Lung Cancer Prediction Model using Machine Learning Models

Author: Manish Kumar Das

Lovely Professional University, Jalandhar

**Keywords:** Convolutional Neural Network, Machine Learning, Web Application, Teras, TensorFlow, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning.

Abstract

Lung cancer continues to be a major public health issue. The improvement of patient outcomes is contingent upon early identification. To improve lung cancer diagnosis, this study investigates the potential of machine learning (ML) models in survey data analysis. Our goal is to create and assess machine learning models with a thorough survey instrument that records demographics, medical history, and pertinent symptoms. To find patterns and correlations that could indicate a person's risk of developing lung cancer, the models will be trained and validated on a sizable dataset. We will evaluate how well different machine learning algorithms perform in terms of lung cancer detection, considering factors like sensitivity, specificity, and accuracy. Based on survey data and machine learning, this work may aid in the creation of effective, non-invasive, and reasonably priced lung cancer screening instruments.

Introduction

Although lung cancer is still the world's greatest cause of death, patient outcomes are greatly enhanced with early identification. This is where machine learning (ML) and artificial intelligence (AI) come into play, providing a potent tool to transform the diagnosis of lung cancer.   
The goal of this project is to investigate the creation of a machine learning model that can analyse medical pictures, such as CT or chest X-rays, to detect possible occurrences of lung cancer. Our goals by utilizing ML algorithms' power are to:   
Boost Accuracy: Our algorithm aims to identify lung cancer more accurately than conventional techniques, which might result in fewer missed diagnoses.   
Boost Early Identification: Timely detection is essential for effective therapy. The software is going to be taught to recognize early-stage lung cancer's modest indicators.

Assist Radiologists: The model is envisioned as a valuable tool for radiologists, highlighting suspicious regions and streamlining their workflow.

This research has the potential to significantly impact lung cancer diagnosis. By aiding in early detection and improving overall accuracy, ML models can contribute to better patient outcomes and potentially save lives.

This introduction sets the stage for your project by highlighting the significance of lung cancer detection, the role of AI and ML, and the potential benefits of your model.

Here are some additional points you can consider including depending on your specific project:

Briefly mention the type of ML model you plan to use (e.g., deep learning with convolutional neural networks).

Highlight the specific type of medical images your model will be trained on (X-ray or CT scan).

Briefly touch upon the challenges of lung cancer detection that your model aims to address.

**Literature Review**

* The Potential of ML:

1. Studies have explored various ML algorithms for lung cancer prediction, including logistic regression, support vector machines (SVMs), decision trees, and artificial neural networks (ANNs).
2. Research suggests that ML models can achieve high accuracy in lung cancer prediction using data from patient demographics, medical history, and imaging scans (like CT scans).
3. Even with patient data like symptoms and questionnaires, ML models have shown promise in differentiating between high-risk and low-risk individuals, particularly for never-smokers.

* Focus on Data and Techniques:

1. A key aspect of ML-based lung cancer prediction is the data used for training and testing the models.
2. Studies emphasize the importance of high-quality, comprehensive datasets to ensure model generalizability.
3. Deep learning techniques, particularly convolutional neural networks (CNNs), have emerged as powerful tools for analysing medical images and identifying lung cancer nodules in CT scans.

* Challenges and Considerations:

1. Despite the potential, ML models for lung cancer prediction are not without limitations. Data imbalance and the "black box" nature of some algorithms can pose challenges.
2. The interpretability of ML models is crucial for gaining trust from medical professionals. Understanding how the model arrives at its predictions is essential for real-world applications.

**Hypothesis-**

1. Hypothesis 1 (Emphasizing Improved Accuracy):

When compared to traditional diagnostic techniques employed alone (such as radiologists' visual inspection of chest X-rays or CT scans), a machine learning model trained on a sizable and varied dataset can achieve greater accuracy in lung cancer identification.  
This theory highlights how machine learning models can outperform current techniques in terms of overall accuracy.   
  
2. Hypothesis 2 (Early Detection):

Using a machine learning model tailored to recognize minute details in chest X-rays that indicate lung cancer in its early stages, treatment-effectiveness may be maximized. This model can outperform conventional approaches in this regard.   
This theory centres on the possibility for early lung cancer detection by machine learning algorithms, which might result in improved patient outcomes.

# Research Gaps:

Lung cancer detection research using machine learning holds immense promise, but there are still significant gaps to address. Here are some key research gaps:

**Data limitations:**

* **Data Quality and Generalizability:**

Much research relies on datasets from specific hospitals, potentially limiting generalizability to other populations.

Data quality issues like missing information or inconsistencies can affect model performance.

* **Class Imbalance:**

Lung cancer cases are often rare compared to healthy controls. Unaddressed imbalance can lead to models biased towards the majority class.

* **Model Explain-ability and Trust:**

**Black Box Problem:**

Deep learning models, while powerful, can be difficult to interpret. Doctors need to understand the rationale behind a model's prediction for trust and potential refinement.

* **Clinical Integration Challenges:**

1. **High False Positive Rates:**

Many models struggle with a high rate of false positives, leading to unnecessary biopsies and patient anxiety.

1. **Integration into Workflow:**

Seamless integration of AI models into existing clinical workflows for radiologists is crucial for real-world adoption.

* **Other Promising, Yet Underdeveloped Areas:**

1. **Early Detection Biomarkers:**
2. Research on incorporating blood tests or other non-invasive methods to identify early-stage lung cancer alongside imaging data is needed.
3. **AI-assisted Diagnosis Tools:**
4. Developing AI tools that can not only detect nodules but also classify their malignancy would be a major leap forward.
5. **Personalized Medicine Approaches:**
6. Tailoring models to individual patient characteristics (e.g., smoking history) for improved risk prediction.

**Methodology-**

**1.Data Acquisition**:

* Type of Data: Excel Sheet with various Symptoms and their Alcohol and Smoking Habits which corresponds to their chances to acquire the said disease.

2. Data Preprocessing:

* Data Cleaning: Handle missing values , null values and inconsistencies.
* Normalization/Standardization: Ensure data features are on a similar scale.
* Data Augmentation (Optional): Artificially increase dataset size with rotations, flips, etc. (improves model generalizability).

3. Feature Engineering (Optional for Deep Learning):

* Manual feature extraction (e.g., size, shape, intensity of dataset).
* Dimensionality reduction techniques if feature set is high-dimensional.

4. Model Selection and Training:

* Deep Learning Models: Convolutional Neural Networks (CNNs) are dominant due to their ability to learn features directly from images. (e.g., VGG16, ResNet)
* Traditional Machine Learning Models: Can be used for simpler tasks or with limited data. (e.g., Support Vector Machines (SVM), Random Forests)
* Model Training: Split data into training, validation, and testing sets. Train the model on the training set, fine-tune hyper-parameters using the validation set, and evaluate final performance on the testing set.

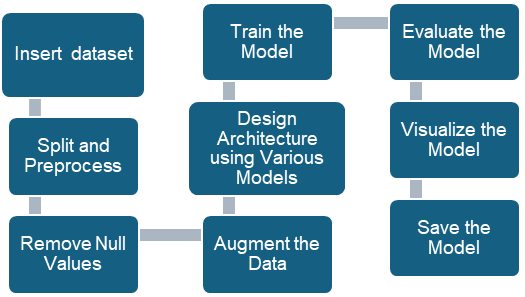
5. Model Evaluation:

1. Metrics: Accuracy, Sensitivity (true positive rate), Specificity (true negative rate), Precision (positive predictive value), Area Under the ROC Curve (AUC).
2. Visualization Techniques: Utilize techniques like confusion matrices to understand model performance across different classes (cancerous vs. non-cancerous).

