Predictive Analysis of Lung Cancer Susceptibility Based on Symptomatic Indicators

ABDALRAHMAN AL-QANNAS[1] and Mariam Biltawi [2]

[1] Al Hussein Technical University, Amman 11831, Jordan
21110285@htu.edu.jo, mariam.biltawi@htu.edu.jo

Abstract
This research aims to evaluate symptomatic indicators including coughing, smoking, and chest discomfort using machine learning techniques to support early and accurate lung cancer diagnosis. While secondary data was obtained from a small database, primary data was gathered by means of surveys aimed at medical and data science experts. Data imbalances were corrected with advanced preprocessing techniques including the Synthetic Minority Over-sampling Technique (SMOTE). Using logistic regression and random forest models, both models display good test accuracy (95.37%). Results highlight the need for diverse data collecting, regular model changes, and using multimodal data for higher predictive accuracy. Future research involves expanding databases containing real-time data and improving models to enhance clinical use and early lung cancer identification.

Keywords: Lung Cancer Prediction, Symptoms Indicators, Predictive Analytics

1. Introduction

"Big data" in the digital world of today refers to huge, complicated sets larger than what could be handled with traditional techniques. Along with the continuous flow of data generated by sensors and other sources, it shows a previously unseen integration of data coming from a variety of sources, including sophisticated online transactions and dynamic social media exchanges. Traditional techniques of data processing, storage, and analysis find great challenges from this amazing mix of volume, fast speed, and variety of data. Big data marks an evolution in how businesses manage data. It changes the way companies use data by letting them gather and examine vast amounts of data as well as obtain insightful analysis that guides their development and enhancement of their processes of decision-making.

By offering huge amounts of data that can be examined for insights, big data has transformed the environment of multiple fields of science, including healthcare. Within the context of lung cancer, one of the main causes of cancer-related mortality globally, big data provides advanced analytics' ability to significantly improve early diagnosis and treatment plans.

The objective of the study is to use machine learning techniques to predict the possibility of lung cancer based on symptoms including coughing, smoking, and chest pain. It looks for correlations in the symptom data that could help with early and accurate identification and better patient outcomes. The main objective is to identify important markers of lung cancer using advanced statistical methods and investigate the correlations between these symptoms to improve prediction accuracy. In the end, the research aims to improve predictive modeling methods,

advance early detection measures, and expand knowledge of symptom relationships in medical diagnostics.

The research structure is described in this paper, which begins with an introduction and a Section 2 related work. In Section 3, data gathering processes are described in detail; in Section 4, analytical approaches are covered. Section 5 of the report discusses the results, while Section 6 provides findings and recommendations for the future.

2. Related Work

Among the main causes of cancer-related deaths worldwide, lung cancer calls for early diagnosis and accurate prediction to raise survival rates. The development of deep learning and machine learning has brought fresh approaches to improve lung cancer prediction's efficiency and precision. Examining many studies that have used machine learning models to predict lung cancer based on symptomatic indicators, risk factors, and medical imaging data, this section of the literature highlights their approaches, results, and areas of weakness in current research.

Ausawalaithong et al. (Ausawalaithong Worawate, 2018) predicted lung cancer from chest X-ray pictures using the DenseNet-121 deep learning network. Using big image datasets to improve performance despite their tiny datasets, their approach included transfer learning. Though the tiny dataset remained an important barrier, their results showed a mean accuracy of 74.43%, therefore highlighting the possibility in clinical applications of the model.

Sowjanya et al. (Sowjanya, 2016) studied examined several machine learning methods for chest X-ray image lung cancer detection. While comprehensive accuracy measures were not stated, their analysis of several algorithms found that select models, especially convolutional neural networks (CNNs) were quite effective in spotting lung cancer indicators in medical imaging.

Radhika et al. (Radhika P R, 2019) investigated logistic regression, naive Bayes, random forest, and support vector machines to find the best successful machine learning method for lung cancer identification. With logistic regression obtaining a mean accuracy of 82%, their comparison research showed that random forest models and logistic regression achieved strong results, so highlighting their effectiveness in predictive modeling.

Ahmed and Mayya (Ahmad, 2020) applied logistic regression and decision tree models to predict lung cancer risk factors like smoking and chronic disease. The tool had 93.33% total accuracy on 1000 medical records. Considering limitations in determining particular ages of onset and other cancer types, their study highlighted the tool's early lung cancer prediction applications.

