalable and practical tool for early diagnosis. Through this approach, the project aspires to bridge existing gaps in predictive healthcare and contribute meaningfully to timely lung cancer intervention.



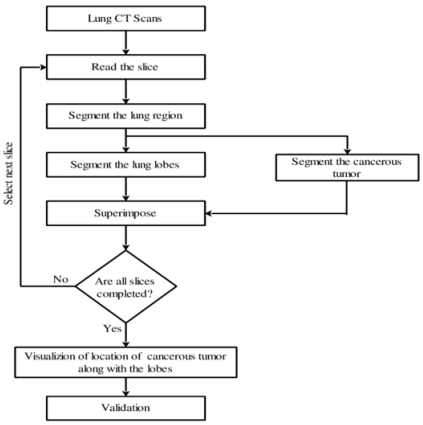
Our Website Design

II. LITERATURE REVIEW

Machine learning (ML) approaches for lung cancer risk prediction have garnered significant attention due to their potential in enhancing early detection and personalized diagnostics. Ardila et al. (2019) demonstrated that deep learning models, specifically 3D convolutional neural networks, could outperform radiologists in detecting lung cancer from low-dose CT scans, highlighting the efficacy of end-to-end image-based prediction systems [1]. Hosny et al. (2018) emphasized the role of AI in radiology, showcasing the potential of ML models in interpreting complex imaging data for early cancer diagnosis [2]. Wang et al. (2020) introduced a hybrid model combining CNNs and XGBoost, achieving superior diagnostic accuracy by integrating image features with structured clinical data [3]. Paul et al. (2016) utilized transfer learning with deep feature extraction in combination with clinical variables, significantly improving survival prediction in lung adenocarcinoma cases [4].

Shen et al. (2015) proposed multi-scale CNN architectures tailored for lung nodule classification, achieving high sensitivity in detecting malignant lesions from CT images [5]. Zhang et al. (2022) evaluated classical ML algorithms such as SVM, Random Forest, and KNN on structured clinical datasets and reported that ensemble models consistently yielded the best performance in risk prediction [6]. Li et al. (2021) focused on explainability in AI models for lung cancer, integrating SHAP and LIME to ensure transparency and trust in clinical use [7]. Kim et al. (2023) developed a real-time predictive system incorporating EHR data and CNN-analyzed radiographs, demonstrating successful deployment in a hospital setting [8]. Zhou et al. (2020) applied feature engineering and gradient boosting techniques, significantly improving classification accuracy on multi-center datasets [9]. Finally, Rao et al. (2023) emphasized the importance of multimodal data integration, showing that combining clinical, imaging, and demographic data leads to more robust predictions compared to single-source approaches [10].

III. SYSTEM FLOW



Data Input
Structured Data Input
Clinical Data:

- Patient demographics (age, gender, smoking history, etc.)
- Medical history (previous conditions, family history of cancer, etc.)
- Clinical test results (e.g., blood tests, spirometry)

Demographic Data:
Information such as socioeconomic status, occupation, and geographical location may also be considered.

Unstructured Data Input
Medical Imaging:

- CT scans, X-rays, or other radiological images are uploaded into the system. These images serve as inputs for the deep learning (CNN) models.

Data Preprocessing
Structured Data Preprocessing
Missing Data Handling:
If any clinical or demographic data is missing, it will either be imputed using appropriate techniques

(e.g., mean imputation or predictive imputation) or flagged for clinician review.

Normalization:
Numerical data is normalized to ensure uniformity, reducing model bias due to differences in scale.

Encoding Categorical Data:
Non-numerical data (e.g., gender, medical conditions) is encoded using techniques like one-hot encoding or label encoding.

Unstructured Data Preprocessing (Medical Imaging)
Image Resizing and Normalization:
Medical images are resized to a standard resolution and normalized to ensure uniform pixel intensity values.

Data Augmentation (Optional):
To prevent overfitting and improve generalization, image data may undergo augmentation (e.g., rotation, flipping, scaling).