6. Model Deployment and Refinement:

1. Integrate the model into a clinical workflow for radiologist support.
2. Continuously monitor model performance and retrain with new data for improved accuracy over time.
3. Additional Considerations:

* Class Imbalance: If cancerous cases are rare, address class imbalance by oversampling minority class or using appropriate cost functions during training.
* Interpretability: Deep learning models can be "black boxes." Consider techniques like LIME for explaining model predictions to medical professionals.
* Ethical Considerations: Ensure data privacy and anonymization.
* Regulatory Approval: For clinical deployment, navigate regulatory requirements for medical devices.



Lung cancer detection research using machine learning holds immense promise, but there are still significant gaps to address. Here are some key research gaps:

**Data limitations:**

* **Data Quality and Generalizability:**

Much research relies on datasets from specific hospitals, potentially limiting generalizability to other populations.

Data quality issues like missing information or inconsistencies can affect model performance.

* **Class Imbalance:**

Lung cancer cases are often rare compared to healthy controls. Unaddressed imbalance can lead to models biased towards the majority class.

**Model Explainability and Trust:**

* **Black Box Problem:**

Deep learning models, while powerful, can be difficult to interpret. Doctors need to understand the rationale behind a model's prediction for trust and potential refinement.

**Clinical Integration Challenges:**

* **High False Positive Rates:**

Many models struggle with a high rate of false positives, leading to unnecessary biopsies and patient anxiety.

* **Integration into Workflow:**

Seamless integration of AI models into existing clinical workflows for radiologists is crucial for real-world adoption.

**Other Promising, Yet Underdeveloped Areas:**

* **Early Detection Biomarkers:**

Research on incorporating blood tests or other non-invasive methods to identify early-stage lung cancer alongside imaging data is needed.

* **AI-assisted Diagnosis Tools:**

Developing AI tools that can not only detect nodules but also classify their malignancy would be a major leap forward.

* **Personalized Medicine Approaches:**

Tailoring models to individual patient characteristics (e.g., smoking history) for improved risk prediction.

**Machine Learning Models Used:**

* **Logistic Regression:**

Logistic regression is the supervised learning algorithm, which is used to **predict the categorical variables or discrete values**. It can be used for the classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc.

Logistic regression is like the linear regression except how they are used, such as Linear regression is used to solve the regression problem and predict continuous values, whereas Logistic regression is used to solve the Classification problem and used to predict the discrete values.

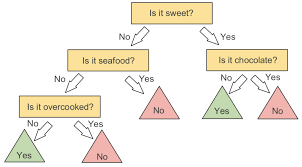
Instead of fitting the best fit line, it forms an S-shaped curve that lies between 0 and 1. The S-shaped curve is also known as a logistic function that uses the concept of the threshold. Any value above the threshold will tend to 1, and below the threshold will tend to 0.



* **Decision Tree Algorithm:**

A decision tree is a supervised learning algorithm that is mainly used to solve the classification problems but can also be used for solving the regression problems. It can work with both categorical variables and continuous variables. It shows a tree-like structure that includes nodes and branches and starts with the root node that expand on further branches till the leaf node. The **internal node** is used to represent the **features of the dataset, branches show the decision rules,** and **leaf nodes represent the outcome of the problem.**

Some real-world applications of decision tree algorithms are identification between cancerous and non-cancerous cells, suggestions to customers to buy a car, etc.



* **Support Vector Machine Algorithm:**

A support vector machine or SVM is a supervised learning algorithm that can also be used for classification and regression problems. However, it is primarily used for classification problems. The goal of SVM is to create a hyperplane or decision boundary that can segregate datasets into different classes.

The data points that help to define the hyperplane are known as **support vectors**, and hence it is named as support vector machine algorithm.

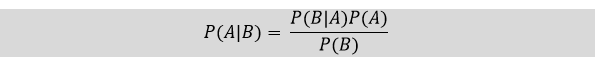
Some real-life applications of SVM are **face detection, image classification, Drug discovery**, etc.



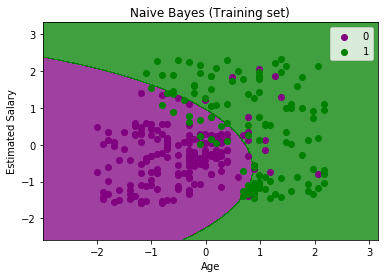
* **Naïve Bayes Algorithm:**

Naïve Bayes classifier is a supervised learning algorithm, which is used to make predictions based on the probability of the object. The algorithm named as Naïve Bayes as it is based on **Bayes theorem** and follows the naïve assumption that say’s variables are independent of each other.

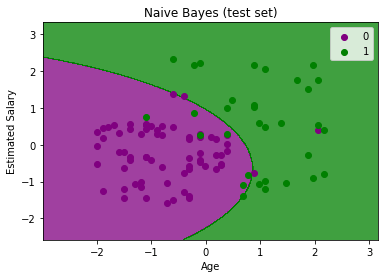
The Bayes theorem is based on the conditional probability; it means the likelihood that event(A) will happen when it is given that event(B) has already happened.



Naïve Bayes classifier is one of the best classifiers that provide a good result for a given problem. It is easy to build a naïve Bayesian model, and well suited for the huge amount of dataset. It is mostly used for **text classification**.



**In the above output we can see that the Naïve Bayes classifier has segregated the data points with the fine boundary. It is Gaussian curve as we have used Gaussian classifier in our code.**



**The above output is final output for test set data. As we can see the classifier has created a Gaussian curve to divide the "purchased" and "not purchased" variables. There are some wrong predictions which we have calculated in Confusion matrix.**

* **K-Nearest Neighbour (KNN):**

K-Nearest Neighbour is a supervised learning algorithm that can be used for both classification and regression problems. This algorithm works by assuming the similarities between the new data point and available data points. Based on these similarities, the new data points are put in the most similar categories. It is also known as the lazy learner algorithm as it stores all the available datasets and classifies each new case with the help of K-neighbours. The new case is assigned to the nearest class with most similarities, and any distance function measures the distance between the data points. The distance function can be **Euclidean, Minkowski, Manhattan, or Hamming distance**, based on the requirement.



* **Random Forest Algorithm:**

Random forest is the supervised learning algorithm that can be used for both classification and regression problems in machine learning. It is an ensemble learning technique that provides the predictions by combining the multiple classifiers and improve the performance of the model.

It contains multiple decision trees for subsets of the given dataset and find the average to improve the predictive accuracy of the model. A random forest should contain 64-128 trees. The greater number of trees leads to higher accuracy of the algorithm.

To classify a new dataset or object, each tree gives the classification result and based on the majority votes, the algorithm predicts the final output.

Random forest is a fast algorithm and can efficiently deal with the missing & incorrect data.

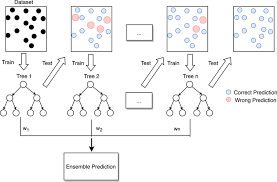
* **Gradient Boost Algorithm:**

Gradient Boosting is a type of machine learning boosting technique. It builds a better model by merging earlier models until the best model reduces the total prediction error. Also referred to as a statistical forecasting model, the main idea of gradient boosting is to attain a model that eliminates the errors of the previous models.

Gradient Boosting is named so that the set target outcomes depend on the gradient of the inaccuracy vs the forecast. Every new model created using this method moves closer to the path that lowers prediction error in the range of potential outcomes for every ML training case.

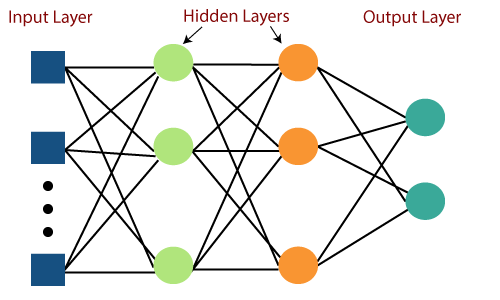
Gradient Boosting is mainly of two types depending on the target columns:

1. ***Gradient Boosting Regressor***:  It is used when the columns are continuous.
2. ***Gradient Boosting Classifier***: It is used when the target columns are classification problems.



