

**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  

---

**SINGAPORE**

**Group Project Report**

**BC2407**

**Analytics II**

**Done by: Team 5, Seminar Group 2**

**Ng Wan Yee**

**Nicholas Ting Jingjie**

**Natalie Teong Ying Er**

**Woon Hui En**

**Patel Dhairya Nayanbhai**

## Table of Contents

<b>Executive Summary.....</b>	<b>4</b>
<b>1.0 Business Problem.....</b>	<b>5</b>
1.1 Stakeholder Impacts.....	5
1.1.1 Patients.....	5
1.1.2 Hospitals.....	5
1.1.3 Government.....	6
<b>2.0 Costs of Problem.....</b>	<b>6</b>
2.1 Indirect Productivity Losses.....	6
2.2 Quality of Care Costs.....	6
<b>3.0 Analytical Problem.....</b>	<b>6</b>
3.1 Project Feasibility.....	7
3.1.1 Predictive Needs - Need for Advanced Diagnostic Methods.....	7
3.1.2 Technological Capabilities.....	7
3.1.3 Availability of Training Data.....	8
<b>4.0 Project Objective.....</b>	<b>8</b>
4.1 Clinical and Operational Outcomes.....	8
4.1.1 Detecting High-Risk Profiles.....	8
4.1.2 Advancing Preventative Care Measures.....	9
4.1.3 Leveraging Advanced Imaging Analysis for Diagnosis.....	9
<b>5.0 Insights from Convolution Neural Network.....</b>	<b>9</b>
<b>6.0 Insights from Random Forest.....</b>	<b>11</b>
<b>7.0 Insights from MARS.....</b>	<b>13</b>
<b>8.0 Insights from XGBoost.....</b>	<b>15</b>
<b>9.0 Insights for correlation matrix.....</b>	<b>17</b>
<b>10.0 Recommendations.....</b>	<b>18</b>
<b>10.1 Technical recommendations.....</b>	<b>18</b>
10.1.1 Association Rules in Lung Cancer Detection.....	18
10.1.2 XGBoost Model for Lung Cancer Detection.....	18
10.1.3 CNN Model for Enhanced Diagnostic Imaging in Lung Cancer Detection.....	19
<b>10.2 Operational Improvements for CNN Model Implementation in Lung Cancer Diagnostics.....</b>	<b>20</b>
10.2.1 Collaboration with Healthcare Institutions and Medical Imaging Centers.....	20
10.2.2 Association with Academic and Research Entities.....	20
10.2.3 Development of Rigorous Testing and Validation Frameworks.....	20
<b>10.3 Company recommendations.....</b>	<b>21</b>
10.3.1 Early Detection and Preventive Care.....	21
10.3.2 Improved Resource Allocation.....	21
10.3.3 Personalised Treatment Plans.....	21
<b>11.0 Post-Implementation: Monitoring Impact of Recommendations.....</b>	<b>21</b>
<b>12.0 Limitations of Model &amp; Concepts in Business context.....</b>	<b>22</b>
<b>12.1 Models.....</b>	<b>22</b>
12.1.1 Association Rules.....	22
12.1.1.1 Rules that lack logical association.....	22

12.1.2 XGBoost Model.....	22
12.1.2.1 High dependence on provided data.....	22
12.1.2.2 Interpretability of XGBoost model.....	23
12.1.3 CNN Model.....	23
12.1.3.1 Interpretability of CNN Model.....	23
<b>12.2 Integration of concepts into Business Context.....</b>	<b>24</b>
12.2.1 Integration with Existing Systems.....	24
12.2.2 Adoption and Trust.....	24
12.2.3 Data Privacy and Security.....	24
<b>13.0 Future Improvements &amp; Implementations.....</b>	<b>24</b>
13.1 Incorporation of additional data types into dataset.....	24
13.2 Expand AI application.....	25
<b>14.0 Conclusion.....</b>	<b>25</b>
APPENDICES.....	28
Appendix 1: Data Cleaning.....	28
Appendix 1.1 Data Cleaning for CNN.....	29
Appendix 2: Data Exploration.....	30
Appendix 3: Description for Logistic Regression Model.....	34
Appendix 4: Description for Random Forest Model.....	34
Appendix 5: Description for Multivariate Adaptive Regression Splines (MARS) Model...	36
Appendix 6: Description for XGBoost Regression Model.....	37
Appendix 6: Top 10 Association Rules From Apriori Algorithm.....	39
Appendix 7: Training and Validation Accuracy for CNN, Inception V3 and VGG16.....	40

## **Executive Summary**

This report seeks to understand the factors that contribute to lung cancer so as to find out the best methods we are able to utilise to identify individuals with high risks of lung cancer.

The majority of lung cancer cases are detected at advanced stages, significantly diminishing the effectiveness of treatments and survival rates. Traditional diagnostic methods, including manual pathology analysis and diagnosing techniques, are plagued by inefficiencies, subjective interpretations, and long diagnosis intervals, necessitating the exploration of more accurate and efficient detection strategies. This, combined with a global shortage of pathologists and increased lung cancer risk factors like smoking, vaping, and air pollution, underscores the need for innovative solutions.

To address this pressing issue, we utilised a dataset found from Kaggle, reflecting the intricacies of lung cancer diagnosis, to develop and assess predictive models utilising cutting-edge machine learning (ML) techniques. The project was initiated with meticulous data cleaning followed by an exploration phase using Tableau and Python. This phase aimed to distil a deeper understanding of the dataset and pinpoint influential variables associated with lung cancer risk.

Our analysis made use of 6 analytical models to aid in our analysis: Association Rules, Linear Regression, Random Forest Generator, XGBoost, Convolutional Neural Networks, and lastly, Multivariate Adaptive Regression Splines (MARS). We also made use of a dataset consisting of images to run our Convolutional Neural Networks. Each model was then evaluated for their diagnostic precision, accuracy and overall ability to detect and aid in the diagnosis of lung cancer.

Key findings from our models demonstrated promising capabilities in identifying lung cancer markers, with the CNN model showcasing exceptional accuracy in image analysis for detecting benign and malignant cases. As well as our XGBoost model excelling at prediction of the presence of Lung Cancer through personal and symptomatic indicators. Furthermore, the study highlights the potential operational benefits of integrating these models into clinical practice, including enhancing diagnostic workflows, facilitating early interventions, and optimising resource allocation.

Moreover, the report outlines strategic recommendations for hospitals to adopt these technologies, including forming partnerships with research institutions, and ensuring continuous model improvement to keep pace with medical advancements. While acknowledging limitations such as model interpretability and integration challenges, the report suggests future suggestions for investments in data interpretation tools and applying AI in prognosis prediction to refine lung cancer diagnostic processes further.

In conclusion, leveraging machine learning for lung cancer detection marks a critical advancement towards transforming healthcare diagnostics. By prioritising early detection and accuracy, the proposed models have the potential to significantly improve treatment outcomes and reduce the healthcare impact of lung cancer.

## **1.0 Business Problem**

Lung cancer is the second most common cancer worldwide (World Cancer Research Fund, 2022). According to the World Health Organisation (2023), lung cancer is the top cause of cancer death worldwide. In Singapore, it is the leading cause of cancer death for men and third-leading cause for women. A key problem is that most cases are found too late, in advanced stages, when the disease is harder to treat and chances of survival are much lower. Often, early signs of lung cancer are hard to detect, making early diagnosis difficult. In Singapore, about 75% of lung cancer cases are detected at advanced stages (stages 3 or 4).

Today, the detection of lung cancer mainly relies on manual Pathology Section Analysis. However, the low efficiency and subjective nature of manual film reading can lead to certain misdiagnoses and omissions. In addition, studies have shown misdiagnosis rates in pathology were at 70.1%, depending on the cancer type and other variables (T Wang, 1993). Misdiagnosis due to manual film reading in lung cancer pathology contributes to a huge portion of diagnostic errors (Mingsi.L, 2023).

Another common method of Lung Cancer detection is CT or chest X-rays. Despite being slightly more accurate than other methods, the time interval from chest CT or chest X-ray imaging to diagnosis, and from specialist consultation to diagnosis ranges from 43 and 72 days (Zigman Suchsland, M., 2022). As a result, every single month of delay in the diagnosis in many cases results in a 6 to 13% higher risk of dying (Hanna, T. P., 2020).

These diagnosis deficiencies could pose various challenges such as reduced treatment effectiveness, lowered survival rates, as well as increased healthcare costs to patients. From a hospital's perspective, it could also cause a strain in resources, especially when there is a high influx of patients.

Therefore, there is a pressing need for earlier, quicker and more accurate diagnosis to reduce these costs, as well as to effectively and efficiently treat patients.

## **1.1 Stakeholder Impacts**

### **1.1.1 Patients**

Late detection of Lung Cancer can lead to higher mortality rates for patients. In this regard, delayed diagnosis and detection of Lung Cancer only at its advanced stages would mean that treatment options are limited, more aggressive, and potentially less effective. Further, advanced stages Lung Cancer usually require more extensive and costly treatment such as Chemotherapy and Radiation Therapy (Kehong et al., 2023), which can lead to increased financial burdens, anxiety, depression and a compromised quality of life for the patient and their family members (Rotter et al., 2019).

### **1.1.2 Hospitals**

Hospitals would also be significantly impacted by the delayed detection of Lung Cancer because the higher prevalence of advanced-stages Lung Cancer could lead to a strain in medical resources, personnel, and facilities. This may lead to overcrowding and also

compromise on the quality of care that hospitals can provide to their patients. As a result, this could lead to more patients experiencing progressively worsening health, and making more return visits to the hospital in the short term (Sartini et al., 2022), which could in turn perpetuate a cycle of strain on healthcare resources and worsened patient outcomes.

Eventually, this could lead to reputational damages for hospitals if they are unable to provide effective and timely care for their patients.

### **1.1.3 Government**

Late diagnosis and higher prevalence of advanced-stage lung cancer could lead to increased government expenditure as well because they would need to subsidise a greater and more frequent amount of medical claims through government schemes given to citizens. The substantial costs associated with treating advanced-stage cancer, coupled with the growing demand for healthcare services, could strain the national healthcare budget and result in less flexibility when allocating budgets to other sectors of the country, impacting overall government spending priorities.

Further, there could be greater public health implications if high mortality rates eventually affect the overall country's health indicators and life expectancy. As a result of the distortion of health indicators from high mortality rates of advanced cancer patients, the economy may experience ripple effects as investor confidence, healthcare sector performance, productivity, and long-term economic growth decreases (Kaveh. G 2008).

## **2.0 Costs of Problem**

### **2.1 Indirect Productivity Losses**

Late detection of Lung Cancer, or detection of Lung Cancer only at advanced stages could lead to indirect economic and productivity losses for the country due to premature mortality and disabilities for the patients, as well as additional caregiving duties presented to their family members.

### **2.2 Quality of Care Costs**

When healthcare resources are stretched thin due to the influx of advanced Cancer patients, there is a risk of decreased healthcare quality for all patients in the hospital. This may be in the form of longer wait times for appointments, reduced time with healthcare providers, and delays in receiving test results and treatment. As a result of lower quality care, healthcare providers may miss out on opportunities for early diagnosis, treatment, and management of conditions for patients whose conditions have not progressed to a severe stage yet.

## **3.0 Analytical Problem**

From an analytical perspective, our project is focused on developing a predictive model that can accurately diagnose lung cancer at early stages using machine learning techniques. Lung cancer, being the leading cause of cancer death globally, presents significant challenges in early detection due to the subtlety of its early symptoms and the limitations of current diagnostic methods. Our model aims to improve the precision and efficiency of lung cancer diagnosis, thereby enabling more effective treatment and potentially reducing the

socio-economic costs associated with the disease. By leveraging machine learning algorithms, we aim to overcome the high false-positive rates and subjective biases inherent in manual film reading and existing diagnostic procedures.

### **3.1 Project Feasibility**

The feasibility of our project is supported by three critical aspects: the need for advanced diagnostic methods, the technological capabilities for machine learning in healthcare, and the availability of relevant data for model training.

#### **3.1.1 Predictive Needs - Need for Advanced Diagnostic Methods**

The current landscape of lung cancer diagnosis presents several challenges that necessitate the development of advanced diagnostic methods. Firstly, the traditional diagnostic process, which heavily relies on manual pathology analysis and imaging techniques, suffers from inefficiencies such as high false-positive rates and subjective interpretation errors. These limitations can lead to delayed or inaccurate diagnoses, significantly impacting patient outcomes, particularly in the case of a disease as aggressive as lung cancer (Shaughnessy, 2017)

Moreover, lung cancer symptoms are often subtle and easily overlooked in the early stages, resulting in a majority of cases being diagnosed at an advanced stage when treatment options are limited and less effective (TimesofIndia, 2023). The socio-economic costs associated with late-stage lung cancer diagnosis are immense, ranging from increased healthcare expenditure to significant emotional and financial burdens on patients and their families (Kutikova et al., n.d.).

The need for advanced diagnostic methods is clear as the tools that can accurately and efficiently detect lung cancer at earlier stages are crucial for improving survival rates, reducing the economic impact of the disease, and ultimately enhancing the quality of life for patients. With their ability to learn from vast amounts of data and identify patterns that may not be visible to the human eye, machine learning models offer a promising solution to these challenges.

#### **3.1.2 Technological Capabilities**

Recent advancements in machine learning (ML) and artificial intelligence (AI) technologies have paved the way for revolutionary changes in healthcare diagnostics. Specifically, ML algorithms such as Convolutional Neural Networks (CNNs) for image analysis and classification algorithms like XGBoost and Random Forest for analysing symptomatic and demographic data have shown great potential in detecting complex patterns and anomalies that are indicative of diseases like lung cancer (Mokoatle et al., 2023).

These technological capabilities enable the processing and analysis of large and complex datasets, including medical images and patient records, to generate accurate and reliable diagnostic predictions. Furthermore, ML models can continuously improve as they are exposed to more data, ensuring that the diagnostic process becomes increasingly precise over

time. The integration of ML into healthcare diagnostics promises to enhance the accuracy of lung cancer detection and reduce the time required for diagnosis, allowing for earlier intervention and better patient outcomes.