Salaken et al. (Salaken, 2021) diagnosed lung cancer in a small dataset using convolutional neural network deep learning. They used deep autoencoders for feature extraction and achieved an AUC of 99.3% and 97.1% accuracy, indicating that deep learning can handle small data sets and improve diagnostic

accuracy.

Nemlander et al. (Nemlander, 2022) applied logistic regression and random forest models to predict lung cancer based on customized e-questionnaire symptoms and smoking status. They found 82% accuracy for never smokers, 77% for current smokers, and 63% for previous smokers. This study highlighted symptom-based data collecting and smoking status's impact on predictive modeling.

Dritsas and Trigka (Dritsas, 2019) enhanced lung cancer prediction using machine learning models on a small dataset. Their investigation produced an accuracy of 97.1%, greatly exceeding random forest models with an accuracy of roughly 90%. This study showed that careful model optimization improves prediction performance

Study	Methods	Dataset Size	Key Metrics (Accuracy)	Objective
Ausawalaithong et al. (2018)	DenseNet-121, Transfer Learning	Small	74.43%	X-ray Images
Sowjanya et al. (2016)	Various ML Techniques, CNNs	Medium	Not specified	X-ray Images
Radhika et al. (2019)	Logistic Regression, Naive Bayes, etc.	Medium	82%	Symptoms
Ahmad and Mayya (2020)	Logistic Regression, Decision Tree	1000 records	93.33%	Risk Factors
Salaken et al. (2021)	Deep Learned Features, CNNs	Small	97.1%	Small Dataset
Nemlander et al. (2022)	Logistic Regression, Random Forest	Varies	Never Smokers: 82%, Current Smokers: 77%, Former Smokers: 63%	Symptoms
Dritsas and Trigka (2019)	Optimized ML Models	Small	97.1%	Small Dataset

Many studies have shown encouraging outcomes from the integration of machine learning algorithms into lung cancer prediction. By using both imaging and non-imaging data, these models present great possibilities for early diagnosis and tailored risk assessment, so enhancing patient outcomes. Refine these predictive models and improve their clinical applicability by means of ongoing research including diverse datasets and advanced algorithms.

3. Data collection and Description

This section presents an overview of the data collecting methods used in this research, describing both primary and secondary data sources. It justifies the choices taken for data gathering and describes the advantages and limits of each type of data.

3.1 Primary Data

Primary data was collected through a survey targeting professionals in oncology, radiology, pulmonology, and data science. The survey aims to obtain insights into the number and importance of various lung cancer symptoms and the performance of machine learning models in diagnosing lung cancer. The total respondents are 34 including 24 data scientists and 10 as (Oncology, Radiology, Pulmonology) which shows us a huge involvement for the data scientist in this research. Below are the details of the survey questions

- Age
- Gender (Male, Female)
- Area of Specialization (Oncology, Radiology, Pulmonology, Data Scientist)
- Symptomatic Indicators:
  How often do you encounter patients with a history of smoking? (Always, Often, Sometimes, Rarely, Never)
  Which symptoms do you consider are the most important indicators of lung cancer? (Shortness of breath, Chest pain, Chronic disease, Smoking, Coughing, Fatigue, Swallowing Difficulty, Peer Pressure, Wheezing, Yellow Fingers, Allergy, Alcohol Consuming, Anxiety)
- Data Scientist Specific Questions:
  Have you ever used or considered using machine learning models in medical diagnostics? (Yes, No)
  How effective do you believe machine learning models are in predicting lung cancer based on symptoms? (Very effective, Effective, moderately effective, not effective)
  What types of machine learning models do you think have been most effective for predicting lung cancer from symptomatic data? (Logistic regression, Decision tree, Random Forest, Gradient boosting) What improvements in data collection or model training techniques do you think could enhance the accuracy of lung cancer predictions?

A recent survey collected responses from 34 professionals in oncology, radiology, pulmonology, and data science to get insights regarding lung cancer symptoms and the importance of machine learning models in identifying and diagnosing the cancer. Key factors collected included age, gender, area of specialization, symptomatic indicators, and decisions on the effectiveness of machine learning in this context. The purpose of the survey is to better understand the current state of lung cancer symptom identification and the potential role machine learning might play in improving patient outcomes.
The choice of implementing a survey was justified by its importance to acquire current, accurate viewpoints from specialists directly involved in lung cancer diagnosis and treatment. This approach allowed for the collection of

quantitative data, providing an accurate representation of present methods and opinions.