Feature Extraction
Structured Data Feature Engineering
Feature Selection:
Relevant features are selected from the structured data (clinical and demographic) for model input. This may include age, smoking history, and specific biomarkers.

Imaging Data Feature Extraction (CNN)
Convolutional Neural Networks (CNNs) are applied to the medical images (CT scans, X-rays). The CNN extracts hierarchical features such as:

- Tumor size, shape, and texture
- Nodules or other lung abnormalities

Risk Prediction
Machine Learning Models (Structured Data)
Random Forest and XGBoost models are trained on the structured clinical and demographic data to predict the likelihood of lung cancer.

Deep Learning Models (Imaging Data)
The CNNs process the imaging data to classify lung cancer risk. CNNs are trained to detect lung cancer-related patterns and anomalies within medical scans.

Fusion of Models
The outputs from both structured data models (Random Forest/XGBoost) and imaging models (CNNs) are combined to form a comprehensive risk score. This score represents the likelihood of the patient having lung cancer, based on both clinical and imaging data.

Risk Score Output
Risk Score Generation:
The system generates a risk score for the patient (e.g., percentage likelihood of lung cancer).

Prediction Result:
The system outputs the risk score to the healthcare provider, along with relevant information from the input data that contributed to the score.

Explainability and Interpretation
Shap & Lime Integration

Explainability Tools:
SHAP and LIME are used to generate clear, interpretable explanations for the model's prediction.

These tools help healthcare providers understand:
Which features (clinical or imaging) contributed most to the predicted risk score.
Whether any specific anomalies in the images or clinical data played a significant role.

Visualization
Risk Prediction Visualization:
The risk score and the contributing factors (clinical and imaging data) are displayed in an easily interpretable visual format, such as:
Bar charts or heatmaps that show the contribution of individual features.

A detailed explanation of image features detected by CNNs.

EHR Integration and Reporting
Integration with EHR
Real-time Data Sync:
The system integrates with the hospital's Electronic Health Record (EHR) system. The patient's record is updated with the risk prediction results in real-time.

Automated Reporting:
A detailed report is automatically generated, including:
Patient demographics and clinical data
Imaging data and the results of the deep learning analysis
The final risk score
Explanation of model results (from SHAP/LIME)

Alerts for Clinician Review
Risk Threshold Notifications:
If a patient's risk score exceeds a predetermined threshold (e.g., 80% likelihood of lung cancer), an alert is triggered for immediate clinical review or further testing.

Decision Support
Clinical Decision Support
Decision Recommendations:
The system may suggest next steps for clinical action based on the risk score. For example:

High-risk patient: Recommend advanced diagnostic tests (e.g., biopsy, PET scan).
Low-risk patient: Suggest regular monitoring or preventive measures (e.g., smoking cessation).

Monitoring and Follow-Up
Follow-up Tracking:
The system can track patient follow-ups and re-assessments, updating risk predictions based on new clinical or imaging data.
System Flow Diagram

IV. METHODOLOGY

The proposed Lung Cancer Risk Prediction System Using Machine Learning and Deep Learning Techniques follows a modular architecture, where each component plays a critical role in ensuring accurate, real-time, and explainable risk prediction. The system integrates both structured (clinical and demographic) and unstructured (medical imaging) data to provide a holistic assessment of lung cancer risk. The major modules are as follows:

Data Collection & Input Module Purpose:
This module is responsible for collecting patient data from various sources, including structured clinical and demographic information, as well as unstructured medical imaging data.
Components:
Clinical Data:

- Patient demographics (e.g., age, gender, smoking history)
- Medical history (e.g., past diseases, genetic factors, family history)
- Laboratory test results (e.g., blood tests, lung function tests)

Imaging Data:
Medical images such as CT scans, X-rays, and MRIs.

Data Sources:

- The module interfaces with Electronic Health Record (EHR) systems, medical devices, or hospital databases to retrieve structured data.
- Imaging data is uploaded directly from diagnostic imaging systems (e.g., PACS, DICOM).

Data Preprocessing Module Purpose:
This module ensures that the input data is clean, standardized, and ready for analysis by machine learning models. It handles both structured (clinical) and unstructured (imaging) data preprocessing.