### **3.1.3 Availability of Training Data**

For developing our predictive model, we will utilise publicly available datasets from Kaggle, which contain detailed information on lung cancer cases, including imaging data and patient demographics. These datasets provide a rich source of information for training and validating our machine-learning algorithms. The comprehensiveness and diversity of these datasets are critical for building a robust model capable of generalising across different populations and improving the detection of early-stage lung cancer.

By addressing these feasibility aspects, our project aims to deliver a machine learning solution that can significantly enhance the accuracy and efficiency of lung cancer diagnosis, contributing to better patient outcomes and reduced healthcare costs.

## **4.0 Project Objective**

This project is dedicated to the development of a machine learning-based analytical model focused on the early detection, precise diagnosis, and efficient management of lung cancer. Our primary goal is to elevate diagnostic accuracy and operational efficiency for healthcare providers, facilitating prompt and effective treatment interventions. Utilising advanced predictive analytics, our aim is to accurately detect lung cancer in its initial stages, thereby diminishing the delay in treatment commencement. This approach is anticipated to significantly lower the mortality and morbidity rates stemming from delayed diagnoses.

### **4.1 Clinical and Operational Outcomes**

Deploying this model is anticipated to yield significant clinical and operational benefits. The first is the refinement of lung cancer diagnosis precision and the streamlining of the diagnostic process, which together are expected to enhance patient treatment outcomes substantially. Secondly, by improving diagnostic accuracy and efficiency, we envisage fostering a positive reputation for healthcare providers and reducing long-term operational costs. Lastly, we will monitor the performance of our Machine Learning model by tracking key metrics, and indicators to evaluate its accuracy by having a robust classification model. For a good model rating, its F1-score should be higher than 70% , precision score of 9.5% and recall score of 70% - 80% (Purva. H, 2023).

#### **4.1.1 Detecting High-Risk Profiles**

Our model meticulously identifies individuals at an elevated risk of developing lung cancer. The model can accurately reveal early-stage lung cancer markers by analysing complex data patterns and pinpointing key diagnostic indicators. This early detection is pivotal, as lung cancer's subtle initial symptoms often lead to delayed diagnoses when treatment options are less effective (Hansen et al., 2011). The model is an advanced tool for healthcare providers to quickly recognise at-risk patients, ensuring they receive timely and adequate medical attention.



#### **4.1.2 Advancing Preventative Care Measures**

With insights derived from our analytical model, healthcare providers can recognise patterns and signs that may precede the development of lung cancer. These insights are crucial for devising targeted screening programs and preventative interventions, tailored to those at an increased risk. The application of such preventative measures aims to mitigate the progression of lung cancer, enhance the health outcomes for patients, and ultimately alleviate the healthcare system's burden by reducing the incidence of advanced-stage lung cancer cases.

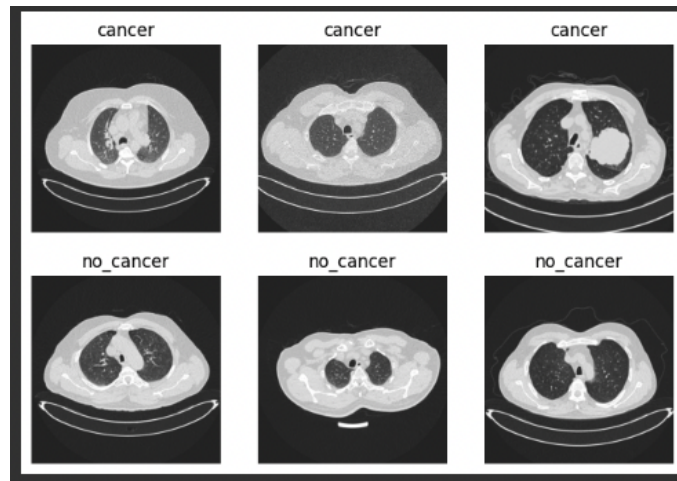
#### **4.1.3 Leveraging Advanced Imaging Analysis for Diagnosis**

Adding to our comprehensive approach, we utilise a Convolutional Neural Network (CNN) model, developed with TensorFlow, to perform detailed chest X-ray and CT-scan image analyses. This model determines the presence of lung cancer through a binary outcome (Yes/No) and, upon detection, further evaluates whether the cancer is benign or malignant. This dual-phase diagnostic process not only refines diagnosis but also provides a nuanced understanding of the patient's condition, enabling more personalised treatment plans. The effectiveness and accuracy of this model, in conjunction with three others devised for this project, will undergo these thorough evaluations to identify the most suitable option for lung cancer prediction. This comparative analysis is vital to our commitment to utilising the most effective diagnostic tools available, ensuring our project's goal of enhancing lung cancer diagnosis and treatment outcomes is met.

#### **5.0 Insights from Convolution Neural Network**

Lung cancer dataset was collected over a period of 3 months. It comprises CT scans of patients diagnosed with lung cancer across various stages, as well as scans from healthy subjects. The dataset has a total of 1,190 CT scan images from 110 cases. These cases are categorised into three categories: normal, benign and malignant. Out of the 110 cases, 40 are diagnosed as malignant, 15 as benign and 55 as normal.

Initially, our code was modified to display sample images for each category - "Benign", "Malignant" and "Normal". To simplify the interpretation and enhance clarity regarding the presence or absence of cancer, the dataset has been reclassified with binary labels "yes" and "no\_cancer" fig 1.1 and fig 1.2.



**Fig. 1.1 - Showing images categorised into cancer and no\_cancer**

	img_path	label
0	/content/drive/MyDrive/Anal 2 CNN project/Beni...	no_cancer
1	/content/drive/MyDrive/Anal 2 CNN project/Beni...	no_cancer
2	/content/drive/MyDrive/Anal 2 CNN project/Beni...	no_cancer
3	/content/drive/MyDrive/Anal 2 CNN project/Beni...	no_cancer
4	/content/drive/MyDrive/Anal 2 CNN project/Beni...	no_cancer

**Fig. 1.2 - Showing path label no\_cancer**

Our dataset is split into three subsets: training, testing and validation. The split ratios are as follows: **Training set:** 70% of the data ensures a substantial amount of data for training the model. **Testing set:** 15% of the data is used to evaluate the model's performance and generalisation capability. **Validation set:** 15% of the data is used to fine-tune the model and avoid overfitting.

After splitting our data into a train vs test set, we have evaluated the 2 different classes' weights. With **cancer: 0.55822** indicates that the weight assigned to the "cancer" class is 0.56. Since "cancer" class is less frequent in the training data, it gets a higher weight to account for lower representation, making the model pay more attention to this class during training. As this is in place to account for the imbalance nature of the dataset, where the "cancer" class is less frequent compared to the "no\_cancer" class.

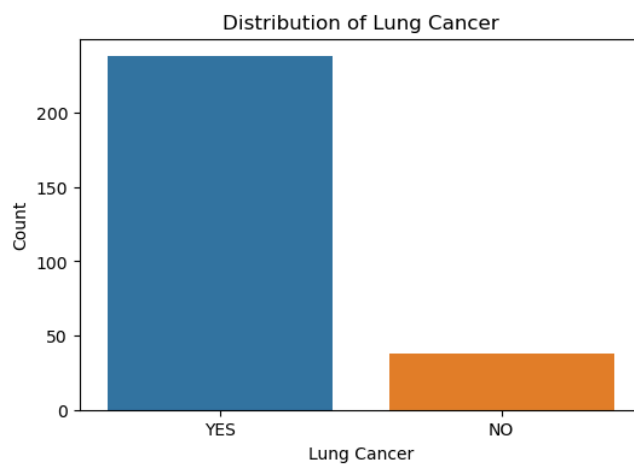
We looked at the performance of the 3 image classification models on the lung cancer test datasets; CNN, Inception V3 and VGG16. All three achieved good accuracy, with test accuracies ranging from 81.8% to 86.7%. This suggests that they all learned the patterns in the training data. VGG16 achieved a good balance between test loss around 0.3 and test accuracy around 86%.

Thus, using this dataset for lung cancer along with the model prediction accuracy and loss. Could help hospitals make more informed and quicker decisions such as medical imaging

analysis. It is faster and forms more objective analysis, analysing medical images like CT scans, X-rays can be time-consuming for human radiologists (Mukesh. S, 2022). CNN models can analyse these images faster and reduce turnaround time for diagnosis. It also helps to reduce the possibility of human error or bias. On top of that, it also helps with early interventions and warning signs to provide better treatment plans for hospitals. As CNN models are developed to identify subtle changes in medical images that might indicate the early stages of disease (Mengfang. L, 2023). It helps to provide real-time decision making to doctors during diagnosis and treatment planning.

## **6.0 Insights from Random Forest**

Predictive modelling techniques were applied to a dataset containing information on individuals' gender, age, physical and psychological symptoms, and lifestyle factors. The primary objective was to develop models capable of accurately identifying positive lung cancer cases based on these input variables.



**Fig. 1.3 - Bar plot showing the imbalance in dataset**

The dataset is imbalanced, containing 238 positive cases and 38 negative cases after removing duplicates, as shown in Fig. 1.3. The training set, comprising 80% of the total data, was used to train and optimise the models, while the remaining 20% of the data formed the testing set, which was used to evaluate the models' performance on unseen data. Stratified sampling was used during the data splitting process to maintain the same ratio of positive to negative cases in both the training and testing subsets. Stratification ensures that the class imbalance present in the original dataset is preserved, mitigating potential bias and providing a more accurate assessment of the models' performance in real-world scenarios.

The Random Forest model demonstrates strong performance with an accuracy of approximately 91%. In diagnosing lung cancer, incorrectly classifying a positive case as negative can have severe consequences, such as delayed treatment and potentially life-threatening outcomes. The true positive rate, also known as recall, is a critical metric in this context. It measures the proportion of actual positive cases that the model accurately identifies as positive. In other words, recall quantifies the model's effectiveness in detecting lung cancer when it is indeed present.

Given the high stakes involved in diagnosing lung cancer, maximising the recall rate should be the primary focus. With a recall of 95.83%, the Random Forest model is excellent at correctly classifying positive cases, which is crucial for a condition where early detection can significantly improve survival rates.

Additionally, F1-score is another important metric to consider, given the imbalance in the dataset used. The model's F1-score of approximately 94.84% indicates a good balance between precision and recall, suggesting that the model performs well in both correctly identifying a large proportion of lung cancer cases while minimising false positives.

Precision is a measure of the proportion of the cases that the model predicts as positive, that are actually positive. This high precision is valuable in a clinical setting, as it minimises false positives. False positives, where the model incorrectly predicts lung cancer in patients who do not actually have the disease, can lead to unnecessary anxiety for patients and a waste of limited healthcare resources, such as further diagnostic tests and specialist consultations. By maintaining a high precision, the model minimises these false positives, allowing healthcare professionals to allocate resources more effectively to patients who genuinely require attention and treatment.

A mean cross-validation score of approximately 89.51% when using 10-fold cross validation suggests that the model is stable and reliable across different subsets of data. This reliability is important for confidence in the model's generalisability to new patient data and to ensure there is no overfitting to a subset of the data.

```
low_risk_threshold = 0.3  
medium_risk_threshold = 0.6
```

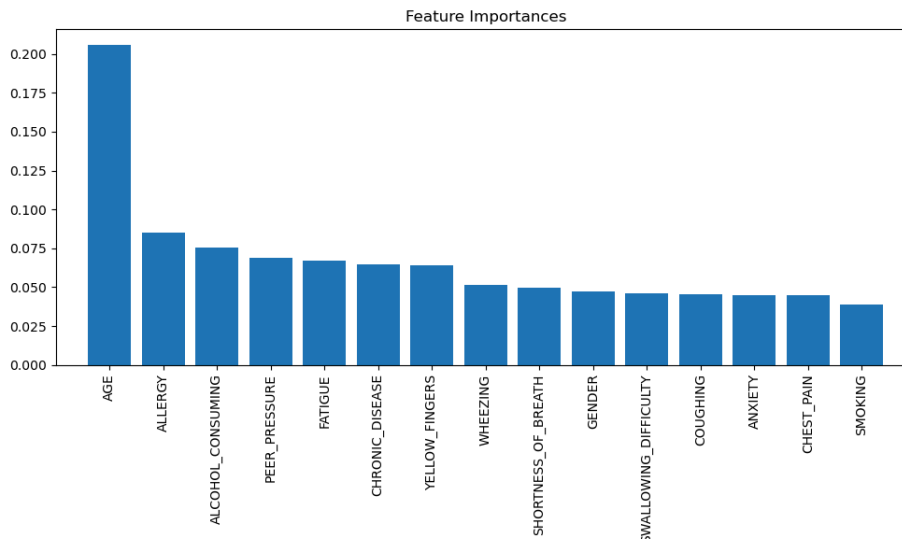
#### **Random Forest Risk Scores:**

Patient 1: Probability = 0.985, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 2: Probability = 0.99, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 3: Probability = 1.0, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 4: Probability = 1.0, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 5: Probability = 0.455, Risk Category = Medium Risk, Actual Lung Cancer = 0

**Fig. 1.4 - Demonstration of categorisation of patients into risk categories based on predicted probabilities using Random Forest**

The Random Forest model effectively separates patients into low, medium and high-risk categories based on predicted probabilities, as shown above. Cases with a predicted probability of lower than 0.3 are classified as low-risk, while those with a predicted probability of between 0.3 and 0.6 are classified as medium-risk. Those with a predicted probability 0.6 and above are classified as high-risk.

High-risk patients can be flagged for immediate attention, while medium-risk patients can be scheduled for regular monitoring, optimising the use of healthcare resources. With the model's ability to identify high-risk patients, there's a significant opportunity for early intervention. This could facilitate earlier treatments, potentially leading to better outcomes and higher survival rates for patients diagnosed with lung cancer.



**Fig. 1.4 - Bar plot of relative importance of variables in Random Forest Model**

In the context of lung cancer diagnosis, identifying the top variables contributing to model decisions is critical as shown in Fig. 1.4. These may include demographic factors, patient history, genetic predispositions, and specific markers identified from diagnostic tests. While the Random Forest model provides high accuracy, combining its output with clinical expertise could lead to more comprehensive diagnostic insights. Healthcare professionals could use the model's predictions as a decision support tool to identify cases requiring further analysis.