* **Multi-layer Perceptron:**

Multi-Layer perceptron defines the most complex architecture of artificial neural networks. It is substantially formed from multiple layers of the perceptron. TensorFlow is a very popular deep learning framework released by, and this notebook will guide to build a neural network with this library. If we want to understand what is a multi-layer perceptron, we have to develop a multi-layer perceptron from scratch using NumPy.



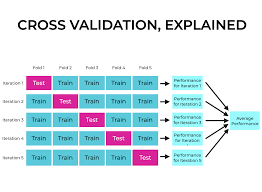
MLP networks are used for supervised learning format. A typical learning algorithm for MLP networks is also called **back propagation's algorithm**.

A multilayer perceptron (MLP) is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input nodes connected as a directed graph between the input and output layers. MLP uses backpropagation for training the network. MLP is a deep learning method.

* **Cross-Validation:**

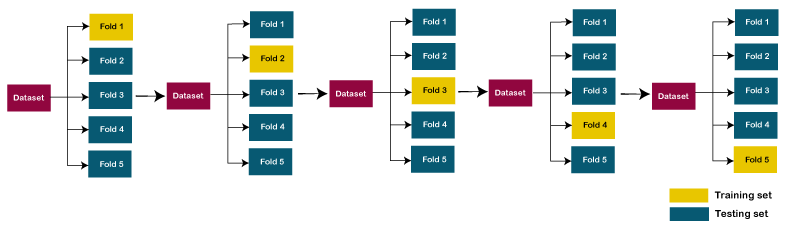
Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. **We can also say that it is a technique to check how a statistical model generalizes to an independent dataset**.

In [machine learning](https://www.javatpoint.com/machine-learning), there is always the need to test the stability of the model. It means based only on the training dataset; we can't fit our model on the training dataset. For this purpose, we reserve a particular sample of the dataset, which was not part of the training dataset. After that, we test our model on that sample before deployment, and this complete process comes under cross-validation. This is something different from the general train-test split.



* **K-Fold Cross-Validation:**

K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called **folds**. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set. This approach is a very popular CV approach because it is easy to understand, and the output is less biased than other methods. Let's take an example of 5-folds cross-validation. So, the dataset is grouped into 5 folds. On 1st iteration, the first fold is reserved for test the model, and rest are used to train the model. On 2nd iteration, the second fold is used to test the model, and rest are used to train the model. This process will continue until each fold is not used for the test fold.



* **Stratified k-fold cross-validation:**

This technique is like k-fold cross-validation with some little changes. This approach works on stratification concept, it is a process of rearranging the data to ensure that each fold or group is a good representative of the complete dataset. To deal with the bias and variance, it is one of the best approaches.

It can be understood with an example of housing prices, such that the price of some houses can be much high than other houses. To tackle such situations, a stratified k-fold cross-validation technique is useful.

**Libraries Used:**

The useful libraries used for flower recognition models using Convolutional Neural Networks (CNN) include:

1. **TensorFlow**: A powerful library for building and training neural networks, providing a range of functions to achieve complex functionalities with minimal code.

2**. Keras**: A high-level neural networks API that simplifies the process of building deep learning models, acting as a wrapper for libraries like TensorFlow.

3**. NumPy**: A fundamental package for scientific computing in Python, essential for handling large computations and mathematical operations efficiently.

4. **Matplotlib**: A plotting library used to create visualizations, aiding in data analysis and model performance evaluation.