Merits:

- Direct insights from experts.
- Flexibility in data collection.
- Specific to research objectives.

Limitations:

- Potential sample bias.
- Subjectivity in responses.

Key findings:

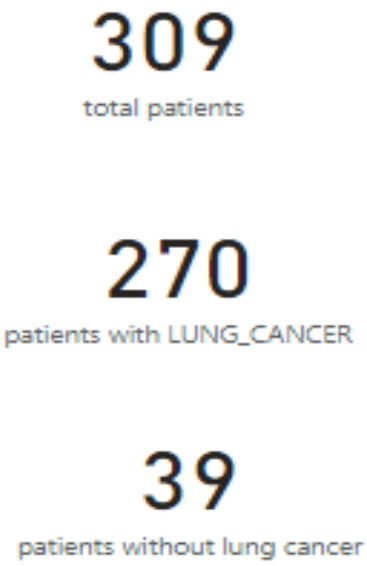
This study reveals a substantial interest among data scientists in employing machine learning for lung cancer diagnosis. Participants showed confidence in existing models, especially logistic regression and decision trees, while highlighting the need for more diverse datasets to boost accuracy. Key symptoms observed by medical specialists included smoking, shortness of breath, coughing, and chest pain. To further improve model reliability, the study advises applying advanced techniques such as transfer learning and continuous model updates.

3.2 Secondary Data

Secondary data comes from a Kaggle dataset with aggregated lung cancer case data including symptoms and medical histories.

Dataset Source: **Kaggle lung cancer dataset**. This dataset has also been used and analyzed in Elias Dritsas and Maria Trigka's paper on lung cancer risk prediction with machine learning algorithms (Dritsas, 2019) , offering a useful reference for validating our methods.

Overview of Secondary Data:



Sample Size: 309 records, containing 270 lung cancer cases and 39 without lung cancer. The dataset includes 15 features and one label ("LUNG_CANCER"),

Key Features:

- Demographics: Age, Gender.
- Symptoms: Shortness of breath, chest pain, smoking history, and more.
  Medical History: Chronic disease status, alcohol consumption, allergies, and more

The dataset was chosen for its detailed structure and the wide range of features it have, making it appropriate for robust statistical analysis and model development. Its application in past studies by Elias Dritsas and Maria Trigka (Dritsas, 2019) adds an extra degree of validation and comparison for our work.

Merits:

Pre-Existing Research: The dataset's use in previous research, such as the one by Elias Dritsas and Maria Trigka, (Dritsas, 2019) enables comparison and enhancing the validity of the findings.

Richness of Data: The dataset contains a wide variety of features, including demographic details, symptom records, and medical history, which improves the study.

Limits:

Data Imbalance: Initially, the dataset showed an imbalance with more lung cancer cases than controls. This could bias the prediction models towards overpredicting lung cancer.

Data Size: The small size of the data can improve developing machine learning algorithms but may limit the reliability of the models for larger datasets.

Key findings:

Patients with pre-existing chronic illnesses and chest pain are more likely to have lung cancer diagnosed. Strong links for symptoms: **shortness of breath**, **smoking history** and **Allergies.**

correlations insights:

0.33 allergies

0.18 Yellow fingers

0.29 alcohol consumption

0.25 Wheezing

0.26 Shortness of breath

Aspect	**Primary Data**	**Secondary Data**
Source	Survey of professionals	Kaggle dataset, referenced in Dritsas and Trigka's study
Key Variables	Age, Gender, Specialization, Symptomatic Indicators, ML Efficacy	Demographics, Symptoms, Medical History
Sample Size	34 responses	309 records
Purpose	Gather expert insights on symptoms and ML models	Investigate symptom relationships for lung cancer prediction
Merits	Direct expert insights, customizable	Comprehensive, detailed, rich in variety
Limitations	Sample bias, subjective responses	Data imbalance, limited size

4. Research Approach and Methodologies

This part includes a full overview of the research methodologies used in this paper. The "Research Onion" model is considered to highlight the layers of research decisions made including the philosophy, strategy, strategies, choices, time spans, and procedures used. It specifically discusses the application of quantitative methodologies, the choice of machine learning algorithms, and the methodical processes for data preprocessing, model training, and evaluation. This comprehensive approach ensures a strict and structured investigation of lung cancer prediction built on symptomatic indicators.