As more patient data become available, the Random Forest model can be retrained to incorporate new insights, potentially discovering novel patterns and associations relevant to lung cancer detection. For clinical acceptance, it's important to make the Random Forest model as transparent as possible. Techniques such as feature importance graphs can help clinicians understand how the model is making its predictions.

## **7.0 Insights from MARS**

The MARS model shows an accuracy and precision comparable to the Random Forest model, with an accuracy of approximately 91% and precision around 92%. It also has an F1-score of 95%. The exceptionally high recall of approximately 97.92% demonstrates the model's ability to correctly identify most of the actual lung cancer cases. This is crucial in a clinical context, as missing a positive case can have serious consequences.

With a mean cross-validation ROC\_AUC (Area Under the Receiver Operating Characteristic Curve) score of around 94.03%, the MARS model demonstrates a reliable performance across multiple data subsets, which suggests good generalisability.

#### MARS Model Summary

Basis Function	Pruned	Coefficient
(Intercept)	No	0.190668
ALLERGY	No	0.138624
SWALLOWING_DIFFICULTY	No	0.109499
COUGHING	No	0.098869
PEER_PRESSURE	No	0.0827809
ALCOHOL_CONSUMING	No	0.204596
FATIGUE	No	0.191978
CHRONIC_DISEASE	No	0.158828
YELLOW_FINGERS	No	0.148101
SMOKING	No	0.0796479
h(AGE-77)	No	-0.040999
h(77-AGE)	Yes	None
ANXIETY	Yes	None
CHEST_PAIN	Yes	None
WHEEZING	Yes	None
SHORTNESS_OF_BREATH	Yes	None
GENDER	Yes	None

MSE: 0.0726, GCV: 0.0824, RSQ: 0.3835, GRSQ: 0.3066

The MARS model also provides clear interpretability through basis functions and their coefficients, which can be crucial for understanding the model's decision-making process in clinical settings. The identification of important variables like "ALLERGY", "SWALLOWING\_DIFFICULTY", "COUGHING", "PEER\_PRESSURE", etc., can inform healthcare providers about significant risk factors and symptoms to consider when diagnosing lung cancer.

On top of that, the F1-score of 94.91% also further proves the model's ability to identify most of the lung cancer cases, which is important in clinical context to avoid missing positive cases (Teemu. K, 2020).

Patient 1: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 2: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 3: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 4: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 5: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 0

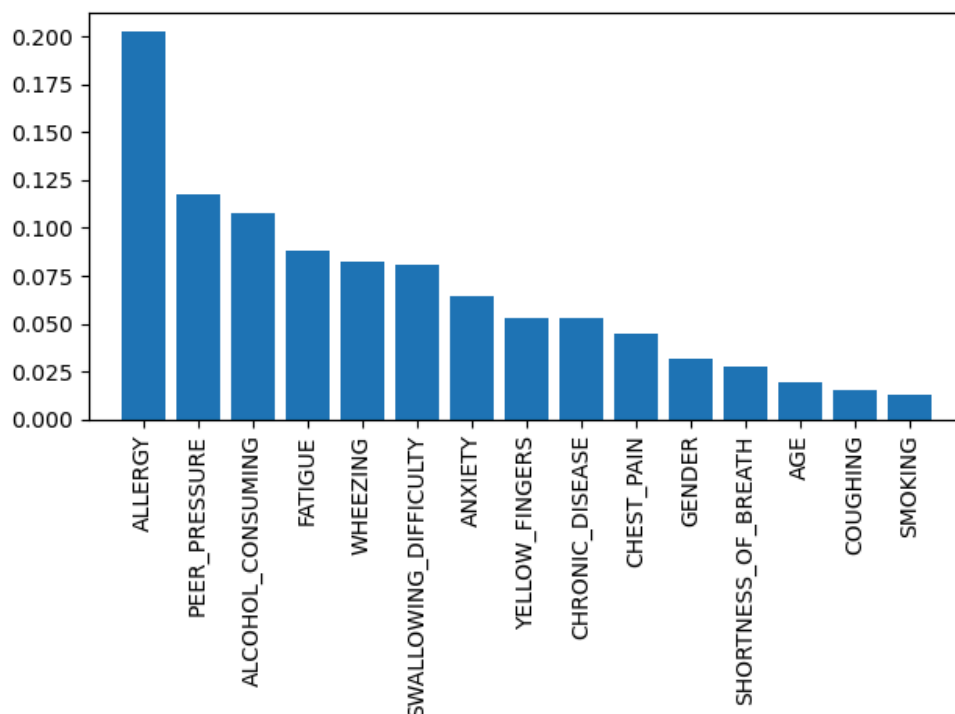
**Fig. 1.5 - Demonstration of categorisation of patients into risk categories based on predicted probabilities using MARS**

The model effectively stratifies lung cancer risk, allowing healthcare providers to categorise patients into distinct risk categories, which is important for prioritising cases and planning treatment. The presence of pruning in the MARS model indicates its potential for scalability and adaptability, allowing for the adjustment of the model complexity based on the available data. The model's ability to incorporate various patient-specific variables suggests potential for personalised risk assessment and treatment planning. By identifying high-risk patients, the MARS model can facilitate earlier interventions, which is crucial for diseases like lung cancer where early detection can significantly affect the prognosis.

### **8.0 Insights from XGBoost**

The XGBoost model demonstrates a high recall rate of approximately 97.92%, which is crucial for lung cancer diagnosis where missing out on true positives can be particularly detrimental. An accuracy of about 89.29% indicates that the model is adept at classifying cases correctly. It also has an F1-score of 95%, indicating that the model maintains a strong balance between precision and recall.

With a mean cross-validation accuracy of approximately 88.06%, the XGBoost model exhibits robustness, which suggests it is not overly sensitive to specific features or outliers in the training data and can potentially generalise well to other datasets.



**Fig. 1.6 - Bar plot of relative importance of variables in XGBoost Model**

The model also provides insights into feature importance, highlighting the factors that contribute most significantly to the predictions. In this case, important features such as "ALLERGY" and "PEER\_PRESSURE" have been identified as primary indicators of lung cancer risk, as shown in Fig. XX. This interpretability aspect of the XGBoost model is particularly valuable for clinicians, as it can guide them in understanding the key risk factors and inform their decision-making process.

Known for its efficiency in handling large datasets, the XGBoost model can quickly process and analyse extensive lung cancer data, making it suitable for clinical settings where time is critical. Its computational efficiency and predictive performance make the XGBoost model a potentially effective tool for increasing operational efficacy in lung cancer screening programs.

Patient 1: Probability = 0.9775846600532532, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 2: Probability = 0.9726433157920837, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 3: Probability = 0.9427878856658936, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 4: Probability = 0.9933772683143616, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 5: Probability = 0.8073779344558716, Risk Category = High Risk, Actual Lung Cancer = 0

**Fig. 1.7 - Demonstration of categorisation of patients into risk categories based on predicted probabilities using XGBoost**

The XGBoost model provides probabilistic outcomes, which can be used to assign a risk score to each patient, similar to the other models (see Fig. 1.7)



## 9.0 Insights for correlation matrix

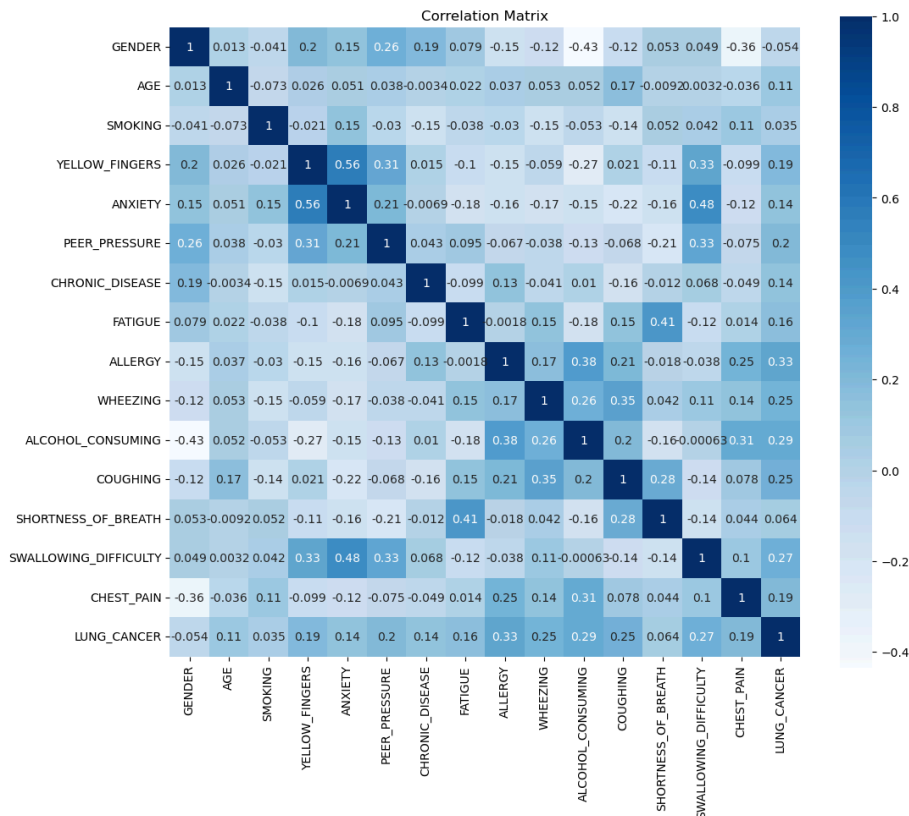


Fig. 1.8 - Correlation Matrix between variables to identify their relationship

We have plotted a heatmap of the correlation between variables to help us identify their relationship, which then helps us decide on the data we should feed to predict our model. After analysing the correlation matrix, our results show that **YELLOW\_FINGERS with ANXIETY** has a moderate correlation at 0.56 which might show underlying factors associated with lung cancer risk such as smoking, which can cause yellowing of fingers and induced anxiety. Apart from that, to identify the correlation between smoking and its variables we have identified that there is a negative correlation between **AGE and SMOKING** even though it is weak it still suggest that age, as represented in this dataset, does not strongly predict smoking status, which is a primary risk factor for lung cancer. Apart from that, there is a positive correlation between respiratory symptoms and lung cancer such as (**WHEEZING, COUGHING, and SHORTNESS OF BREATH**). It is relevant since they can be symptoms of lung cancer (Claudia. B, 2013). The lifestyle factor that has correlation is between **PEER\_PRESSURE and ALCOHOL\_CONSUMING** it might be an indication of social lifestyle factors that could correlate with smoking habits. Such as, if someone experiences peer pressure to drink alcohol, they might be in a social group where smoking is common (Alberto. V, 2018).

## **10.0 Recommendations**

### **10.1 Technical recommendations**

#### **10.1.1 Association Rules in Lung Cancer Detection**

Association rules can be instrumental in healthcare settings, particularly in understanding complex relationships between various risk factors and lung cancer incidence. In the context of lung cancer detection, our analysis employing the Apriori algorithm has revealed strong associations between certain lifestyle factors, environmental exposures, and the diagnosis of lung cancer. For instance, if we consider that a particular combination of symptoms and exposure to specific pollutants is frequently present in lung cancer patients, these patterns can guide medical professionals in early screening and diagnosis.

Detecting lung cancer poses significant challenges due to the often subtle and non-specific nature of its symptoms. Unlike conditions with more pronounced and easily identifiable signs, lung cancer's indicators can be easily overlooked or mistaken for less severe illnesses (Cellina et al., 2023). This complexity requires a nuanced approach to diagnosis, moving beyond static criteria to more adaptive, data-driven models that can interpret the subtle nuances of symptom presentation and variability among individuals.

By running association rule analysis periodically on updated datasets, including patient symptoms, demographics, and environmental data, healthcare providers can ensure they consider the most current and relevant factors when assessing lung cancer risk. This adaptability allows for more precise and timely interventions. With insights from the latest data, clinicians can prioritise high-risk patients for further diagnostic tests, and policymakers can design more effective public health initiatives to combat lung cancer.

#### **10.1.2 XGBoost Model for Lung Cancer Detection**

The adoption of the XGBoost machine learning model for lung cancer detection is proposed, leveraging the model's proven success in various medical fields. Given its effectiveness in predicting outcomes for Acute Ischemic Stroke (AIS) and its prowess in diabetes detection through breath analysis, XGBoost is a versatile tool capable of interpreting complex medical data to provide actionable insights (Chung et al., 2023).

#### **Challenges in Current Detection Frameworks:**

The reliance on medical professionals' subjective judgement and the resource-intensive nature of lung cancer screening underscore the need for more efficient, data-driven approaches. Traditional methods are often constrained by diagnostic inaccuracies and inefficiencies, underscoring the urgency for innovation.

#### **Advantages of XGBoost in Lung Cancer Detection**

##### **1. Operational Efficiency and Enhanced Detection:**

XGBoost addresses these challenges by automating risk assessments, thus freeing up valuable resources and time for direct patient care. This level of automation fosters more streamlined patient management and can significantly improve healthcare

outcomes. By processing complex datasets rapidly, XGBoost excels in identifying early-stage lung cancer patterns that might be overlooked with conventional diagnostics, thereby enhancing detection capabilities.

2. Accuracy, Precision, and Data-Driven Insights:

A key advantage of XGBoost lies in its high predictive accuracy and precision, which are crucial for reducing false negatives and positives and ensuring timely and appropriate patient care. The model's ability to uncover subtle patterns in patient data enables the discovery of new diagnostic markers, enhancing lung cancer diagnosis.

3. Adaptive Learning for Continuous Improvement:

The model's adaptive learning capability ensures it remains aligned with the latest medical insights, evolving alongside new patient data. This adaptability is critical for maintaining the relevance and effectiveness of lung cancer detection methods.

### **Real-World Applications and Efficacy:**

XGBoost's proven efficacy in other medical domains, such as acute ischemic stroke (AIS) diagnosis and diabetes detection, lays a solid foundation for its application in lung cancer detection, suggesting potential for significant improvements in early detection rates and patient outcomes (Chung et al., 2023; Paleczek et al., 2021). Its successful integration into lung cancer detection processes can be a significant step towards a more data-driven, patient-centric approach in healthcare diagnostics, ultimately enhancing operational efficiency and improving healthcare delivery.

