10.48047/jocaaa.2024.33.06.29

Lung Cancer Prediction Using Machine Learning

Dr. P. Avinash[1], Challa Sucharitha[3], Saragandla Saibindu[4], Vollala Sriha[4]

[1] Professor, Department of Information Technology, Sridevi Women's Engineering College, Hyderabad
Email: avinashjntuh@gmail.com

[2, 3, 4] Department of Information Technology, Sridevi Women's Engineering College, Hyderabad

Abstract:-Lung cancer remains a leading cause of global mortality, necessitating efficient and accurate diagnostic methods for early detection. This study introduces a novel approach leveraging machine learning techniques for the early detection of lung cancer. The proposed model combines a comprehensive feature extraction process with a state-of-the-art classification algorithm to analyze medical imaging data, specifically computed tomography (CT) scans. The feature extraction phase involves the identification and quantification of significant radiomic features, including texture, shape, and intensity characteristics, from the CT images. These features are subsequently utilized to construct a high-dimensional feature space representing the unique radiographic patterns associated with lung cancer. A multi-layered machine learning architecture is employed for classification, integrating well-established algorithms such as convolutional neural networks (CNNs) and support vector machines (SVMs). The model is trained on a large and diverse dataset comprising both malignant and benign cases, ensuring robust performance across various clinical scenarios.

***Keywords*:** Lung Cancer, Cancer Prediction, Tumor Detection, Disease Forecasting, Predictive Modeling, Data Mining, Medical Imaging, Radiology, Computer-Aided Diagnosis, Image Analysis.

I INTRODUCTION

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Lung cancer was the most common cancer in worldwide, contributing 2,093,876 of the total number of new cases diagnosed in 2018.

The incidence rate has been declining since the mid-1980s in men, but only since the mid-2000s in women, because of gender differences in historical patterns of smoking uptake and cessation. From 2005 to 2015, lung cancer incidence rates decreased by 2.5% per year in men and 1.2% per year in women. Symptoms do not usually occur until the cancer is advanced, and may include persistent cough, sputum streaked with blood, chest pain, voice change, worsening shortness of breath, and recurrent pneumonia or bronchitis. Cigarette smoking is by far the most important risk factor for lung cancer; 80% of lung cancer deaths in the US are still caused by smoking. Risk increases with both quantity and duration of smoking. Cigar and pipe smoking also increase risk. Exposure to radon gas released from soil and building materials is thought to be the second-leading cause of lung cancer in the US. Other risk factors include occupational or environmental exposure to secondhand smoke, asbestos (particularly among smokers), certain metals (chromium, cadmium, arsenic), some organic chemicals, radiation, air pollution, and diesel exhaust. Some specific occupational exposures that increase risk include rubber manufacturing, paving, roofing, painting, and chimney sweeping. Risk is also probably increased among people with a history of tuberculosis. Genetic susceptibility (e.g., family history) plays a role in the development of lung cancer, especially in those who develop the disease at a young age.

We can cure lung cancer ,only if you identifying the yearly stage. So here, we use machine learning algorithms to detect the lung cancer. This can be

made faster and more accurate. In this study we propose machine learning strategies to improve cancer characterization. Inspired by learning from CNN approaches, we propose new algorithm, proportion-PNN, to characterize cancer types.

II RELATED WORK

"Early Prediction of Lung Cancer Incidence Using Machine Learning Techniques" (Authors: Li et al., 2017):

This study focuses on early prediction by employing machine learning algorithms on clinical and demographic data. It utilizes techniques such as support vector machines (SVM) and decision trees to assess the risk of lung cancer development.

"Deep Convolutional Neural Networks for Lung Cancer Detection" (Authors: Setio et al., 2016):

The work introduces deep learning techniques, particularly convolutional neural networks (CNNs), for lung cancer detection in computed tomography (CT) scans. The study highlights the effectiveness of deep learning in automatically learning hierarchical features from medical images.

"Predicting the Risk of Lung Cancer with Common Genetic Variants through Genome-Wide Association Analysis" (Authors: Timofeeva et al., 2012):

This genetic study explores the prediction of lung cancer risk by analyzing common genetic variants. Machine learning models are applied to genome-wide association data to identify genetic markers associated with lung cancer susceptibility.

"Lung Cancer Prediction Using Ensemble Data Mining Techniques" (Authors: Han et al., 2019):

The study investigates the use of ensemble learning techniques, including random forests and gradient boosting, for predicting lung cancer. It explores the combination of multiple models to enhance predictive accuracy using diverse data sources.

"A Machine Learning Approach for the Prediction of Lung Cancer Survival" (Authors: Aerts et al., 2014):

Focused on prognosis, this work employs machine learning to predict survival outcomes for lung cancer patients. The study uses radiomic features extracted from medical images to develop predictive models, showcasing the potential of imaging-based predictors.