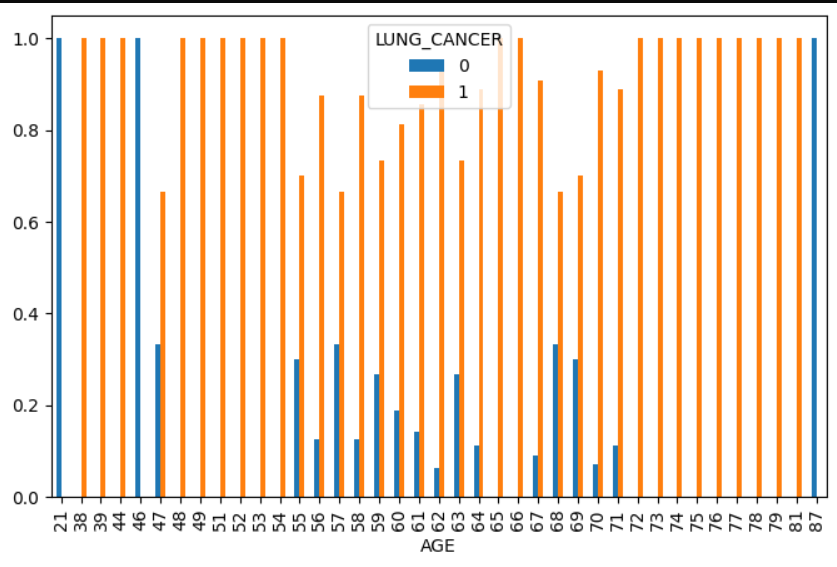
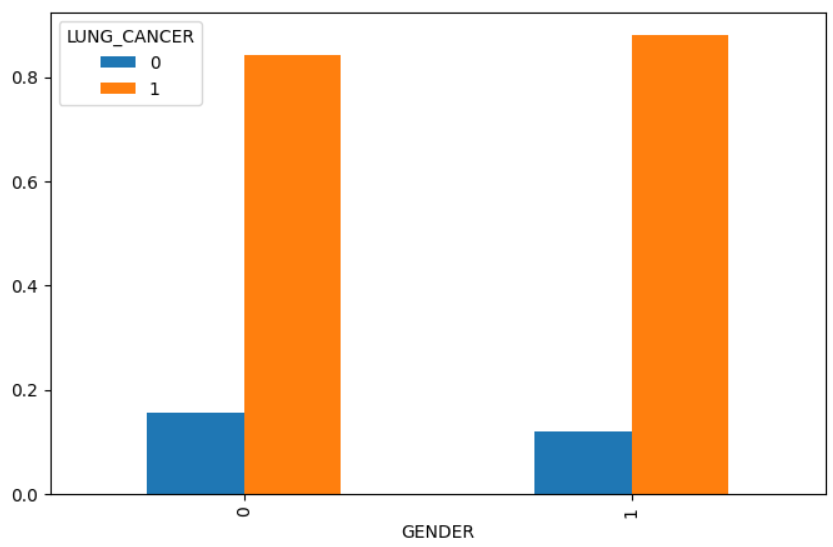
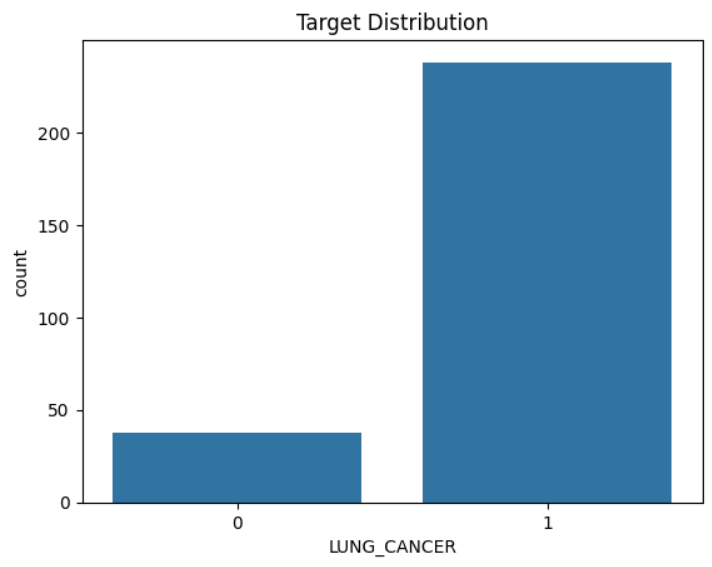
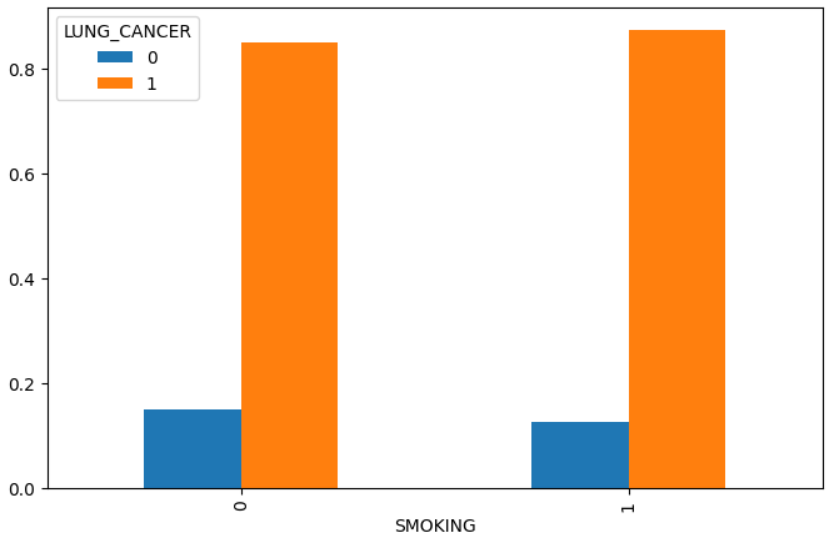
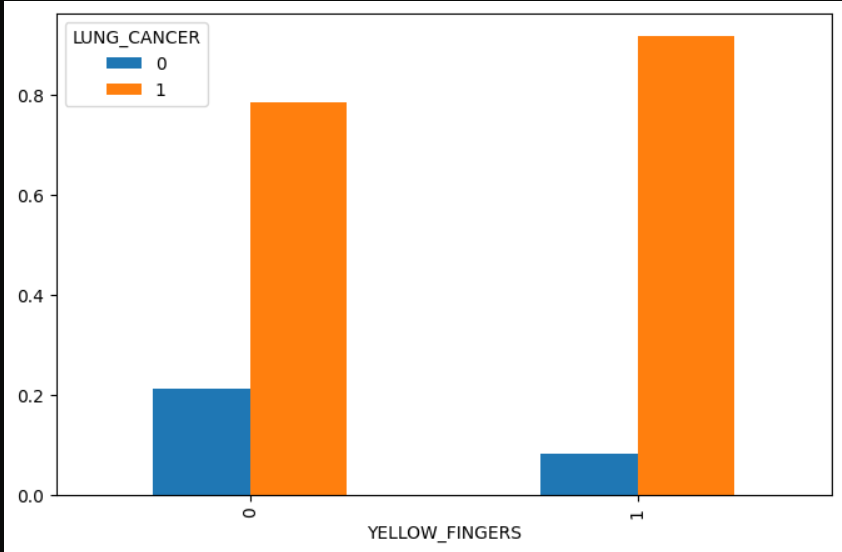
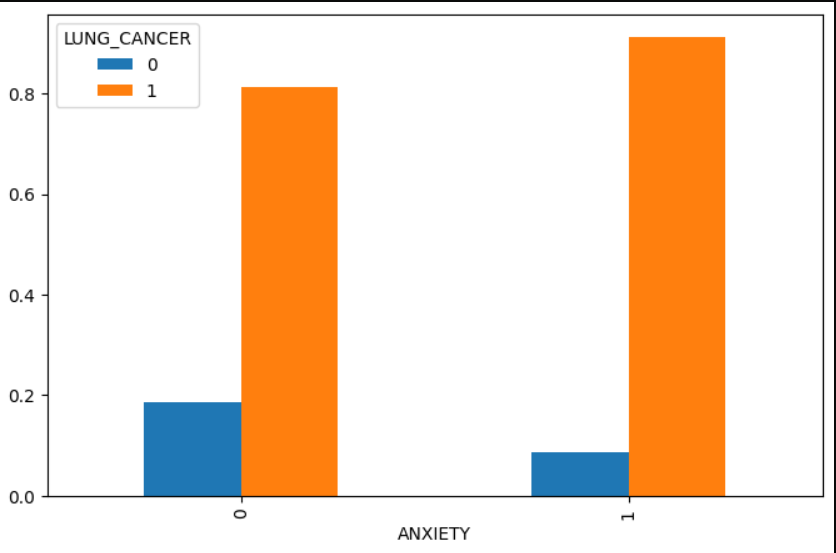
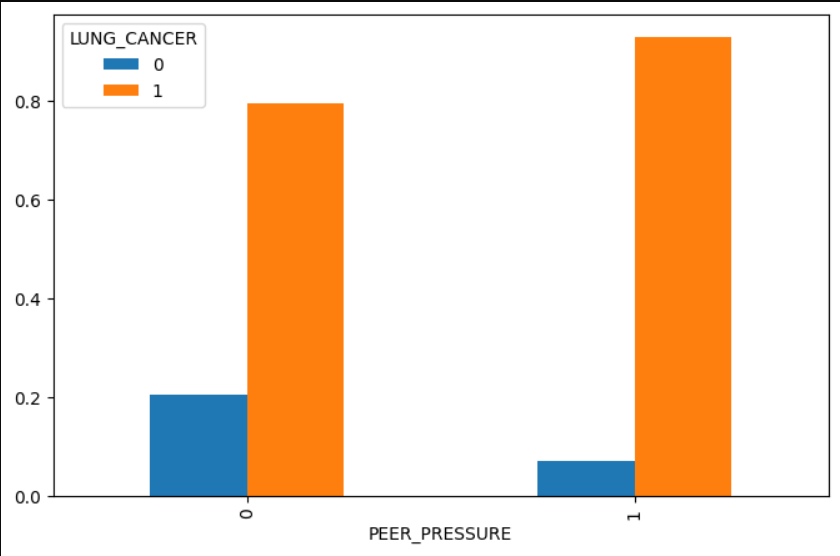