#### **10.1.3 CNN Model for Enhanced Diagnostic Imaging in Lung Cancer Detection**

Currently, the process of diagnosing lung cancer relies heavily on the expertise of radiologists who review and interpret medical images. While this method has been the cornerstone of lung cancer diagnosis, it comes with certain limitations:

1. Subjectivity in Interpretation: Radiologists' assessments can vary, leading to differing interpretations of imaging studies, which could result in inconsistencies in early detection rates (Onder et al., 2021).
2. Volume of Imaging Data: With the increasing number of imaging studies being performed, it is becoming increasingly challenging for radiologists to meticulously review every case, leading to potential delays in diagnosis and treatment (Kwee & Kwee, 2021).
3. Complexity of Lung Cancer Presentation: Lung cancer can present with subtle changes in early stages that may be difficult to discern, thus requiring a level of detail that sometimes surpasses human detection capabilities (Rampinelli et al., 2016).

To mitigate these issues, we recommend implementing a Convolutional Neural Network (CNN) model designed specifically for lung cancer detection from imaging data. This advanced model aims to complement the radiologists' work by pre-screening images to highlight areas of concern and provide preliminary risk assessments, thereby enhancing the diagnostic process.

### **Benefits of CNN Model Integration:**

1. Consistency and Accuracy: Trained on extensive datasets, CNN models excel in identifying complex patterns within lung imaging data, significantly reducing the subjectivity of human interpretation. This leads to more objective and reliable diagnostics, offering consistent and quantifiable metrics that improve the precision and accuracy of lung cancer detection.
2. Efficiency in Reviewing Cases: By automatically analysing images and flagging areas for further review, CNNs enable radiologists to prioritise their focus on cases with potential concerns, streamlining the diagnostic workflow and reducing the time to diagnosis.
3. Early Detection and Deep Learning Insights: The CNN model's ability to detect subtle features indicative of early-stage lung cancer is crucial for improving prognosis through timely intervention. Its deep learning capabilities also unearth complex patterns and relationships in imaging data, providing valuable insights into lung cancer characteristics.
4. Dynamic Learning and Scalability: As the model encounters new data, it continuously learns and improves, ensuring its approaches remain current with the latest imaging technologies and medical knowledge. The scalability of CNN models means that once validated, they can be deployed across various healthcare settings, potentially democratising access to sophisticated diagnostic tools, especially in resource-constrained environments.

By integrating the CNN model into the lung cancer diagnostic workflow, we can significantly address current limitations, enhancing diagnostic efficiency, accuracy, and ultimately, patient outcomes. This approach marks a pivotal shift towards leveraging AI to support and augment the capabilities of radiologists in lung cancer detection.

## **10.2 Operational Improvements for CNN Model Implementation in Lung Cancer Diagnostics**

### **10.2.1 Collaboration with Healthcare Institutions and Medical Imaging Centers**

By forming partnerships with healthcare institutions and medical imaging centres, the project can leverage a large dataset for training the CNN model, which is essential for its accuracy and reliability. A robust dataset ensures that the model is exposed to various manifestations of lung cancer, which is crucial for its ability to generalise across unseen cases. Such a partnership can be modelled after successful collaborations like the one between Google and Northwestern scientists, which improved lung cancer detection through AI (Paul, 2019).

### **10.2.2 Association with Academic and Research Entities**

Engagement with academia can fuel innovation in CNN models, continuously drawing on cutting-edge research to refine algorithms. Academic partnerships can also facilitate peer-reviewed studies to validate the efficacy of the CNN model, similar to those published in prominent journals such as "Nature Medicine" (Ardila et al., 2019), highlighting the use of AI in medical imaging for cancer detection.

### **10.2.3 Development of Rigorous Testing and Validation Frameworks**

Developing rigorous testing and validation frameworks for the CNN model is crucial, akin to benchmarking diagnostic tools against pathology reports to ensure AI-driven diagnostics' accuracy and reliability. Continuous and comprehensive validation processes are necessary to maintain the model's effectiveness over time, adapting to new data and evolving standards in medical diagnostics (Hosny et al., 2018).

## **10.3 Company recommendations**

### **10.3.1 Early Detection and Preventive Care**

**Insight:** The CNN and XGBoost models demonstrate high accuracy in detecting lung cancer from CT scans and patient data, enabling earlier diagnosis and intervention.

**Recommendation:** Implement these models as part of a comprehensive lung cancer screening program, targeting high-risk individuals such as heavy smokers and those with a family history of lung cancer. By proactively screening and identifying cases at an early stage, healthcare providers can initiate timely treatment and improve patient outcomes, contributing to the objective of increasing the proportion of lung cancer cases detected at Stage 1 or 2 from the current 15-25% to 40% within one year of implementing the CNN and XGBoost models (Daniela. A 2023).

### **10.3.2 Improved Resource Allocation**

**Insight:** The predictive models can help prioritise cases based on the likelihood of lung cancer, allowing healthcare providers to allocate resources more effectively.

**Recommendation:** Integrate the model outputs into the clinical decision-making process, using the risk scores to triage patients for further diagnostic tests and specialist consultations. This will help ensure that patients with the highest risk receive prompt attention and care, optimising the use of limited healthcare resources. By focusing resources on high-risk patients, healthcare providers can contribute to achieving the objective of increasing the participation rate in lung cancer screening programs for smokers aged 55-80 from the current 43.7% to 50% within 18 months of implementing the risk prediction models (Di Liang, 2021).

### **10.3.3 Personalised Treatment Plans**

**Insight:** The models provide insights into the factors contributing to an individual's lung cancer risk, enabling healthcare providers to develop tailored treatment plans.

**Recommendation:** Use the model insights to create personalised treatment plans that consider each patient's specific risk factors, such as smoking history, age, and comorbidities. By adopting a more targeted approach to treatment, healthcare providers can improve the effectiveness of interventions and enhance patient outcomes. This personalised approach can contribute to achieving the objective of a 90% accuracy rate in identifying lung cancer from CT scans and patient data, with a false positive rate below 5% and a false negative rate below 3%, as the models will help guide treatment decisions based on individual risk profiles.

## **11.0 Post-Implementation: Monitoring Impact of Recommendations**

After implementing the recommended technical strategies focused on leveraging advanced data analytics and machine learning models like XGBoost and CNN for lung cancer

detection, it is crucial to assess the impact of these innovations on diagnostic processes. This involves monitoring key performance indicators such as:

1. Detection Rates: A pivotal measure will be evaluating early detection rates of lung cancer, comparing pre-implementation data with post-implementation outcomes. Specifically, we aim to define a target increase in the percentage of lung cancer cases detected at an early stage after implementing the predictive models. For example, we seek to "Increase the proportion of lung cancer cases detected at Stage 1 or 2 from the current 22.2% to 40% within one year of implementing the CNN and XGboost models (Paul. F, 2018)". This target underscores the importance of early detection in improving patient prognosis and the potential of our chosen technical strategies to make a significant impact.
2. Diagnostic Accuracy: Beyond the detection rates, measuring the accuracy and precision of lung cancer diagnoses is essential to ensure reductions in false positives and negatives. This goal highlights our commitment to increasing lung cancer detection at early stages and ensuring that our diagnostic processes minimise the chance of misdiagnosis, optimising patient care and resource allocation.
3. Operational Efficiency: Assessing improvements in the efficiency of diagnostic workflows and the allocation of healthcare resources is another critical factor. One tangible target is to set goals for increasing participation in lung cancer screening programs among high-risk populations. These goals highlight operational efficiency's role in maximising the reach and impact of lung cancer screening efforts.
4. Health Outcomes: Finally, monitoring patient outcomes, including survival rates and quality of life, will provide a comprehensive view of the impact of earlier and more accurate diagnoses. This holistic approach ensures that the benefits of technological advancements translate into tangible improvements in patient care and healthcare delivery.

## **12.0 Limitations of Model & Concepts in Business context**

### **12.1 Models**

#### **12.1.1 Association Rules**

##### **12.1.1.1 Rules that lack logical association**

Association rules might detect correlations that are not logically associated with lung cancer, and this may pose a challenge for users in terms of comprehension and extracting insights from them. This can stem from the lack of context of the model and subjective nature of interpretability, which the machine learning algorithm may not adequately account for (García et al., 2007).

In order to mitigate this limitation, the hospital can consider hiring subject matter experts specialising in Lung Cancer, ensuring that only pertinent results and associations are extracted.

### **12.1.2 XGBoost Model**

#### **12.1.2.1 High dependence on provided data**

As a supervised Machine Learning technique, the predictive capabilities of the XGBoost model are contingent upon the characteristics of lung cancer previously detected. As a result, the model may fail to accurately predict cases with unusual or rare characteristics that were not present in the training data, which may eventually lead to misclassifications and late diagnosis in the hospital setting.

That said, this challenge can be gradually improved on as the model accumulates more Lung Cancer symptoms and data over time. By integrating these new data and re-training the existing model, the model's ability to analyse a broader range of Lung Cancer characteristics improve. Such iterative processes highlights and enhances the model's inclusivity, enabling it to identify previously undetected Lung Cancer cases that were overlooked by other Lung Cancer detection mechanisms.

As a result of these enhancements, the occurrence of erroneous lung cancer diagnoses is expected to decrease over time.

#### **12.1.2.2 Interpretability of XGBoost model**

While XGBoost models excel at predicting outcome variables, the hospital may find its results hard to interpret, especially if they possess knowledge of Machine Learning concepts. Further, since the model deals with a fairly large number of variables, it may capture several intricate relationships that make the resulting statistics difficult to read.

In order to address this, the hospital can consider investing in post-implementation integration techniques. Some instances of these techniques can include employing visualisation tools like plots and dashboards to simplify the results of the XGBoost model's output. Additionally, the hospital can also look into providing basic training or learning materials to staff whose job rely heavily on the results of these data, so that they can understand and leverage on the insights of these data more effectively.

### **12.1.3 CNN Model**

#### **12.1.3.1 Interpretability of CNN Model**

CNN Models are often criticised for their black-box nature, as it can be challenging to interpret and understand how it arrived at a certain conclusion. For example, if a CNN model concludes that an X-ray indicates presence of Lung Cancer, we will not be able to tell which part, or parts of the image or X-ray helped the model to arrive at that result. In a clinical setting where transparency and trust are paramount, the lack of interpretability may be a significant barrier to adoption, and staff may be less comfortable in using such models.

In order to mitigate this, the hospital can employ interpretability techniques such as class activation maps, saliency maps, and feature visualisation. These techniques can help users to gain greater insights into how the CNN model makes its predictions by highlighting regions



or most salient parts of the input image that contributes most to the final prediction made, making it more credible and user-friendly for hospital staff.

## **12.2 Integration of concepts into Business Context**

### **12.2.1 Integration with Existing Systems**

Integrating these predictive models into existing hospital information systems and workflow may pose some challenges due to the complexity and time-consuming nature of the process.

In order to address these challenges, the hospital could engage with IT professionals and system vendors early in the implementation phase to identify potential integration challenges and develop detailed plan and solutions for seamless incorporation. Such an approach minimises any technical difficulties or compatibility issues, reducing disruptions to daily operations and optimising the integration process.

### **12.2.2 Adoption and Trust**

Healthcare providers may be hesitant to rely on ML models for decision-making as they might prefer to trust their clinical judgement instead.

Hospitals can work towards mitigating this challenge by conducting educational sessions and workshops to demonstrate the models accuracy and reliability through usage of case studies or real-world examples. It is essential to emphasise that these models are used to support clinical patients and providers, rather than to replace them. By providing clear evidence of the model's effectiveness and aligning their implementation with existing clinical practices, healthcare providers can gradually overcome their scepticism and foster greater trust towards the usage of ML models when it comes to enhancing patient care.

### **12.2.3 Data Privacy and Security**

Implementing ML predictive models in healthcare may raise concerns regarding patients' data privacy and security.

In order to address these concerns, it is important for the hospital to ensure strict compliance with data protection regulations like HIPAA (Health Insurance Portability and Accountability Act), as well as implement robust security measures to safeguard patient information. Such measures would include encryption, access controls, and regular audits to monitor data handling practices. In addition, obtaining necessary consents from patients and being transparent about how their data will be used in these ML models to improve lung cancer diagnosis and treatment is essential. By prioritising patient data privacy and security, the hospital can build trust with their patients and stakeholders, ensuring responsible and ethical usage of ML models in healthcare settings.

## **13.0 Future Improvements & Implementations**

### **13.1 Incorporation of additional data types into dataset**

In order to improve the model's accuracy in the future, the hospital could consider collecting other types of data from the patient - genomic, metabolomic, and radiomic.



Firstly, the integration of genomic data presents a promising avenue for refining the predictive models. By integrating information on specific mutation or gene expression profiles associated with Lung Cancer susceptibility and progression, the hospital can identify high-risk individuals more accurately and tailor personalised treatment strategies accordingly.

In examining additional metabolite profiles in various biological samples such as blood, urine, or breath, the hospital can identify unique metabolic characteristics associated with Lung Cancer, aiding in early detection and monitoring treatment response.

Lastly, leveraging on radiomic data like quantitative features from medical images such as CT scans and PET scans and incorporating them into our predictive models allows the hospital to characterise tumour heterogeneity, stage tumours more accurately, and assess response to therapy, enhancing the model's ability to diagnose Lung Cancer.

Addition of these data types not only expand the breadth of our project but also hold the potential to significantly improve patient outcomes through more precise risk assessments and tailored treatment approaches

### **13.2 Expand AI application**

In order to enhance the scope and impact of our predictive models, a future improvement could be incorporating prognosis prediction into our models, where we aim to develop sophisticated AI models capable of predicting patient outcomes such as survival rates, recurrence risk, and treatment response. By leveraging a comprehensive dataset encompassing clinical, imaging and molecular data, these models can provide valuable insights to clinicians, aiding in treatment decision-making and resource allocation, which ultimately improves patient care outcomes.