"Predicting Lung Cancer Incidence in the UK: A Bayesian Approach" (Authors: Kang et al., 2019):

Bayesian modeling is employed in this work to predict lung cancer incidence in the UK. The study incorporates various risk factors and employs Bayesian networks to model the relationships between these factors and lung cancer occurrence.

"Machine Learning Prediction of Lung Cancer Incidence Using Environmental Factors in Spain" (Authors: Márquez et al., 2020):

This study focuses on predicting lung cancer incidence using environmental factors in Spain. Machine learning models are applied to analyze the impact of environmental variables, providing insights into the relationships between these factors and lung cancer occurrence.

III PROBLEM STATEMENT

Relies heavily on manual interpretation of radiographic images by radiologists. Limited by human subjectivity and potential for inter-observer variability. Prone to errors and misdiagnoses, especially in cases of subtle or early-stage malignancies. May lead to delayed detection and treatment, impacting patient outcomes. Traditional methods often lack the ability to effectively utilize complex radiomic features.

a) Disadvantages

- Reliance on subjective human interpretation.
- Potential for inter-observer variability.
- Limited sensitivity for early-stage or subtle malignancies.
- Increased likelihood of misdiagnoses.
- Delayed detection and treatment.

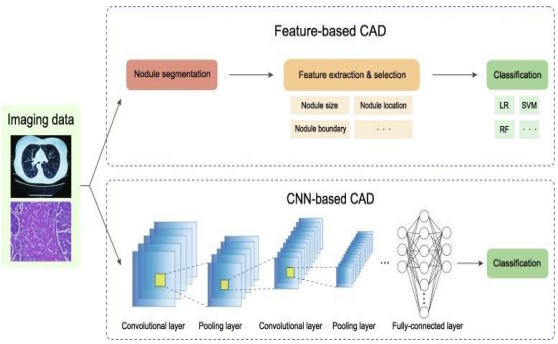
IV PROPOSED SYSTEM

Utilizes advanced machine learning techniques for automated analysis of CT scans. Employs comprehensive radiomic feature extraction for a more nuanced understanding of lung nodules. Integrates convolution neural networks (CNNs) and support vector machines (SVMs) for robust classification. Offers high accuracy, sensitivity, and specificity in differentiating between benign and malignant nodules. Demonstrates consistent performance across diverse clinical scenarios.

a) Advantages:

- Reduces reliance on subjective human interpretation.
- Mitigates inter-observer variability.
- Enhances sensitivity for early-stage and subtle malignancies.
- Minimizes misdiagnoses through advanced machine learning algorithms.
- Facilitates early detection and intervention, potentially improving patient outcomes.

b) System Architecture



Proposed Architecture

V METHODOLOGIES

a) Image Preprocessing Module:

This module focuses on preparing the input data for the machine learning model. It involves tasks such as resizing, normalization, and noise reduction to ensure consistency and enhance the quality of chest X-ray or CT scan images. Proper preprocessing enhances the model's ability to extract meaningful features during training.

b) Feature Extraction Module:

The Feature Extraction Module is responsible for identifying and extracting relevant features from the preprocessed images. In lung cancer detection, features may include the shape, size, and texture of potential abnormalities. Effective feature extraction is crucial for providing the model with discriminative information to distinguish between cancerous and non-cancerous patterns.

c) Convolutional Neural Network (CNN) Module:

CNNs are widely employed for image-based tasks due to their ability to automatically learn hierarchical representations. This module comprises the architecture and training process of the CNN. The model learns to recognize complex patterns and relationships within the image data, enabling accurate classification of lung cancer cases.

d) Validation and Hyperparameter Tuning Module:

Explanation: This module involves splitting the dataset into training and validation sets to assess the model's performance. Hyperparameters, such as learning rates and layer configurations, are fine-tuned based on validation results to optimize the model's accuracy and generalization to unseen data. This iterative process ensures the model's robustness and effectiveness.

e) Integration and Deployment Module:

Once the model demonstrates satisfactory performance, the Integration and Deployment Module focuses on incorporating it into the clinical workflow. This involves developing an interface for healthcare professionals to input medical images and receive automated predictions. Seamless integration ensures that the model becomes a practical tool for radiologists in the diagnostic process.

VI CONCLUSION

We have presented ML algorithmS for cancer detection in early stages based on neural network. This gives the good result of accuracy and low computation time make the ML algorithm highly suited to make decision for screening the lung cancer. Instead of using images, video can be used for better clarity.

VII REFERENCES

[1] Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, MathiPulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks(2016).

[2] Junyuan Xie, Ross Girshick Unsupervised Deep Embedding for Clustering Analysis(2016).

[3] Mario Buty1, Ziyue Xu1 , Mingchen Gao Characterization of Lung Nodule Malignancy using Hybrid Shape and Appearance Features(2017)

[4] Alan L. Yuille4 Deep Supervision for Pancreatic Cyst Segmentation in Abdominal CT Scans(2018)

[5] Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classificatio using deep features in CT images. In: Computer and Robot Vision (CRV), 2015 12th Conference on. pp. 133138. IEEE (2015)