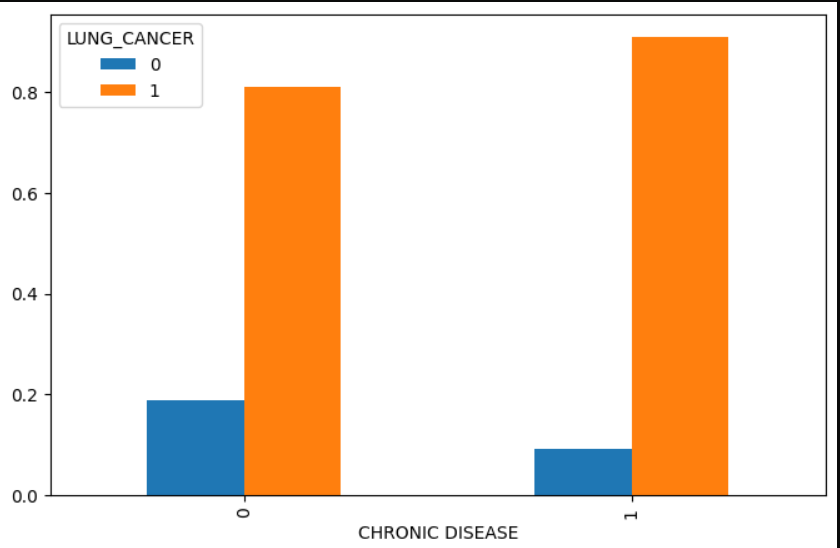
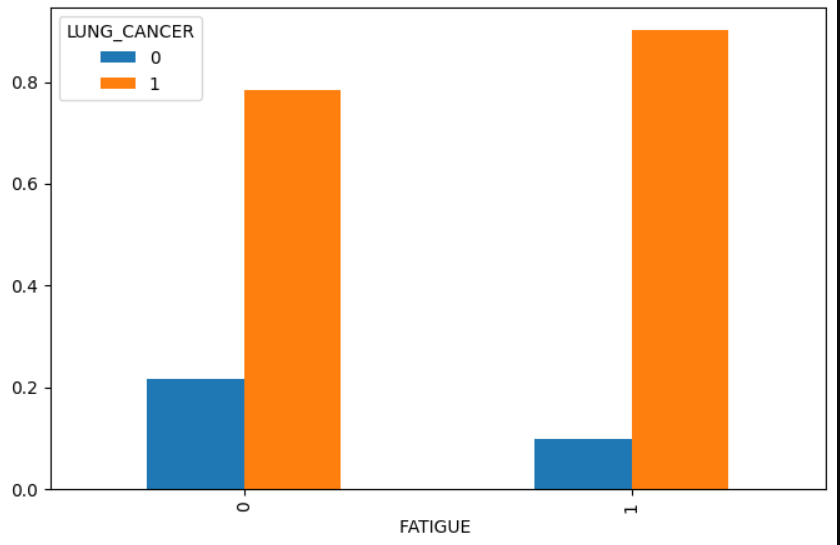
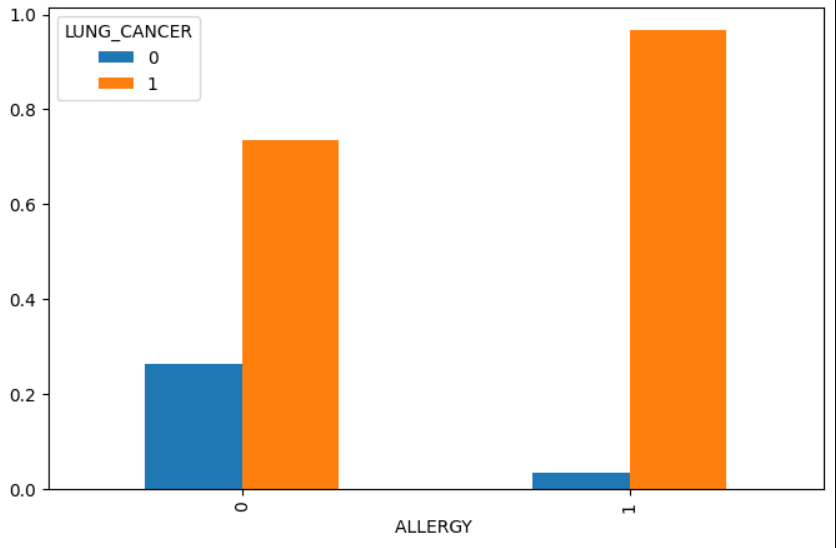
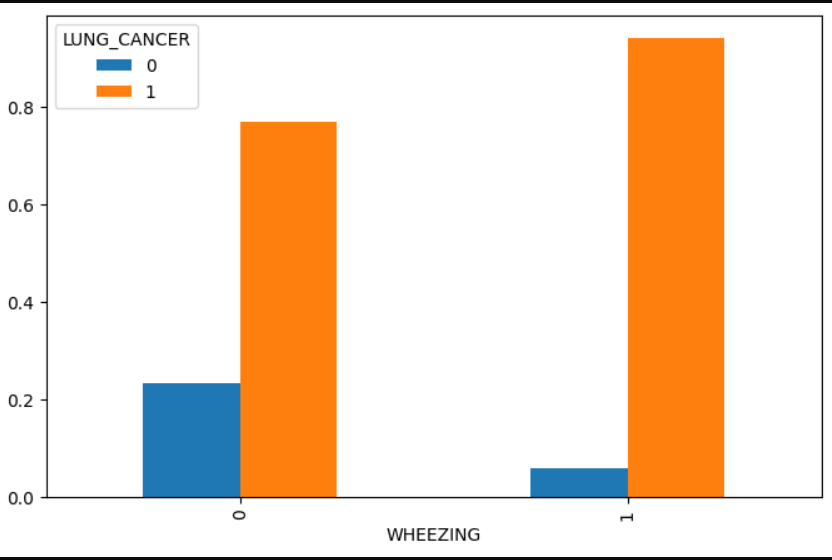
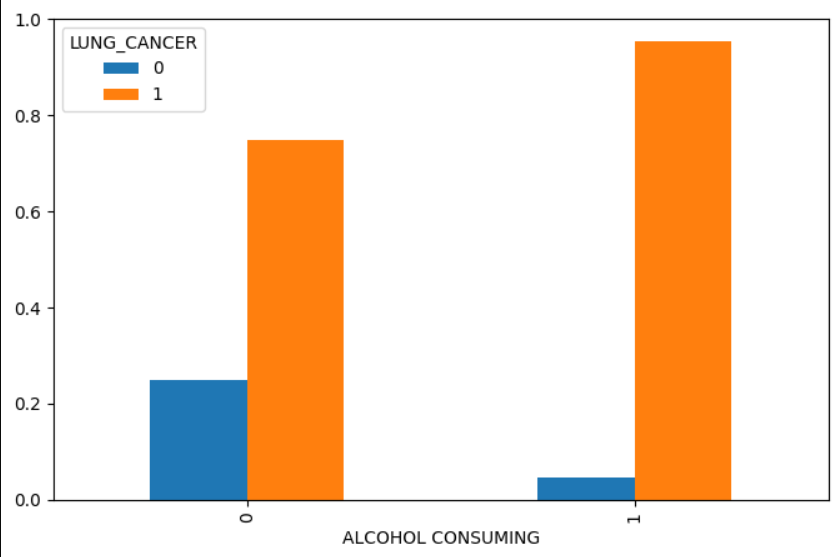
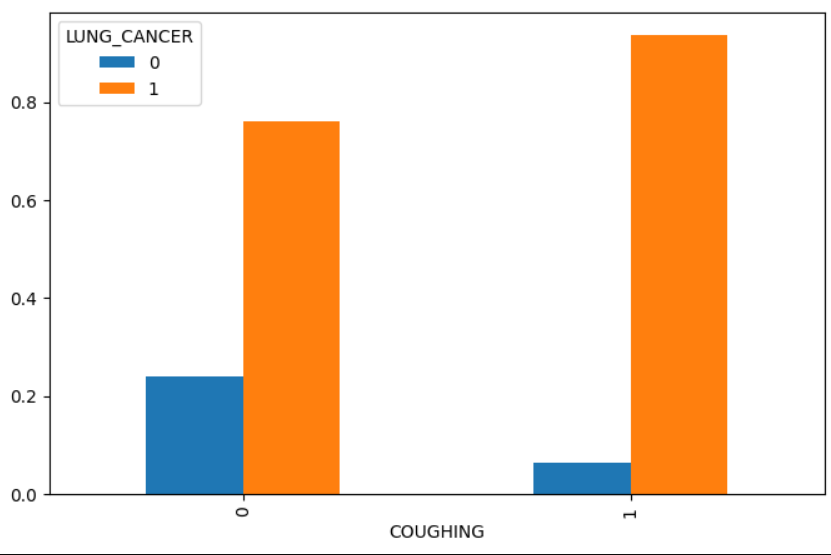
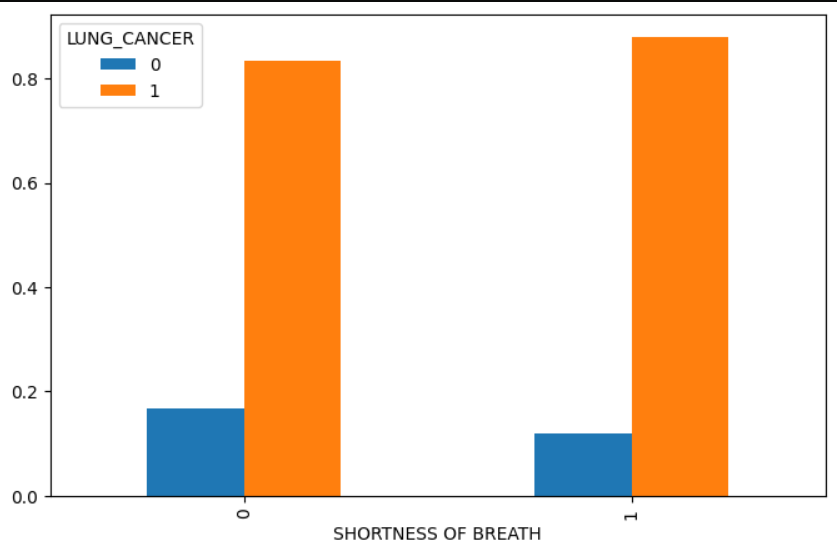