### **14.0 Conclusion**

In conclusion, our chosen XGBoost model, alongside the CNN model, will help hospitals to more accurately predict the presence of Lung Cancer to reduce the instances of late cancer detection. While the XGBoost model primarily uses personal and symptom indicators to predict the presence of Lung Cancer, the CNN uses patients' X-ray images to check for Lung Cancer.

Incorporating both the XGBoost and CNN models provides a robust strategy for lung cancer detection, as it offers a safety net in case either model's prediction is inaccurate. By leveraging the strengths of both approaches, healthcare providers can enhance diagnostic accuracy and confidence, thereby ensuring comprehensive coverage and reducing the risk of missed diagnoses.

This dual-model approach maximises the likelihood of early detection and intervention, ultimately leading to improved patient outcomes and quality of care provided by hospitals, helping us achieve our project's ultimate objective.

## References:

- Amicizia, D., Piazza, M. F., Marchini, F., Astengo, M., Grammatico, F., Battaglini, A., Schenone, I., Sticchi, C., Laveri, R., Di Silverio, B., Andreoli, G. B., & Ansaldi, F. (2023, July 21). *Systematic review of lung cancer screening: Advancements and strategies for implementation*. Healthcare (Basel, Switzerland). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10379173/>
- Ardila, D., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6), 954-961. <https://doi.org/10.1038/s41591-019-0447-x>
- Cellina, M., Cacioppa, L. M., Cè, M., Chiarpenello, V., Costa, M., Vincenzo, Z., Pais, D., Bausano, M. V., Rossini, N., Bruno, A., & Floridi, C. (2023, August 30). *Artificial Intelligence in lung cancer screening: The future is now*. MDPI. <https://www.mdpi.com/2072-6694/15/17/4344>
- Bausewein, C., & Simon, S. T. (2013b, August). *Shortness of breath and cough in patients in palliative care*. Deutsches Arzteblatt international. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3782037/>
- Guan, X., Du, Y., Ma, R., Teng, N., Ou, S., Zhao, H., & Li, X. (2023, June 13). *Construction of the XGBOOST model for early lung cancer prediction based on metabolic indices - BMC Medical Informatics and Decision making*. BioMed Central. <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02171-x>
- Hansen, R. p et al. (2011) *Lung Cancer Statistics | How Common is Lung Cancer?* . Available at: <https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-11-284> .
- Hosny, A., et al. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), 500-510. <https://doi.org/10.1038/s41568-018-0016-5>
- Kehong, Y., Xinhai, Z., Shu Chuen , L., Congyan, Y., Wenting, W., & Susan, H. (2023, October 31). *Economic burden of advanced lung cancer patients treated by gefitinib alone and combined with chemotherapy in two regions of China*. Journal of medical economics. <https://pubmed.ncbi.nlm.nih.gov/37855437/>
- Kutikova , Lee, Chang, Long, Obasaju, & H Crown. (n.d.). *The economic burden of lung cancer and the associated costs of treatment failure in the United States*. <https://pubmed.ncbi.nlm.nih.gov/16112249/>
- McKinney, S. M., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94. <https://doi.org/10.1038/s41586-019-1799-6>
- Mokoatle, M., Marivate, V., Mapiye, D., Bornman, R., & Hayes, V. M. (2023, March 23). *A review and comparative study of cancer detection using machine learning:*

*SBERT and SimCSE application.* BMC Bioinformatics.  
<https://doi.org/10.1186/s12859-023-05235-x>

Kehong, Y., Xinhai, Z., Shu Chuen, L., Congyan, Y., Wenting, W., & Susan, H. (2023, October 31). Economic burden of advanced lung cancer patients treated by gefitinib alone and combined with chemotherapy in two regions of China. *Journal of medical economics*. <https://pubmed.ncbi.nlm.nih.gov/37855437/>

Kwee, T. C., & Kwee, R. M. (2021, June 29). *Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: Growth Expectations and role of artificial intelligence - insights into imaging*. SpringerOpen.  
<https://insightsimaging.springeropen.com/articles/10.1186/s13244-021-01031-4>

Li, & Jiang. (n.d.). (2023, November). *Medical image analysis using deep learning algorithms*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10662291/>.

Liang, D., Shi, J., Li, D., Wu, S., Jin, J., & He, Y. (2022, January 10). *Participation and yield of a lung cancer screening program in Hebei, China*. *Frontiers in oncology*.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8784378/>

Paleczek, A., Grochala, D., & Rydosz, A. (2021b, June 18). *Artificial breath classification using XGBoost algorithm for diabetes detection*. MDPI.  
<https://www.mdpi.com/1424-8220/21/12/4187>

Pinsky, P. F. (2018, June). *Lung cancer screening with low-dose CT: A world-wide view*. Translational lung cancer research.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6037972/>

PubMed Central. Mukesh. S, Ajay. K, K, Suresh Babu. *Convolutional neural network based CT scan classification method for COVID-19 test validation*.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9188200/>.

Rampinelli, C., Calloni, S. F., Minotti, M., & Bellomi, M. (2016, May 17). *Spectrum of early lung cancer presentation in low-dose screening CT: A Pictorial Review - insights into imaging*. SpringerOpen.  
<https://insightsimaging.springeropen.com/articles/10.1007/s13244-016-0487-4>

Rotter, J., Spencer, J. C., & Wheeler, S. B. (2019, April). *Financial toxicity in advanced and metastatic cancer: Overburdened and underprepared*. *Journal of oncology practice*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6494243/>

Sartini, M., Carbone, A., Demartini, A., Giribone, L., Oliva, M., Spagnolo, A. M., Cremonesi, P., Canale, F., & Cristina, M. L. (2022, August 25). *Overcrowding in Emergency Department: Causes, Consequences, and Solutions—A Narrative Review*. MDPI Open Access Journals.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9498666/>

Onder, O., Yarasir, Y., Azizova, A., Durhan, G., Onur, M. R., & Ariyurek, O. M. (2021, April 20). *Errors, discrepancies and underlying bias in radiology with case examples:*

*A pictorial review - insights into imaging.* SpringerOpen.  
<https://insightsimaging.springeropen.com/articles/10.1186/s13244-021-00986-8>

Shaughnessy, A.F. (2017) *High false-positive rate with lung cancer screening, American Family Physician.* Available at:  
<https://www.aafp.org/pubs/afp/issues/2017/0715/p128a.html>

Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A., & Abdani, S. R. (2023, August 8). *A review of deep learning techniques for lung cancer screening and diagnosis based on CT images.* MDPI. <https://www.mdpi.com/2075-4418/13/16/2617>

Shojania, K. G., & Forster, A. J. (2008, July 15). *Hospital mortality: When failure is not a good measure of success.* CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2443229/>

Teemu Kanstr & eacute;n, T. (2023a, August 4). *A look at precision, recall, and F1-score.* Medium.  
<https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec>

Timesofindia (2023) *Lung cancer survivor: All you need to know if you're a cancer survivor, The Times of India.* Available at:  
<https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/lung-cancer-4-warning-signs-that-may-appear-first-thing-in-the-morning/photostory/99693639.cms>

U.S. Food and Drug Administration. (2021). *Artificial intelligence and machine learning in software as a medical device.* Retrieved from  
<https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>

Varela, A., & Pritchard, M. E. (2011). *Peer influence: Use of alcohol, tobacco, and prescription medications.* Journal of American college health: J of ACH.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5916837/>

Wang, Y., Cai, H., Pu, Y., Li, J., Yang, F., Yang, C., Chen, L., & Hu, Z. (2022, March 30). *The value of AI in the diagnosis, treatment, and prognosis of malignant lung cancer.* Frontiers. <https://www.frontiersin.org/articles/10.3389/fradi.2022.810731/full>

## APPENDICES

### Appendix 1: Data Cleaning

Data cleaning is essential in data analysis, as it ensures that the dataset's quality is upheld and that subsequent analyses are based on accurate and consistent information, to derive meaningful insights.

The table below is a summary of the overall data cleaning process for survey lung cancer.csv:

Variable	Action	Reasoning																																
All	Check for Missing Data	Output indicates no missing values <pre>print(df.isnull().sum())</pre> <table><tr><td>GENDER</td><td>0</td></tr><tr><td>AGE</td><td>0</td></tr><tr><td>SMOKING</td><td>0</td></tr><tr><td>YELLOW_FINGERS</td><td>0</td></tr><tr><td>ANXIETY</td><td>0</td></tr><tr><td>PEER_PRESSURE</td><td>0</td></tr><tr><td>CHRONIC_DISEASE</td><td>0</td></tr><tr><td>FATIGUE</td><td>0</td></tr><tr><td>ALLERGY</td><td>0</td></tr><tr><td>WHEEZING</td><td>0</td></tr><tr><td>ALCOHOL_CONSUMING</td><td>0</td></tr><tr><td>COUGHING</td><td>0</td></tr><tr><td>SHORTNESS_OF_BREATH</td><td>0</td></tr><tr><td>SWALLOWING_DIFFICULTY</td><td>0</td></tr><tr><td>CHEST_PAIN</td><td>0</td></tr><tr><td>LUNG_CANCER</td><td>0</td></tr></table> <pre>dtype: int64</pre>	GENDER	0	AGE	0	SMOKING	0	YELLOW_FINGERS	0	ANXIETY	0	PEER_PRESSURE	0	CHRONIC_DISEASE	0	FATIGUE	0	ALLERGY	0	WHEEZING	0	ALCOHOL_CONSUMING	0	COUGHING	0	SHORTNESS_OF_BREATH	0	SWALLOWING_DIFFICULTY	0	CHEST_PAIN	0	LUNG_CANCER	0
GENDER	0																																	
AGE	0																																	
SMOKING	0																																	
YELLOW_FINGERS	0																																	
ANXIETY	0																																	
PEER_PRESSURE	0																																	
CHRONIC_DISEASE	0																																	
FATIGUE	0																																	
ALLERGY	0																																	
WHEEZING	0																																	
ALCOHOL_CONSUMING	0																																	
COUGHING	0																																	
SHORTNESS_OF_BREATH	0																																	
SWALLOWING_DIFFICULTY	0																																	
CHEST_PAIN	0																																	
LUNG_CANCER	0																																	
	Checking for Duplicated Values	<pre>print("\nDuplicate rows:") print(df.duplicated().sum())</pre> Duplicate rows: 33																																
	Remove Duplicated Values	<pre>df.drop_duplicates(inplace=True) print("\nDuplicate rows:") print(df.duplicated().sum())</pre> Duplicate rows: 0																																

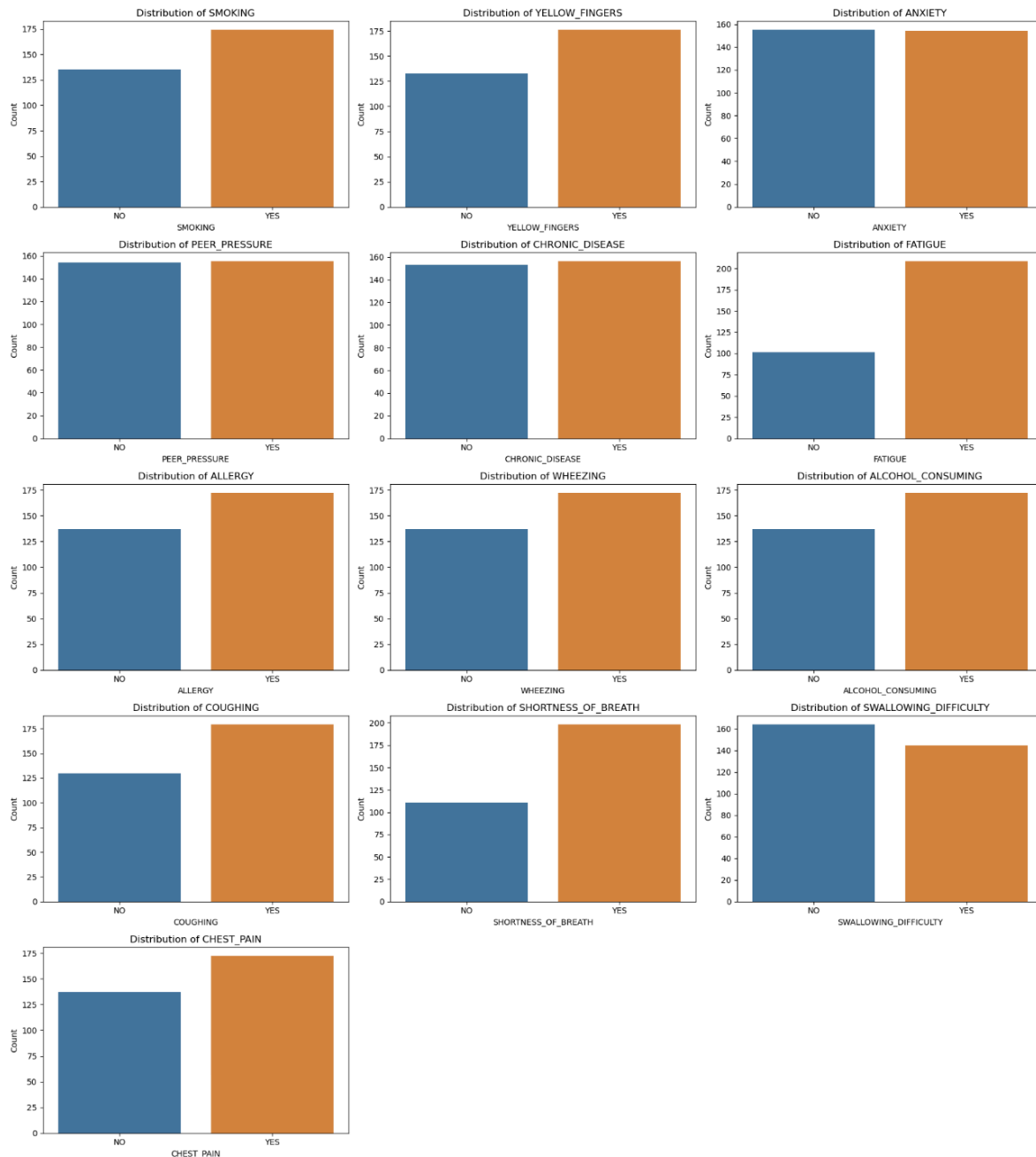
### Appendix 1.1 Data Cleaning for CNN

When preparing dataset for training Convolutional Neural Network (CNN), some steps are taken for data cleaning and preprocessing. We have to handle inconsistent data for “train\_test\_spilt” to handle inconsistent data issues in “train\_test\_spilt”, we have to duplicate the shorter array to match the length of the longer one. The “balance\_data” function ensures

