Lung Cancer Analysis and Prediction

Aditya Choudhary aditya20489@iiitd.ac.in
Akhil Sharma akhil20171@iiitd.ac.in
Shrugal Tayal shrugal20408@iiitd.ac.in
Vaibhav Rajpal vaibhav20146@iiitd.ac.in

Abstract

This paper provides an overview of the analysis of factors causing lung cancer and discusses the development of a machine learning model to forecast lung cancer during its early stages. We will do this with the aid of early-stage symptoms and causes as features for our dataset. We will also discuss some challenges in developing and validating our techniques and outline the path to clinical adoption.

1. Introduction and Motivation

Lung cancer is one of the leading causes of mortality in every country, affecting both men and women. Lung cancer has a low prognosis, resulting in a high death rate due to its late stage detection.

1.1. Language

All manuscripts are in English.

1.2. Data Acquisition effort

We will be using a Kaggle dataset hosted publicly under the link:
kaggle.com/datasets/h13380436001/h-lung-cancer

1.3. Project Timeline

Week No.	Deliverable Task
1	Pre Processing Data, Exploring New Training Models, Feature Extraction for baseline Model
2	
3	
4	Training Models, Analysis and their working. Preparing for the interim Presentation.
5	
6	
7	Working on accuracy and corrections. Exploring the untouched models for further evaluations
8	
9	
10	Compiling the Observations and Concluding the Result

1.4. Individual Tasks

Aditya Choudhary: Pre-Processing of data, Exploring Model Training approaches.
Akhil Sharma: Feature extraction for baseline, Model analysis.
Shrugal Tayal: Evaluation of respective models with their Accuracy Prediction.
Vaibhav Rajpal: Evaluation of K-fold cross validation while finding its best model/approach.

1.5. Related Work

To develop and inspect the Machine learning model and the accuracy ratios of three different classifiers such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Convolutional Neural Network (CNN). To achieve this, we explore feature extraction techniques like Histogram of Oriented Gradients(HoG) coupled with a prediction model like Support Vector Classifier to give comparable results. Therefore we review progress made toward developing lung cancer prediction models.

1.6. Final Outcome

From this ML model, we are trying to predict if a person has lung cancer or not based on various habits and symptoms they have. Our Main goal from this project is to predict if reducing harmful habits like smoking and more reduce the chance of Lung Cancer. We are diagnosing based on their early symptoms and will see if it would work or not.

1.7. References :

1) Abdulazeez ,A .,Swallow ,A ., Abdullah ,D. Lung Cancer Classification based on Correlation Selection. Retreived from https://www.researchgate.net/publication/351889513_Lung_cancer_Prediction_and_Classification_based_on_Correlation_Selection_method_Using_Machine_Learning_Techniques

2) Kadir ,T .,Gleeson ,F. Lung Cancer Prediction using ML and Advanced Imaging Techniques. Retrieved from https://tlcr.amegroups.com/article/view/21998/16754

3) Liu ,S .,Yao ,W .Prediction of Lung Cancer using Gene Expression and Deep Learning with KL Divergence Gene Selection. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04689-9