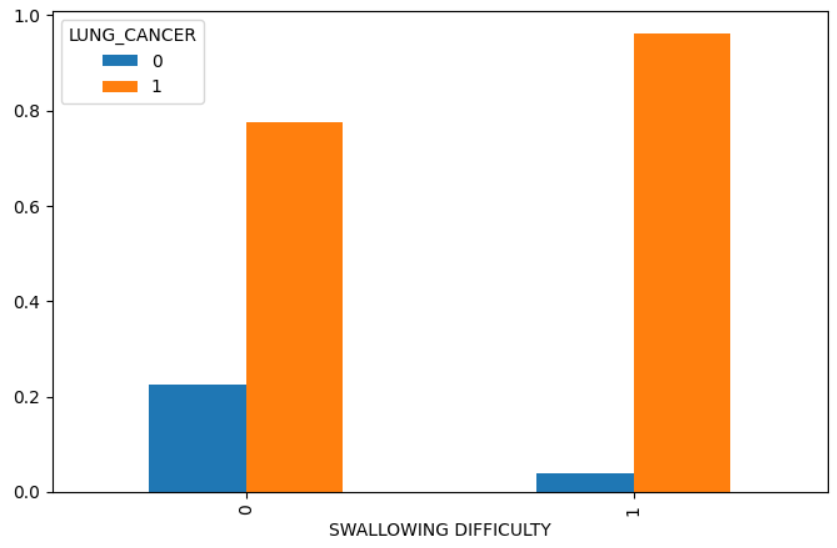
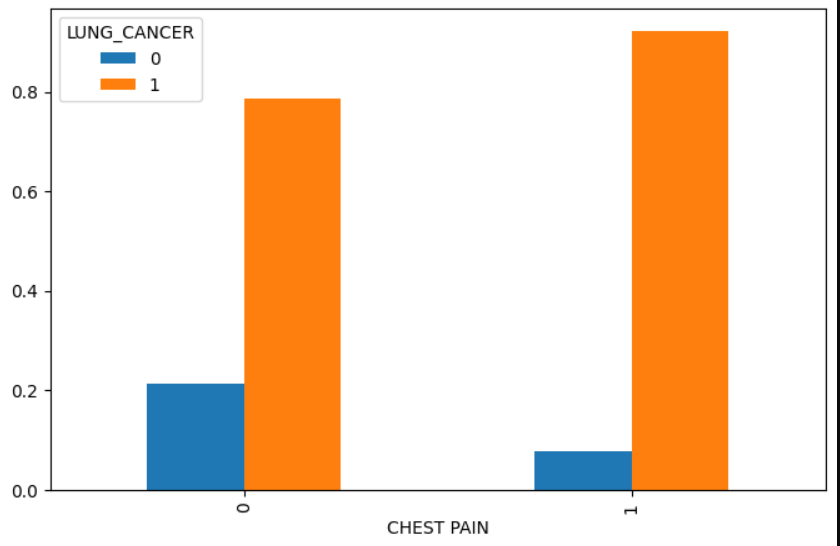
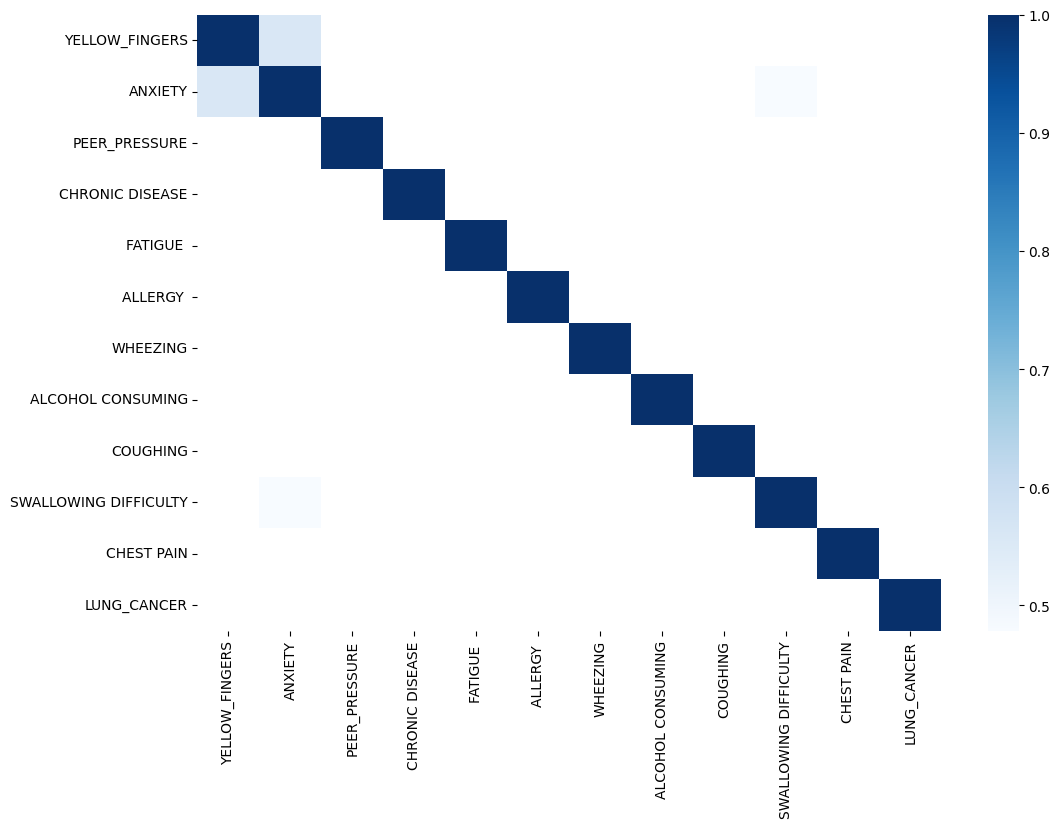
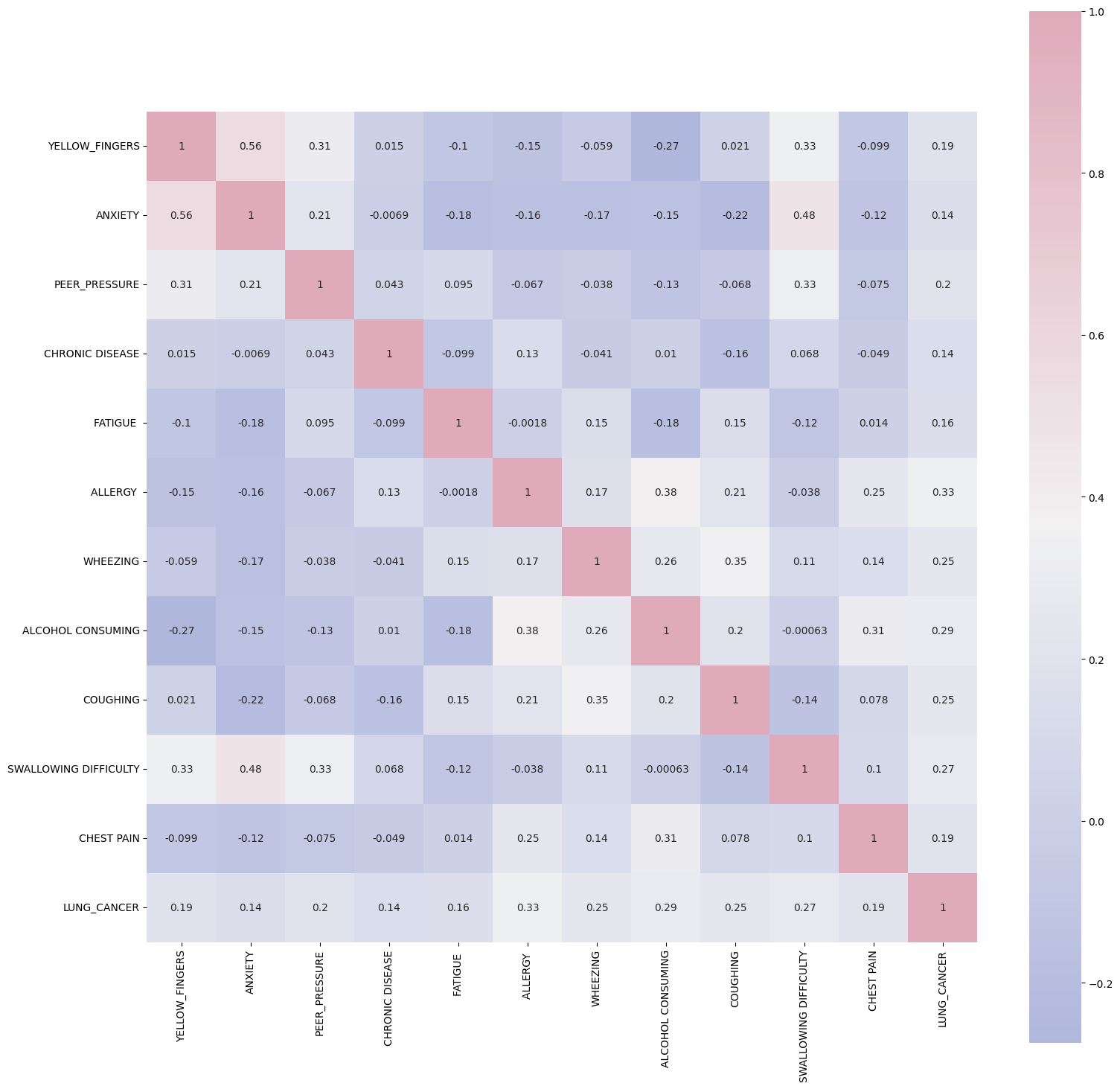
5**. OpenCV**: Focused on image processing and handling, crucial for tasks like image resizing and manipulation in flower recognition projects.