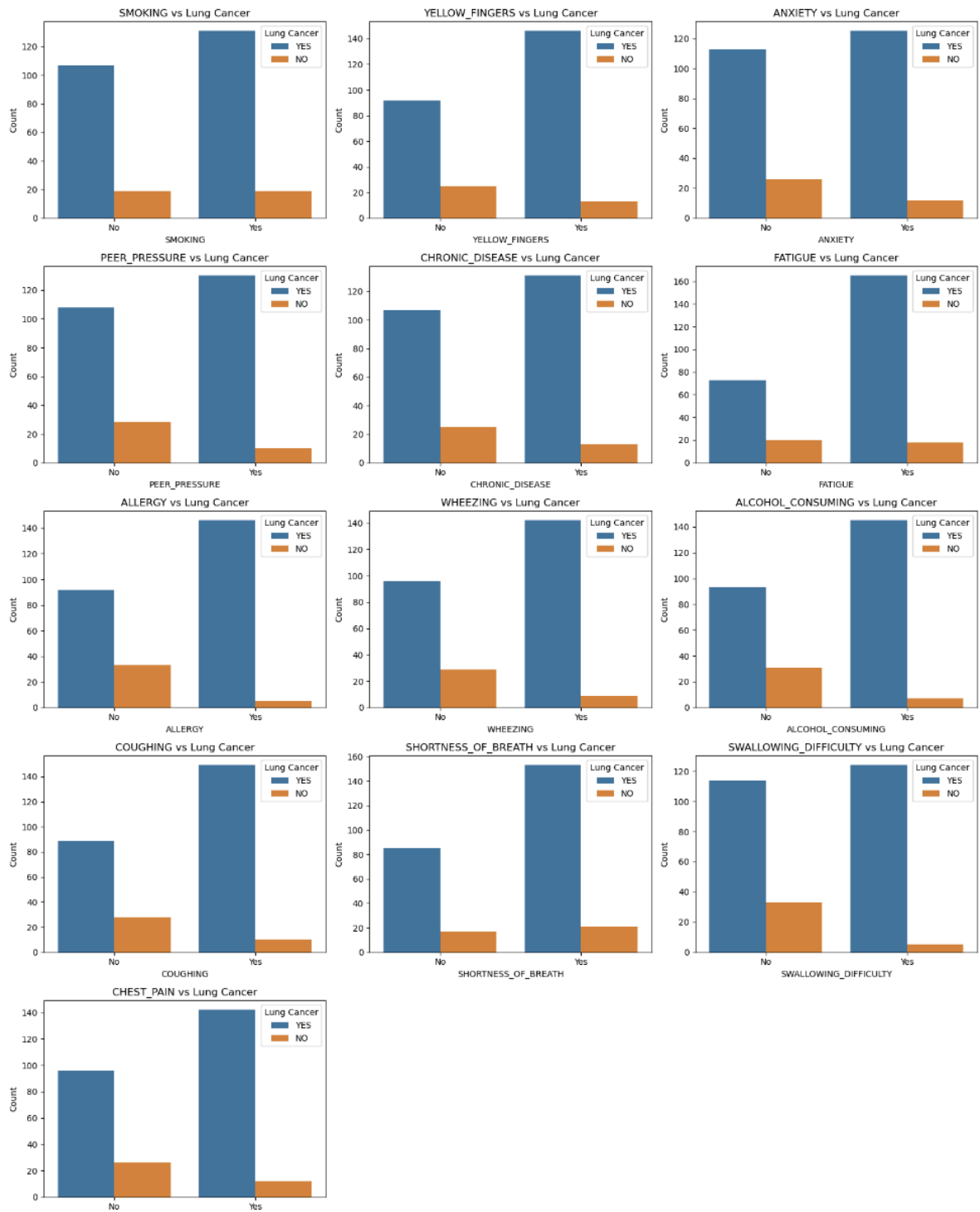
that both “X\_train\_paths” and “Y\_train\_labels” have the same number of samples for each class. It does so by duplicating the samples of the minority class. In addition, the CNN data set consists of both .jpg and .png thus we would have to amend the code accordingly.

## Appendix 2: Data Exploration

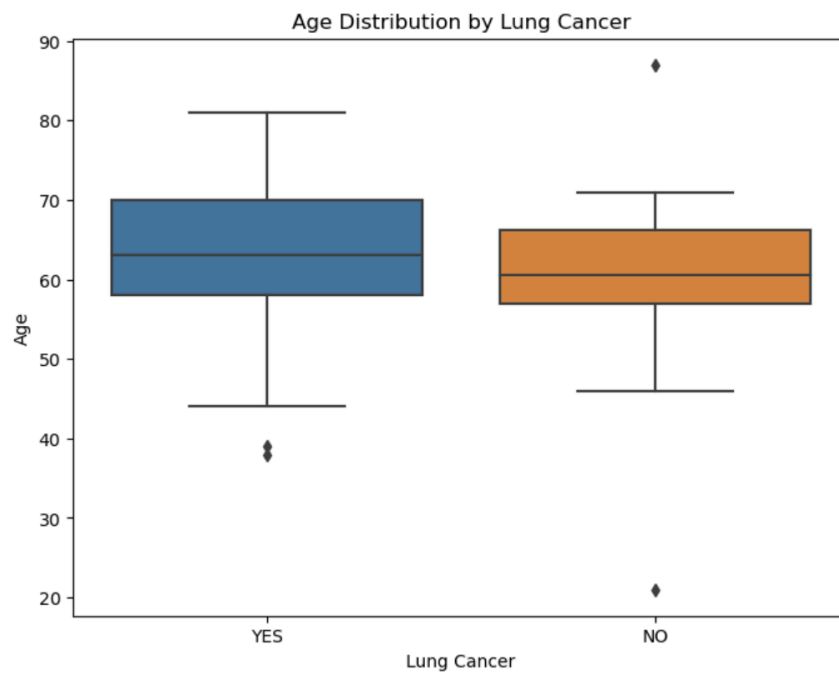
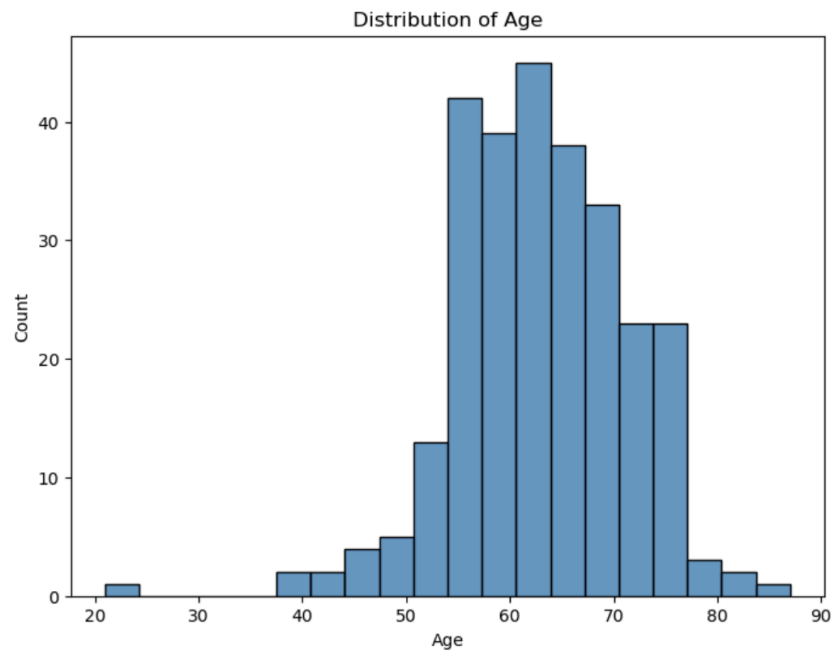
In order to have a better idea of the relationship between the variables, we have performed various plots among the various variables. Firstly, we plotted out the distribution of the categorical variables



Next, we performed various plots between the lung cancer category and various Y variables.

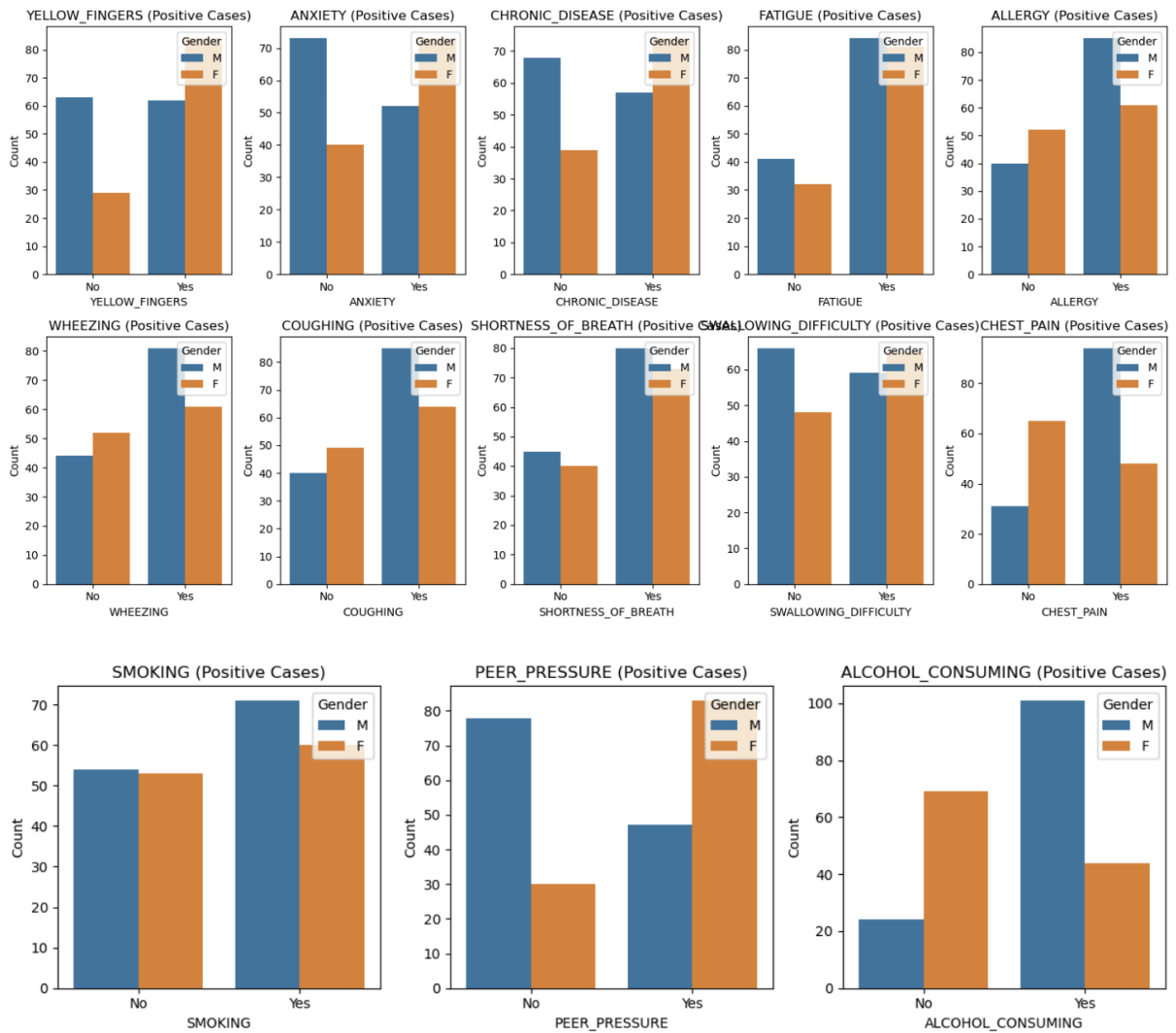
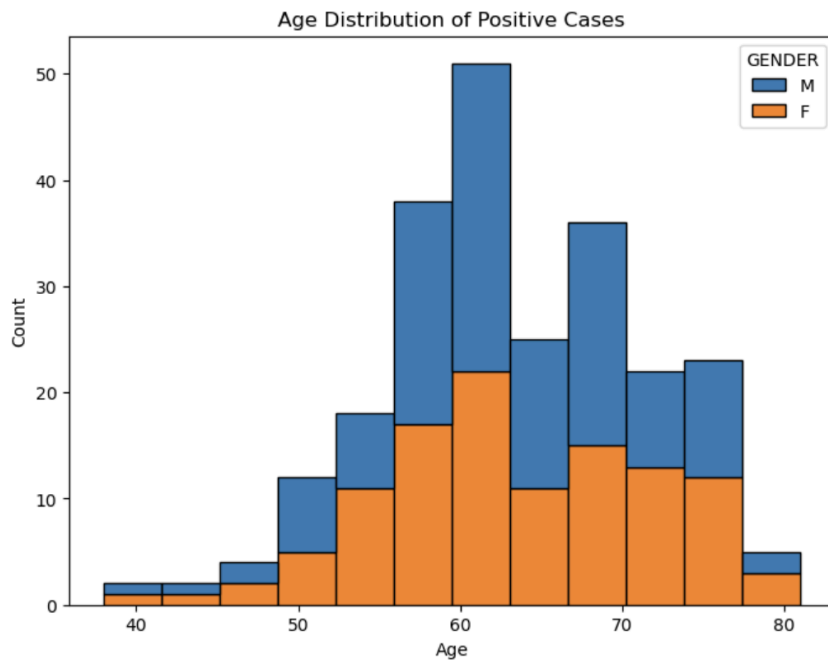


We also explored the distribution of the numerical variable (“Age”) across the dataset as well as the distribution between ‘Age’ and ‘LUNG\_CANCER’ to better visualise the relationship between the two variables.



Furthermore, we performed plots on the age distribution of positive cases by gender as well to observe the effects on gender and age on positive cases.