4.1 Onion Model

The Onion Model presents a detailed structure to conduct research as well as an organized approach to remove layers on research philosophy, approaches, decisions, and methods. This model is crucial for logic and systematically organizing the research process.

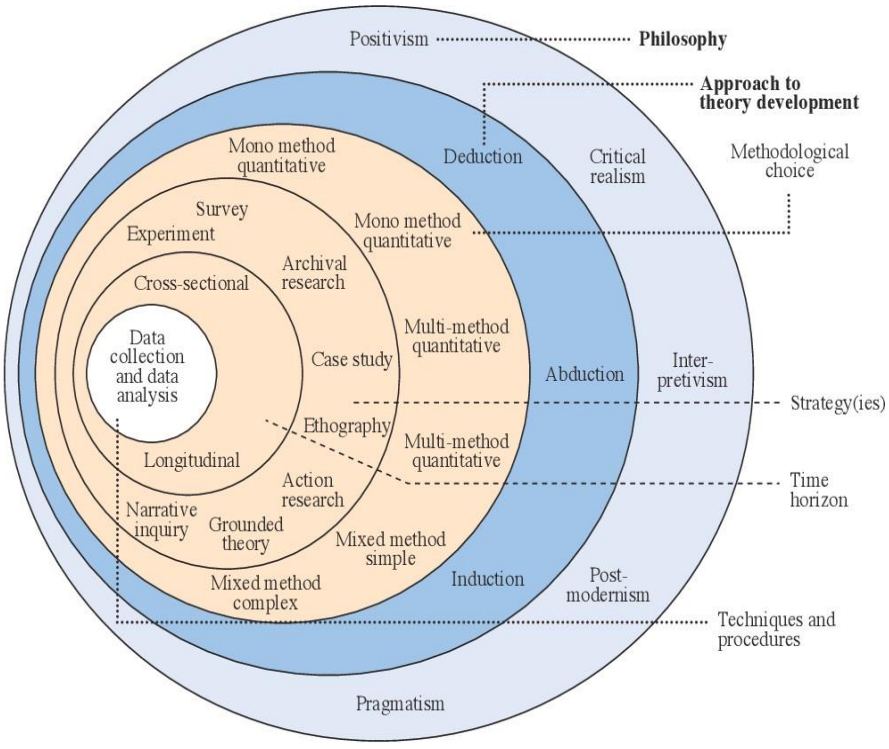


Figure 1: The structure of the Onion Research Model [1].

4.1.1 Philosophy

The research onion's philosophy layer includes positivism, interpretivism, critical realism, and pragmatism, which affect knowledge representation and learning. In this research Positivism focuses on direct proofs and facts and apparent occurrences, choosing quantitative methods and statistical analysis for evaluating objectives This philosophical perspective is suitable since it offers objectivity and consistency in assessing the relationships between lung cancer and symptomatic indicators.

4.1.2 Theory Development Approach

The theory development is about addressing how theories are constructed and tested method uses deduction, induction, and abduction. The research applies deductive reasoning to test lung cancer prediction theories against evidence to validate predictive models.

4.1.3 Methodological Choice

The methodological choice layer involves selecting among quantitative, qualitative, or mixed methods. The research conducted selecting a quantitative methodology to enable accurate measurement and robust statistical analysis for investigating connections between features relevant to lung cancer prediction. Focusing just on numerical data and statistical analysis, a mono-method quantitative method was applied. This decision ensures equality and accuracy, so enabling an obvious understanding of the correlations between features and the prediction model effectiveness.

4.1.4 Research Strategy

The research strategy involves choosing a data gathering approach, such as surveys, experiments, case studies, etc. This study combines a survey for primary data collecting with secondary data analysis, incorporating ideas from field experts and detailed data analysis to increase the accuracy of predictive models for lung cancer