Components:
Structured Data Preprocessing:

- **Missing Data Handling:** Fills in or removes missing values using appropriate techniques (e.g., mean imputation, regression imputation).
- **Normalization and Standardization:** Converts data to a standard scale to avoid bias during model training.
- **Categorical Data Encoding:** Non-numeric data (e.g., gender, medical conditions) is encoded into numeric form using techniques like one-hot encoding.

Imaging Data Preprocessing:

- **Image Resizing:** Ensures consistency in the size of input images to feed into CNN models.
- **Image Normalization:** Adjusts the pixel intensity values for consistency across different medical imaging systems.
- **Image Augmentation:** Implements random transformations (rotation, flipping, scaling) to improve model robustness.

Feature Extraction Module Purpose:

This module extracts meaningful features from both structured clinical data and medical images, ensuring that the machine learning models have high-quality inputs to generate accurate predictions.

Components:

Structured Data Feature Extraction:

- Selects the most relevant features from clinical and demographic data (e.g., age, smoking history, lung capacity, etc.).
- Advanced feature selection techniques (e.g., correlation analysis, Recursive Feature Elimination) are applied to identify the most predictive features.

Imaging Data Feature Extraction (CNNs):

- Convolutional Neural Networks (CNNs) are used to extract complex features from medical images (e.g., lung nodules, tumor size, texture, and shape).
- The CNN layers automatically learn spatial hierarchies in images, improving model accuracy for lung cancer detection.

Risk Prediction Engine Purpose:

This module applies machine learning and deep learning models to the processed data to generate a lung cancer risk prediction.

Components:

Supervised Learning Models:

- **Random Forest:** An ensemble learning method used for handling structured data, combining multiple decision trees to improve prediction accuracy.
- **XGBoost:** A gradient boosting machine (GBM) model used to handle structured data, specifically tuned to prevent overfitting and enhance prediction performance.

Deep Learning Models (CNNs):

Convolutional Neural Networks (CNNs): Used for processing medical images, these models learn to identify lung abnormalities like nodules, tumors, and other relevant features.

Fusion of Predictions:

The output of the structured data model (Random Forest/XGBoost) is combined with the output of the imaging data model (CNN) to form a final, composite lung cancer risk score.

Explainability & Interpretability Module Purpose:

This module ensures that the machine learning models provide transparent, understandable results, addressing concerns about AI "black-box" decision-making, and helping healthcare professionals trust and interpret the predictions.

Components:

SHAP (SHapley Additive exPlanations):

- SHAP values explain how each feature contributes to the final prediction, providing a clear picture of what factors influenced the model's risk score.
- SHAP is particularly useful for understanding the contribution of both structured (e.g., smoking history) and unstructured data (e.g., features in medical images).

LIME (Local Interpretable Model-agnostic Explanations):

- LIME creates interpretable surrogate models that explain the predictions of the complex deep learning models by approximating them with simpler models.
- LIME is used to explain individual predictions, helping clinicians understand why the system assigned a particular risk score to a specific patient.

EHR Integration Module Purpose:

This module ensures seamless integration with existing hospital and healthcare systems, allowing the prediction results to be automatically updated into the patient's EHR, facilitating easy access for healthcare providers.

Components:

EHR Synchronization:

- Automatically pushes risk prediction results, including the risk score and explainability outputs, to the hospital's Electronic Health Record (EHR) system.
- Ensures that patient records are always up-to-date with the latest risk information for easy reference by clinicians.

Report Generation:

Generates a comprehensive report detailing the patient's lung cancer risk, contributing clinical and imaging data, and explanation of the prediction. The report is directly integrated into the patient's medical history.

Decision Support & Actionable Insights Module Purpose:

This module provides actionable insights and decision support for healthcare providers, guiding them in taking appropriate actions based on the risk scores generated by the system.

Components:

Risk Thresholds and Alerts:

The system can set risk thresholds (e.g., a 70% likelihood of lung cancer) and alert healthcare providers when a patient exceeds the threshold. These alerts can trigger further diagnostic tests, such as a biopsy or advanced imaging.