### Appendix 3: Description for Logistic Regression Model

Optimization terminated successfully.  
Current function value: 0.140998  
Iterations 10

Logit Regression Results						
Dep. Variable:	LUNG_CANCER	No. Observations:	249			
Model:	Logit	Df Residuals:	233			
Method:	MLE	Df Model:	15			
Date:	Sun, 31 Mar 2024	Pseudo R-squ.:	0.6463			
Time:	05:51:38	Log-Likelihood:	-35.108			
converged:	True	LL-Null:	-99.262			
Covariance Type:	nonrobust	LLR p-value:	4.557e-20			
	coef	std err	z	P> z	[0.025	0.975]
const	-5.8154	3.023	-1.924	0.054	-11.740	0.109
GENDER	1.1429	0.834	1.370	0.171	-0.492	2.778
AGE	-0.0527	0.050	-1.064	0.287	-0.150	0.044
SMOKING	2.0671	0.807	2.560	0.010	0.485	3.650
YELLOW_FINGERS	1.0879	0.811	1.341	0.180	-0.502	2.678
ANXIETY	1.2503	0.949	1.317	0.188	-0.610	3.111
PEER_PRESSURE	2.1574	0.806	2.677	0.007	0.578	3.737
CHRONIC_DISEASE	3.5577	1.103	3.226	0.001	1.396	5.719
FATIGUE	3.7743	1.032	3.657	0.000	1.751	5.797
ALLERGY	1.8996	0.911	2.085	0.037	0.114	3.686
WHEEZING	0.9744	0.930	1.048	0.295	-0.848	2.796
ALCOHOL_CONSUMING	1.2865	0.903	1.425	0.154	-0.483	3.056
COUGHING	4.2248	1.370	3.083	0.002	1.539	6.911
SHORTNESS_OF_BREATH	-0.7154	0.832	-0.860	0.390	-2.347	0.916
SWALLOWING_DIFFICULTY	3.3758	1.369	2.466	0.014	0.692	6.059
CHEST_PAIN	0.5750	0.814	0.706	0.480	-1.021	2.171

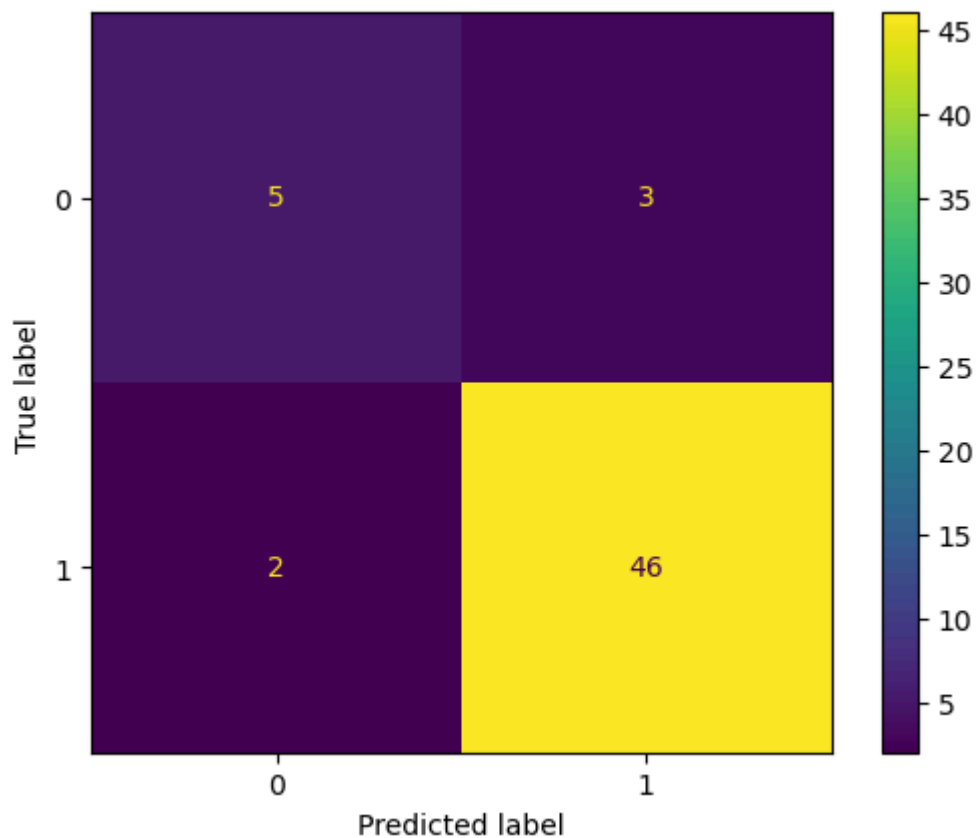
Accuracy: 0.8888888888888888  
Precision: 0.9166666666666666  
Recall: 0.9565217391304348  
F1-score: 0.9361702127659575

### Appendix 4: Description for Random Forest Model

Random Forest Accuracy: 0.9107142857142857  
Random Forest Precision: 0.9387755102040817  
Random Forest Recall: 0.9583333333333334  
Random Forest F1-score: 0.9484536082474226  
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.71	0.62	0.67	8
1	0.94	0.96	0.95	48
accuracy		0.91		56
macro avg	0.83	0.79	0.81	56
weighted avg	0.91	0.91	0.91	56

### **Random Forest Confusion Matrix:**



### **Using 10-fold cross-validation to check for overfitting**

Cross-validation scores: [0.92857143 0.92857143 0.85714286 0.85714286  
0.82142857 0.92857143 0.85185185 0.96296296 0.92592593 0.88888889]  
Mean cross-validation score: 0.8951058201058201

### **Random Forest Risk Scores:**

Patient 1: Probability = 0.985, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 2: Probability = 0.99, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 3: Probability = 1.0, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 4: Probability = 1.0, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 5: Probability = 0.455, Risk Category = Medium Risk, Actual Lung Cancer = 0  
Patient 6: Probability = 0.995, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 7: Probability = 0.97, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 8: Probability = 0.995, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 9: Probability = 0.995, Risk Category = High Risk, Actual Lung Cancer = 1  
Patient 10: Probability = 0.955, Risk Category = High Risk, Actual Lung Cancer = 1

## Appendix 5: Description for Multivariate Adaptive Regression Splines (MARS) Model

MARS Accuracy: 0.9107142857142857

MARS Precision: 0.9215686274509803

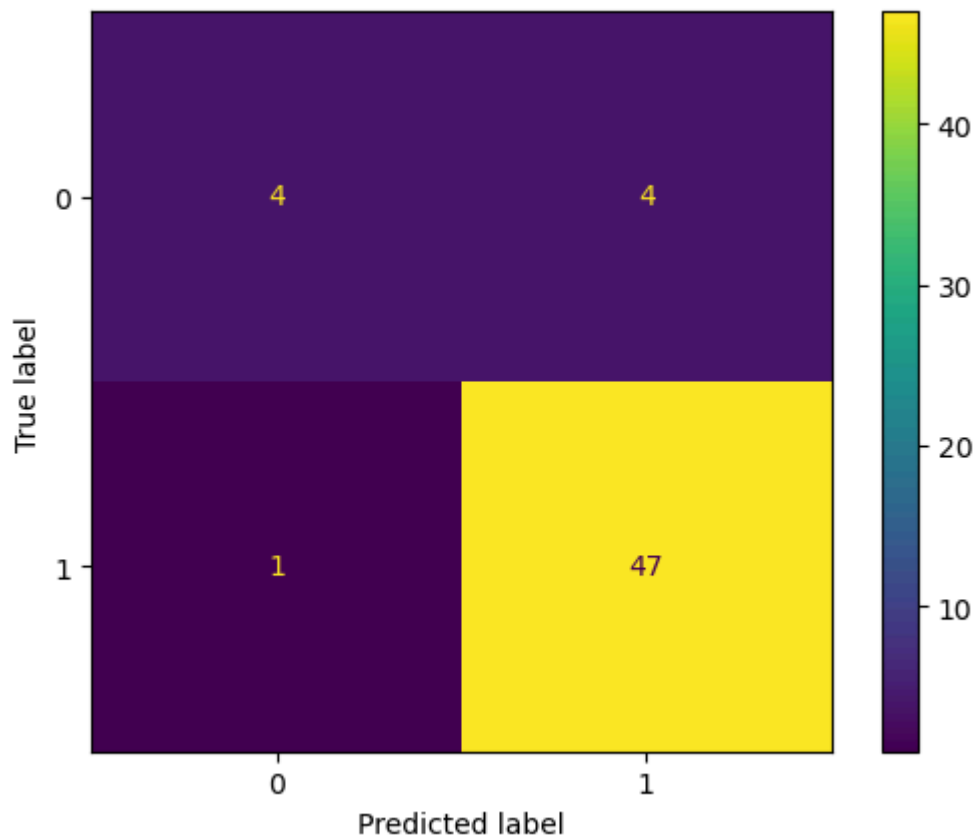
MARS Recall: 0.9791666666666666

MARS F1-score: 0.9494949494949495

MARS Classification Report:

	precision	recall	f1-score	support
0	0.80	0.50	0.62	8
1	0.92	0.98	0.95	48
accuracy			0.91	56
macro avg	0.86	0.74	0.78	56
weighted avg	0.90	0.91	0.90	56

### MARS Confusion Matrix:



Using 10-fold cross-validation to check for overfitting

**Cross-validation ROC\_AUC scores:** [0.921875, 0.984375, 1.0, 0.96875, 0.9635416666666667, 0.765625, 0.9722222222222222, 0.9791666666666667, 0.9347826086956521, 0.9130434782608696]

**Mean cross-validation ROC\_AUC score:** 0.9403381642512076

#### **MARS Risk Scores:**

Patient 1: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 2: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 3: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 4: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 5: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 0  
 Patient 6: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 7: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 8: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 9: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1  
 Patient 10: Probability = 1, Risk Category = High Risk, Actual Lung Cancer = 1

#### **Appendix 6: Description for XGBoost Regression Model**

XGBoost Accuracy: 0.8928571428571429

XGBoost Precision: 0.9038461538461539

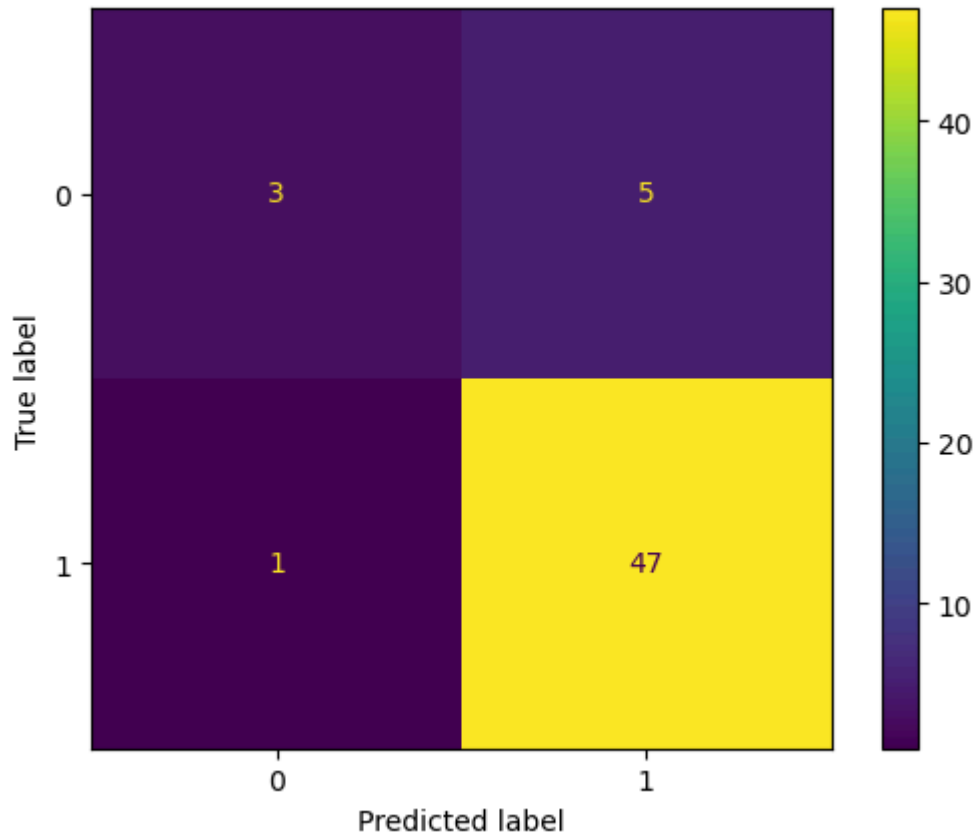
XGBoost Recall: 0.9791666666666666

XGBoost F1-score: 0.94

#### **XGBoost Classification Report:**

	precision	recall	f1-score	support
0	0.75	0.38	0.50	8
1	0.90	0.98	0.94	48
accuracy			0.89	56
macro avg	0.83	0.68	0.72	56
weighted avg	0.88	0.89	0.88	56

#### **XGBoost Confusion Matrix:**



#### Using 10-fold cross-validation to check for overfitting

**XGBoost Cross-validation accuracy scores:** [0.89285714, 0.85714286, 0.89285714, 0.82142857, 0.89285714, 0.89285714, 0.81481481, 1, 0.88888889, 0.85185185]

**XGBoost Mean cross-validation accuracy scores:** 0.8805555555555555

#### XGBoost Risk Scores:

Patient 1: Probability = 0.9775846600532532, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 2: Probability = 0.9726433157920837, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 3: Probability = 0.9427878856658936, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 4: Probability = 0.9933772683143616, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 5: Probability = 0.8073779344558716, Risk Category = High Risk, Actual Lung Cancer = 0

Patient 6: Probability = 0.9577101469039917, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 7: Probability = 0.9832395911216736, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 8: Probability = 0.9637102484703064, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 9: Probability = 0.9847689867019653, Risk Category = High Risk, Actual Lung Cancer = 1

Patient 10: Probability = 0.9484269618988037, Risk Category = High Risk, Actual Lung Cancer = 1