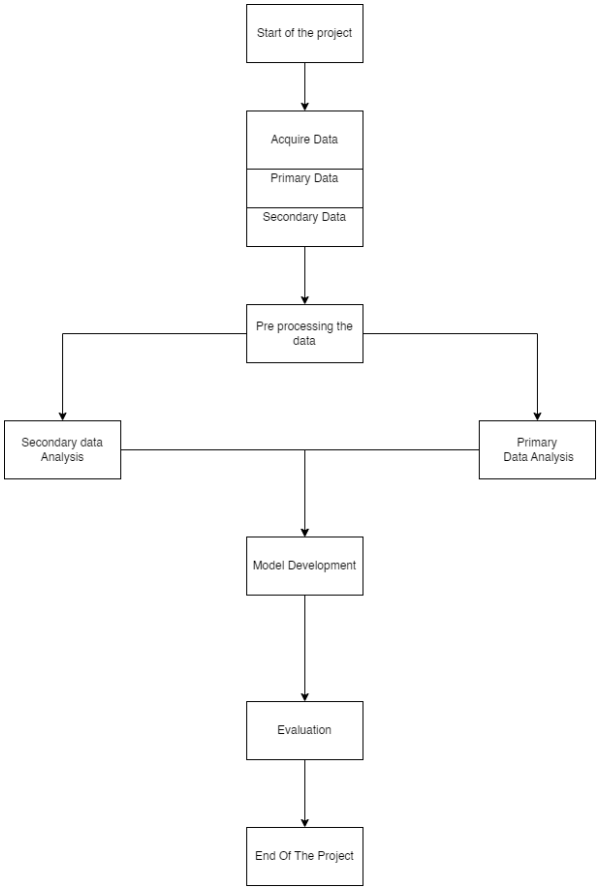
4.1.5 Time Horizons

Time Horizons in the Onion Model relate to the timeframe of the investigation, dividing between cross-sectional and longitudinal studies. Cross-sectional studies study data at a particular point in time, while longitudinal studies observe changes over time. According to my research A Cross-sectional design is used quickly to analyze current correlations and patterns in lung cancer symptomatology and prediction modeling.

4.1.6 Techniques and Procedures

This layer covers the specific procedures used for data gathering, analysis, and validation, such as data preprocessing, development of models, and evaluation approaches. Effective data management strategies and advanced statistical approaches are necessary for handling and evaluating huge datasets. In this research, comprehensive Data Preprocessing and Model Development approaches using comparison between 3 models and SMOTE technique are implemented to ensure the predictive models created are both robust and reliable.

4.2 Research Methodology



Start of the project
Project start-up requires setting research objectives and hypotheses. This critical phase establishes the foundations for the research.

Acquire Data:
Primary and secondary data are divided in data acquisition. Surveys aimed at medical and data science experts provide primary data to help understand lung cancer symptoms and diagnosis techniques. secondary data derived from comprehensive data on lung cancer cases and associated symptoms.

Primary Data
Conducting a detailed survey aimed at oncology, radiology, pulmonology, and data science professionals offered useful research insights. Expert opinions were directly collected to identify major themes and validate data from other sources. The responses revealed key patterns, confirmed symptom significance hypotheses.

Secondary Data:
discovering and securing supplementary datasets that support the main survey data was essential. in the context of the research i seek to locate dataset that has an information about lung cancer symptoms such as smoking, chest pain and other associated common factors related to the diagnosing the lung cancer Having secondary data is necessary, and the reason for this is because the primary data is not always enough, and sometimes it needs to be supported by secondary data. using pre-collected secondary data saves significant time and resources, particularly when it aligns closely with the research objectives

Preprocessing Data:
Data preprocessing is an essential step in the research, focused on cleaning and modifying the raw data to ensure its quality and validity for analysis. In this phase, I normalized the features to a consistent range to reduce any potential biases in the model's performance. Additionally, addressing class imbalance with approaches like SMOTE oversampling proved essential to guaranteeing a fair representation of all classes in the dataset. These preprocessing methods are crucial, since datasets often require extensive modification to be useful for predictive modeling. thorough the collection of data ensures it is optimally

prepared for the training and testing phases, consequently highlighting this step as one of the most critical in the overall research workflow.

Primary data analysis:
The analysis of survey data involved identifying key patterns and related symptomatic indicators of lung cancer. This stage was essential for separating a huge variety of expert opinions into practical findings and examining how symptoms affect lung cancer risks. Each theme was carefully examined to identify the most important symptoms reported by healthcare professionals.

Secondary Data Analysis:
Analyzing secondary datasets requires exploring the data to reveal trends, correlations, and patterns relevant to the research objectives. This step provides additional insights into the topic under research and enhances the overall analysis. Analyzing this data can be used many tools and methodologies like Power Bi, Python This step is essential in the research since with it we will be able to discover and analyze the features, know more about the dataset, and what are the important elements that should be taken into consideration. Also, examining it could help in the preprocessing part in advance.