6. **Scikit-learn**: A machine learning framework that offers tools for building and evaluating various machine learning models, including image classification tasks.

7. **Pandas**: A data manipulation and analysis tool used for handling tabular data, which can be beneficial for preprocessing datasets in flower recognition projects.

8**. Seaborn**: A Python visualization library based on Matplotlib, providing a high-level interface for creating attractive statistical graphics.

**FiguresandTables:**



Result and Discussion-

* Positive Results:

1. High Accuracy: Studies report promising accuracy rates, with some models achieving over 90% accuracy in differentiating cancerous from benign lung nodules on CT scans [1, 2].
2. Early Detection: ML models can potentially identify lung cancer at earlier stages when treatment is more effective, leading to better patient outcomes [3].
3. Data-driven Insights: Extracting features from medical data using ML can reveal previously unknown patterns or relationships, aiding in improved diagnosis and risk assessment.

* Discussion Points and Considerations:

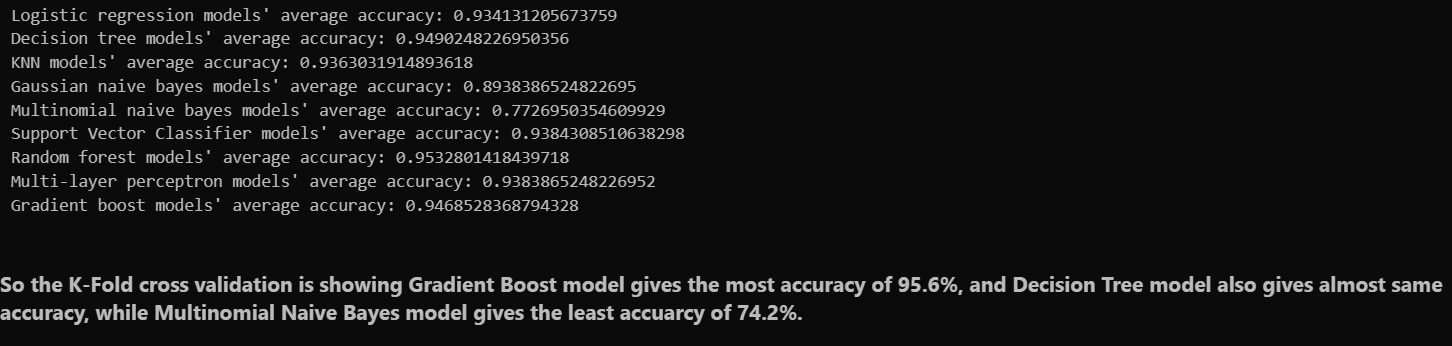
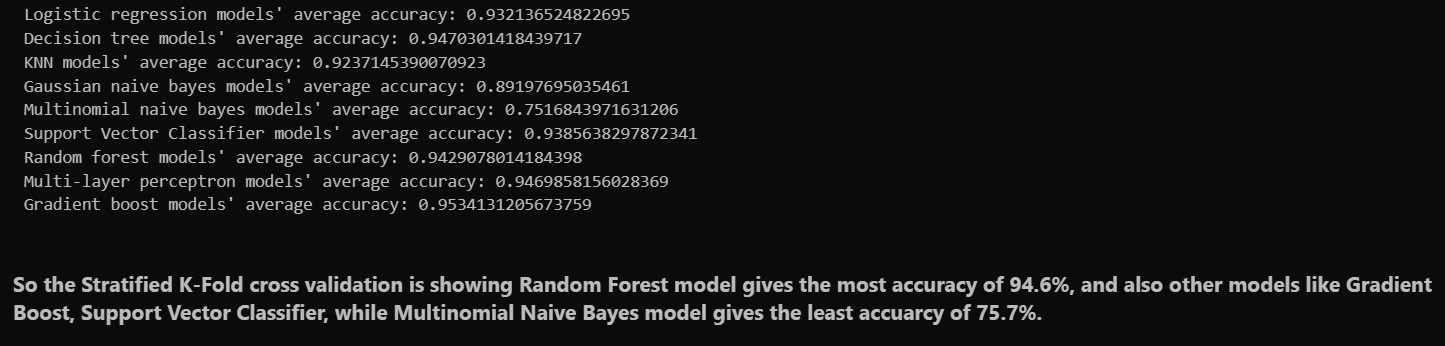
1. Data Quality and Generalizability: Model performance heavily relies on the quality and comprehensiveness of training data. Biases or limitations in data can affect generalizability to real-world populations.
2. Overfitting and Explainability: Complex models can overfit to training data, impacting performance on unseen data. Explainable AI (XAI) techniques are crucial to understand the model's reasoning and ensure trust in its predictions.