#### **Appendix 6: Top 10 Association Rules From Apriori Algorithm**

	antecedents \
22419	(ANXIETY, WHEEZING, PEER_PRESSURE)
22394	(SWALLOWING_DIFFICULTY, GENDER, COUGHING, YELLOW_FINGERS)
22976	(GENDER, SWALLOWING_DIFFICULTY, FATIGUE, YELLOW_FINGERS)
22973	(ANXIETY, SMOKING, SHORTNESS_OF_BREATH, PEER_PRESSURE)
22379	(ANXIETY, PEER_PRESSURE, WHEEZING, YELLOW_FINGERS)
22434	(SWALLOWING_DIFFICULTY, GENDER, COUGHING)
22399	(SWALLOWING_DIFFICULTY, GENDER, WHEEZING, YELLOW_FINGERS)
22414	(ANXIETY, PEER_PRESSURE, COUGHING)
22267	(GENDER, SWALLOWING_DIFFICULTY, FATIGUE, YELLOW_FINGERS)
22294	(ANXIETY, SHORTNESS_OF_BREATH, PEER_PRESSURE)

	consequents
22419	(SWALLOWING_DIFFICULTY, GENDER, COUGHING, YELLOW_FINGERS) 5.400000
22394	(ANXIETY, WHEEZING, PEER_PRESSURE) 5.400000
22976	(ANXIETY, SMOKING, SHORTNESS_OF_BREATH, PEER_PRESSURE) 5.062500
22973	(GENDER, SWALLOWING_DIFFICULTY, FATIGUE, YELLOW_FINGERS) 5.062500
22379	(SWALLOWING_DIFFICULTY, GENDER, COUGHING) 4.983520
22434	(ANXIETY, PEER_PRESSURE, WHEEZING, YELLOW_FINGERS) 4.983520
22399	(ANXIETY, PEER_PRESSURE, COUGHING) 4.772727
22414	(SWALLOWING_DIFFICULTY, GENDER, WHEEZING, YELLOW_FINGERS) 4.772727

22267 (ANXIETY, SHORTNESS\_OF\_BREATH, PEER\_PRESSURE)

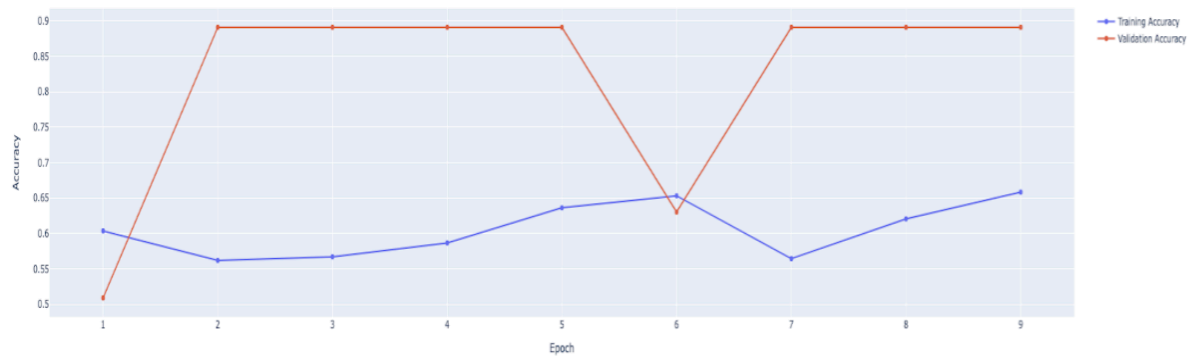
4.632353

22294 (GENDER, SWALLOWING\_DIFFICULTY, FATIGUE,  
YELLOW\_FINGERS) 4.632353

## **Appendix 7: Training and Validation Accuracy for CNN, Inception V3 and VGG16**

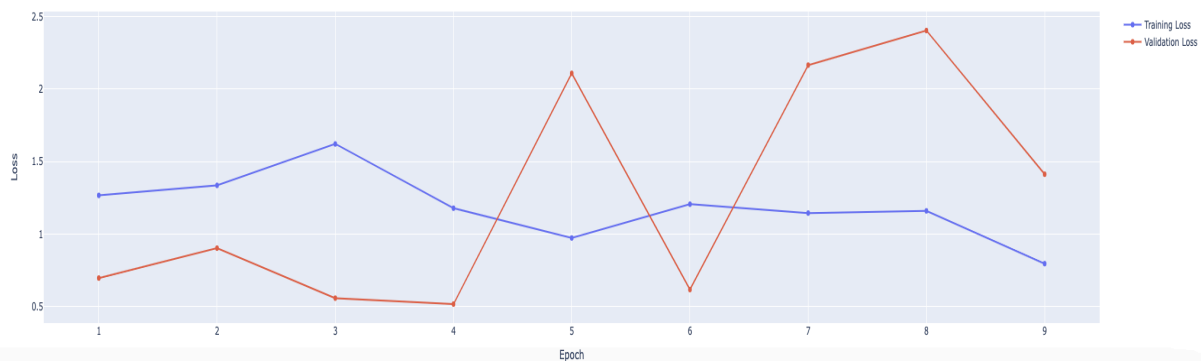
### Training and Validation Accuracy for CNN

Our training accuracy increases steadily over the epochs, which means that the model's performance on the training data is improving. Validation accuracy also increases over the epochs, but not to the same extent as training accuracy. This suggests that the model may be overfitting the data.



### Training and Validation Loss for CNN

Both training and validation loss decreases over the epochs, it means that the model's performance is improving and training loss decreases more rapidly than validation loss. It means that the model may be overfitting the data.



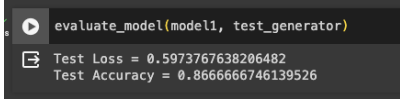
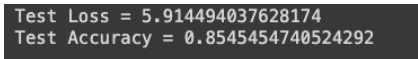
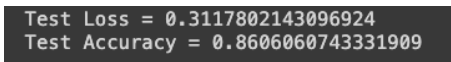
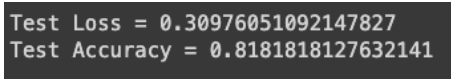
#### **CNN Model**

#### **Test loss**

**0.5973767638206482**

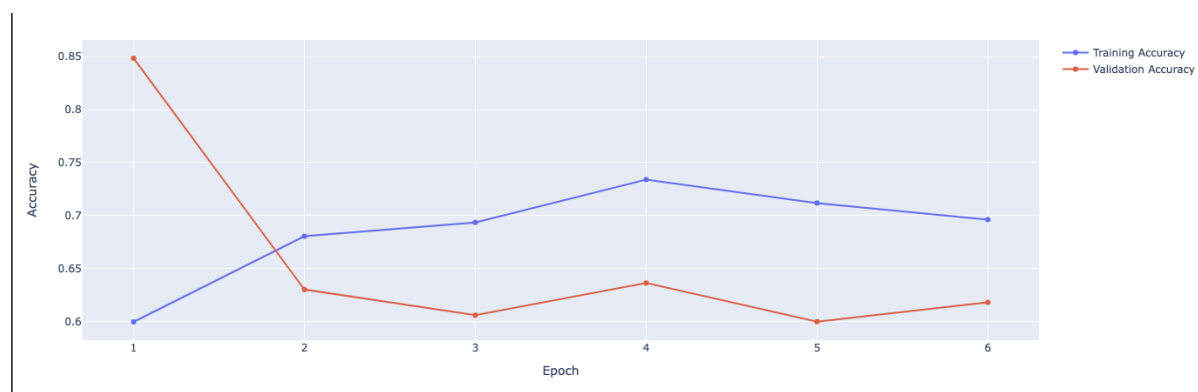
**Test Loss:** The test loss of approximately 0.5974 is moderate,



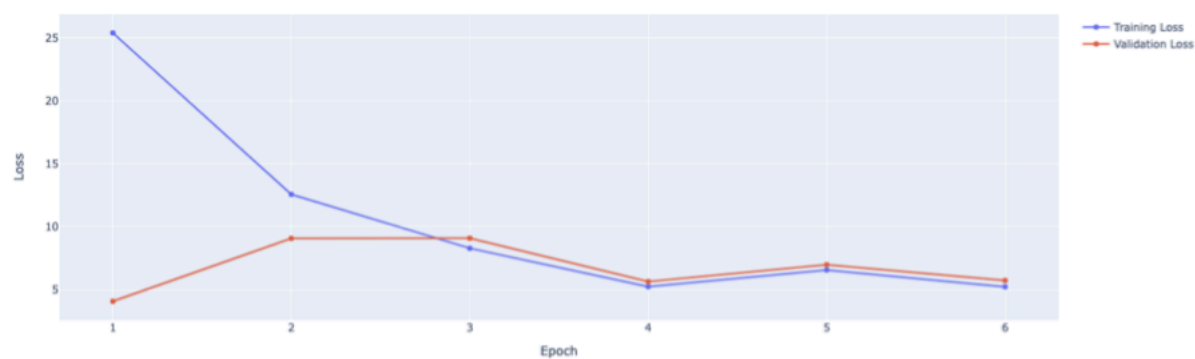
	<p><b><u>Test Accuracy</u></b> <b>0.8666666746139526</b></p> 	<p>suggesting that the model is making some errors in its predictions but is not severely overfitting or underfitting the data.</p> <p><b>Test Accuracy:</b> An accuracy of 86.67% is relatively high, indicating that the model is performing well in classifying the CT scan images into "cancer" and "no_cancer" classes.</p> <p>The test accuracy of 86.67% suggests that the model is performing reasonably well in distinguishing between cancerous and non-cancerous CT scan images.</p> <p>The test loss of 0.5974 indicates that there is still room for improvement in reducing the prediction errors.</p>
<b><u>Inception V3</u></b>		<p>Inception V3 test loss is higher than CNN test loss. Lower loss indicates a better performance on the training data. Thus, CNN seems to be fitting the training data better.</p> <p>Likewise, CNN has a higher test accuracy compared to inception V3.</p>
<b><u>VGG16</u></b>	 <p><b><u>After fine tuning</u></b></p> 	<p>VGG16 test loss is at 0.311 it shows how well a model performs on a task.</p> <p>In addition, having a test accuracy of 0.86 is higher than inception v3 and CNN</p> <p>After fine tuning, the model shows</p>

		a slight improvement in fitting the fine-tuning data.
--	--	---

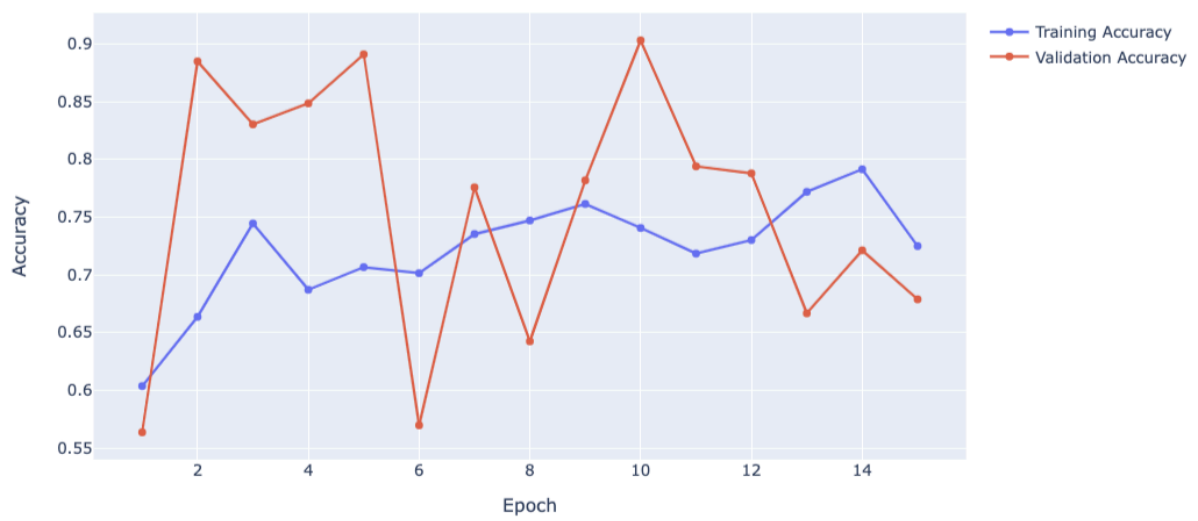
### Training and Validation Accuracy for Inception V3



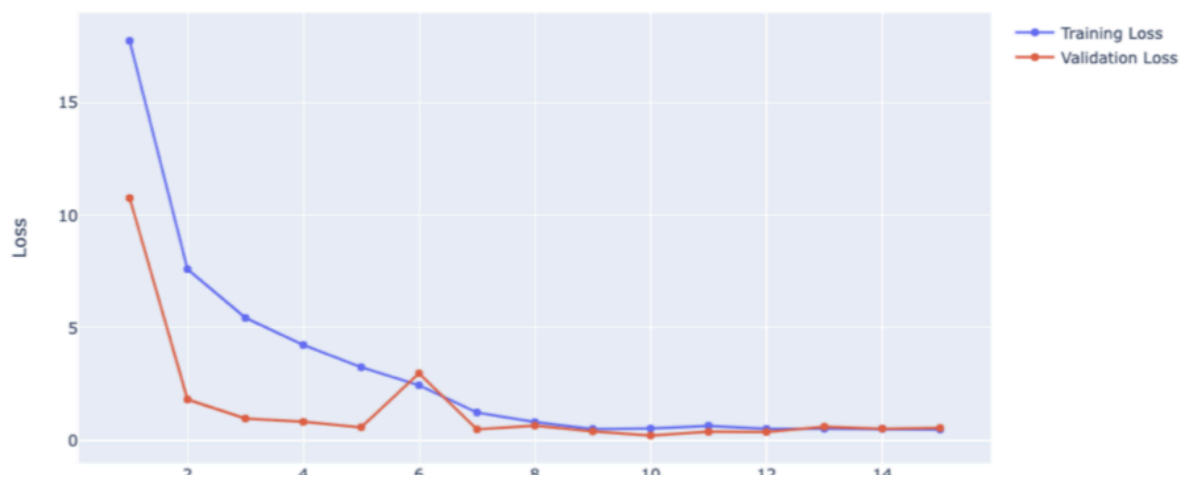
### Training and Validation Loss for Inception V3



### Training and Validation Accuracy for VGG16



Training and Validation loss for VGG16



After fine tuning VGG16