Model Development:
Constructing predictive models is a vital part in the research, when machine learning algorithms are applied to the preprocessed data to predict lung cancer potential according to symptomatic indications. For this purpose, I deployed strong algorithms such as Logistic Regression, Decision Trees, and Random Forests. Logistic Regression, known for its efficiency in binary classification applications. Cross-validation procedures were performed to ensure the model's generalizability across diverse subsets of data. This step is vital in my research as these prediction models create the primary framework through which the correlations between symptoms and lung cancer risks are statistically assessed, making this phase crucial for meeting the research objectives.

Evaluation:
The evaluation process looks into the overall plan and study objectives completion. These covers looking at data collecting, analysis, development of models, according to the study, statistical techniques can identify lung cancer from symptoms, accordingly, enabling early identification and improved patient outcomes. The findings explain diagnosis techniques and suggest next studies to enhance them.

End of Project

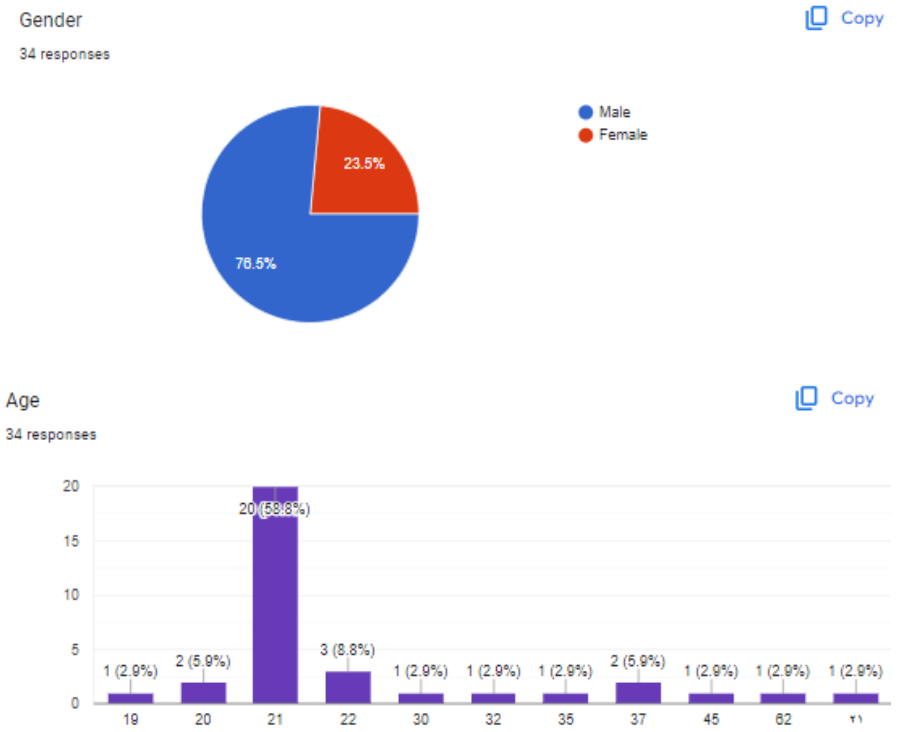
5 Results and Discussion

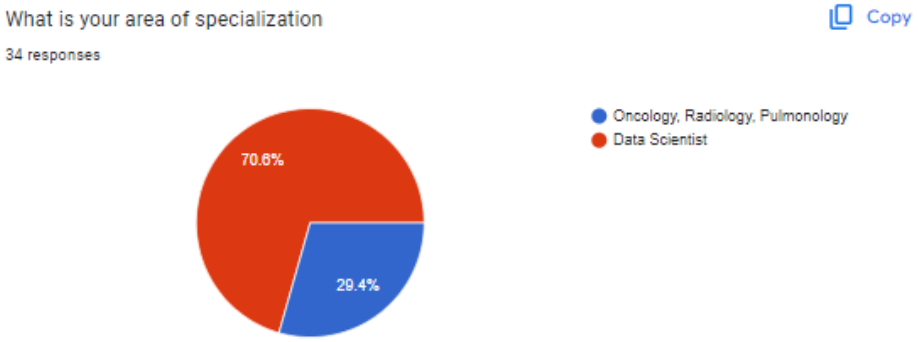
This section provides a comprehensive discussion for the analysis and provide the results for both analysis primary and secondary and the limits and merits and discussion about the model development