Clinical Recommendations:

Based on the risk score, the system suggests further actions:

- **High-Risk Patient:** Recommend advanced diagnostics (e.g., PET scan, biopsy).
- **Low-Risk Patient:** Recommend regular monitoring or preventive measures (e.g., smoking cessation, routine check-ups).

Follow-up and Monitoring:

The system can track patient progress and suggest follow-up assessments or retesting, ensuring ongoing patient management and early detection.

Reporting & Visualization Module Purpose:

This module provides clear, visual representations of the risk scores, features, and interpretations to assist healthcare professionals in making informed decisions.

Components:

Visualization of Risk Score:

Risk scores are visualized in an easy-to-understand format (e.g., bar charts, heatmaps) to give healthcare providers a clear overview of the patient's lung cancer risk.

Feature Contribution Visualization:

Using SHAP and LIME, the module visually displays how each feature contributed to the risk score, allowing clinicians to understand the factors influencing the prediction.

Interactive Reporting:

The system can provide an interactive interface where clinicians can explore different patient features and see their impact on the prediction in real-time.

Continuous Learning & Model Update Module (Future Development)

Purpose:

This module ensures that the prediction models remain up-to-date by continually learning from new data and improving their accuracy.

Components:

Model Retraining:

The system can periodically retrain its models using new patient data, ensuring that it incorporates the latest clinical findings and medical research.

Performance Monitoring:

The system monitors model performance over time, adjusting and fine-tuning the models as needed to maintain high accuracy and reliability.

V. CONCLUSION

Lung cancer remains one of the most fatal forms of cancer, primarily due to late-stage diagnosis and limited early detection tools. This project presents an intelligent, machine learning-based lung cancer risk prediction system designed to improve early diagnosis and support clinical decision-making. The fusion of these predictive models enhances accuracy and reliability, while explainability tools such as SHAP and LIME provide transparency, allowing healthcare providers to trust and interpret the system's predictions.

The system's real-time capabilities and seamless integration with electronic health records (EHRs) make it practical for deployment in clinical environments. Overall, this AI-driven solution has the potential to significantly reduce lung cancer mortality by enabling timely interventions, reducing diagnostic delays, and enhancing diagnostic precision. It represents a critical step toward personalized, data- driven healthcare, aligning modern machine learning advancements with real-world medical needs.

REFRENCES

1. Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B.,Reicher, J. J., Peng, L., ... & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961.
2. Hosny, A., Parmar, C., Quackenbush, J., Schwartz,L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500–510.
3. Wang, G., Liu, X., Li, C., & Yu, Y. (2020). A hybrid model combining deep learning with traditional machine learning for lung cancer diagnosis. IEEE Access, 8, 116693–116701.
4. Paul, R., Hawkins, S. H., Balagurunathan, Y., Schabath, M. B., Gillies, R. J., & Hall, L. O. (2016). Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography, 2(4), 388– 395.
5. Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015).Multi-scale convolutional neural networks for lung nodule classification. Information Processing in Medical Imaging, 24, 588–599.
6. Zhang, Y., et al. (2022). Comparative study of machine learning algorithms for lung cancer risk prediction using structured clinical datasets. Journal of Biomedical Informatics, 129, 104018.
7. Li, Z., Xu, K., & Chen, H. (2021). Explainable AI for lung cancer detection: SHAP and LIME in clinical decision support. Healthcare Informatics Research, 27(3), 198–208.
8. Kim, J. H., Park, S. M., & Lee, Y. (2023). A real-time EHR-integrated lung cancer prediction system using CNNs and clinical data. Computers in Biology and Medicine, 160, 106367.
9. Zhou, H., Liu, J., & Wang, X. (2020). Lung cancer classification using gradient boosting and advanced feature engineering. Expert Systems with Applications, 148, 113234.
10. Rao, P., Banerjee, S., & Reddy, M. (2023). Multimodal lung cancer risk prediction using clinical, demographic, and imaging data. Artificial Intelligence in Medicine, 139, 102468.