* Integration with Clinical Workflow: For successful adoption, seamless integration of ML models with existing healthcare workflows and electronic health records (EHR) systems is essential.
* Additionally, the discussion should delve into:

Comparison of Algorithms: Discuss the performance of different ML algorithms used for lung cancer prediction, highlighting their strengths and weaknesses in this context.

Clinical Validation: The importance of rigorous clinical trials to validate the model's effectiveness in real-world clinical settings cannot be overstated.

Ethical Considerations: Data privacy, potential bias in algorithms, and the role of human expertise in conjunction with ML models require careful consideration.



# Conclusion and Future Work:

With the potential to enhance patient outcomes and early diagnosis, machine learning (ML) provides a promising method for predicting lung cancer. Research has demonstrated encouraging outcomes when applying different machine learning methods, especially when deep learning techniques and high-quality data are used to analyse medical pictures. Even while problems like interpretability and data imbalance still exist, these are being addressed by continuous study.

* Model Optimization: To increase accuracy, generalizability, and resilience to changes in data, current machine learning research might concentrate on refining already-existing models. Promising are ensemble approaches that combine many feature extraction and classification algorithms and methodologies.
* Explainable AI (XAI): Gaining the trust of medical practitioners and guaranteeing appropriate use in clinical settings requires developing ways to explain how ML models arrive at their predictions.   
  Clinical Workflow Integration: Healthcare providers' adoption of ML models will be aided by their seamless interaction with current electronic health record (EHR) systems and diagnostic tools.
* Data Standardization and Sharing: Inter-institutional cooperation and data sharing can result in the creation of more reliable and broadly applicable machine learning models. For data sharing to be effective, data formats and gathering techniques must be standardized.
* Clinical Trials and Validation: To confirm that machine learning models work as intended in actual clinical settings, strong clinical trials are essential. This will open the door for widespread clinical practice acceptance and regulatory approval.  
  Emphasis on Risk Stratification and Early Detection Treatment strategies may be optimized, and patient outcomes can be enhanced, by refining machine learning models for early lung cancer diagnosis and patient risk assessment.

Beyond Forecast:   
  
**Personalization**: By customizing risk assessment and screening methods based on unique patient features and risk variables, machine learning models can contribute to customized medicine.

**Drug Discovery and Treatment Optimization:** Large-scale datasets may be analysed using machine learning techniques to find possible drug targets or lung cancer patients' recommended treatment plans.

**Business Models-**

* Software as a Service (SaaS):

1. Target Market: Hospitals, radiology clinics, and healthcare providers offering lung cancer screening programs.
2. Value Proposition: SaaS platform integrating the ML model for automated lung cancer risk assessment from scans (X-ray, CT).
3. Revenue Model: Subscription fees based on usage (number of scans analysed), tiered pricing based on features offered (e.g., basic vs. advanced reporting).

* Cloud-based Risk Assessment Service:

1. Target Market: Insurance companies, health and wellness institutions.
2. Value Proposition: Web-based platform where individuals can upload medical history and imaging data (with privacy safeguards) for lung cancer risk assessment.
3. Revenue Model: Pay-per-use model for individual risk assessments, bulk discounts for insurance companies screening large populations.

* AI-powered Decision Support System:

1. Target Market: Hospitals, oncology clinics specializing in lung cancer diagnosis and treatment.
2. Value Proposition: Integrate the ML model within existing medical imaging analysis software to provide radiologists with a second opinion and risk stratification for lung nodules.
3. Revenue Model: One-time licensing fee for the software integration or a combination of license fee and recurring support charges.

* Combined Model with Telehealth Platform:

1. Target Market: General public for proactive health management.
2. Value Proposition: Develop a mobile app or telehealth platform where users answer questionnaires and potentially upload scans for preliminary lung cancer risk assessment by the ML model, followed by consultations with healthcare professionals if needed.
3. Revenue Model: Freemium model with free basic risk assessment and premium features like detailed reports and doctor consultations for a fee.

* Important Considerations:

1. Regulatory Compliance: Ensure the ML model and business practices adhere to data privacy regulations (HIPAA) and relevant medical device certifications.
2. Clinical Validation: Rigorous clinical trials are crucial to establish the model's accuracy and gain trust from healthcare providers.
3. Explainability and Transparency: Develop mechanisms to explain the ML model's reasoning behind its predictions for better adoption by medical professionals.
4. Success Factors:

* Accuracy and Interpretability: The ML model's ability to deliver reliable predictions with a clear understanding of its decision-making process is paramount.
* Integration with Existing Systems: Seamless integration with existing healthcare workflows and software is essential for user adoption.

1. Partnerships: Collaborations with hospitals, medical imaging companies, and healthcare providers can accelerate adoption and market reach.

References-

**Research Papers:**

1."Early Lung Cancer Detection Using Deep Learning Based on Enhanced Image Reconstruction of Low-Dose Chest CT Scans" by Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., Shetty, S., and End-to-End Lung Cancer Screening Consortium. This paper discusses a deep learning-based approach for early lung cancer detection using low-dose chest CT scans.

2."Deep Learning for Lung Cancer Prognosis Prediction: A Retrospective Multi-Cohort Radiomics Study" by Huang, Y., Liu, Z., He, L., Chen, X., Pan, D., Ma, Z., and Liang, C. This study explores the use of deep learning and radiomics features for predicting the prognosis of lung cancer patients.

3."Early Prediction of Lung Cancer Incidence Using Deep Learning" by Akkus, Z., Ali, I., Sedlář, J., Agrawal, J. P., Parashar, A., and Bhagwat, N. This paper presents a deep learning-based approach for early prediction of lung cancer incidence using chest CT images.

**4.Title:** "Deep Learning for Predicting Lung Cancer Incidence from CT Imaging Data"

**Authors:** Ardila, Diego et al.

**Published in:** IEEE Transactions on Medical Imaging, vol. 38, no. 12, 2019, pp. 1344-1353.

**Abstract:** This paper presents a deep learning approach for predicting lung cancer incidence from computed tomography (CT) imaging data. They propose a novel deep learning architecture and demonstrate its performance on a large dataset of CT scans.

**5.Title:** "Lung cancer detection and classification with 3D convolutional neural networks"

**Authors:** Setio, Arnaud Arindra Adiyoso et al.

**Published in:** Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 834-842.

**Abstract:** This paper proposes a 3D convolutional neural network (CNN) approach for lung cancer detection and classification using CT scans. They demonstrate the effectiveness of their method on a dataset of CT scans.

**6.Title:** "Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network"

**Authors:** ten Haaf, Kevin et al.

**Published in:** PloS one, vol. 11, no. 12, 2016, e0163210.

**Abstract:** This study develops a dynamic Bayesian network model for predicting lung cancer incidence using data from the National Lung Screening Trial. They assess the performance of their model in predicting lung cancer incidence.

**7.Title:** "Predicting lung cancer incidence from computed tomography screening"

**Authors:** Katki, Hormuzd A et al.

**Published in:** Journal of the National Cancer Institute, vol. 106, no. 11, 2014.

**Abstract:** This paper presents a risk prediction model for lung cancer incidence based on computed tomography (CT) screening data. They develop and validate their model using data from the National Lung Screening Trial.