5.1 Primary data analysis

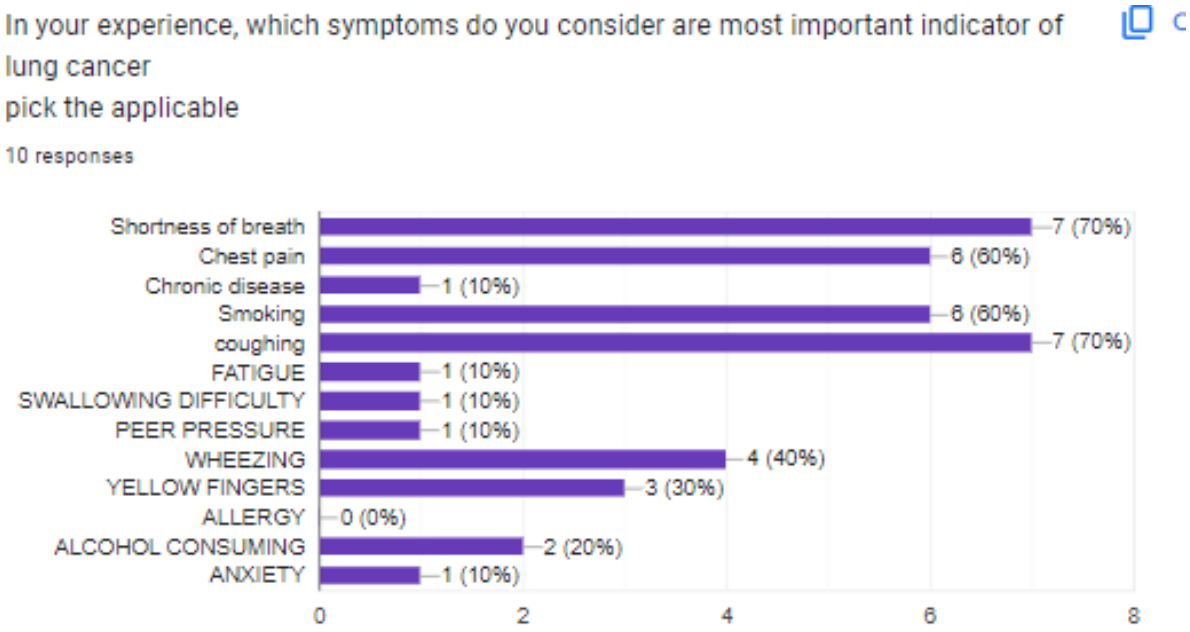
This research gathered primary data from professionals in oncology, radiology, pulmonology, and data science through the use of a survey. The survey aimed to find the prevalence and importance of certain lung cancer symptoms. Examined also was secondary data to support the primary data results.

The demographic data show that among the respondents, 76.5% were men and 58.8% were 21 years of age. The professional backgrounds show that 70.6% were data scientists and 29.4% were oncologists, radiologists, and pulmonologists. This distribution shows a remarkable involvement of data scientists in the survey, which suggests the important role of data science in the research.

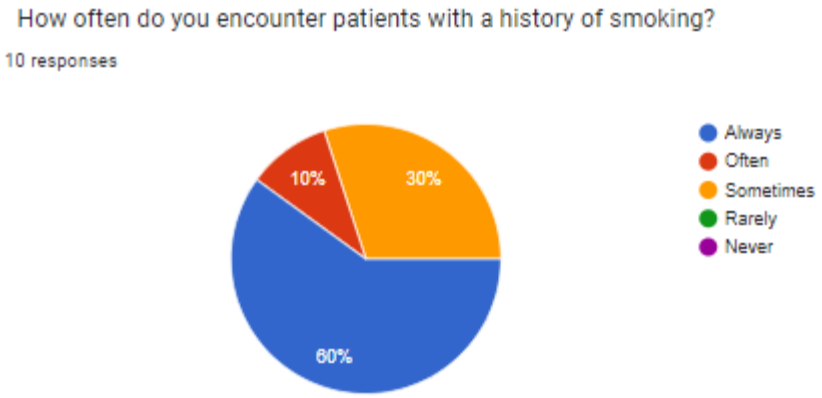




The most typically indicated symptoms were coughing, chest pain and smoking history. Medical doctors identified lung cancer mostly depending on these symptoms. This is in line with the body of studies highlighting these symptoms as major indicators of lung cancer.

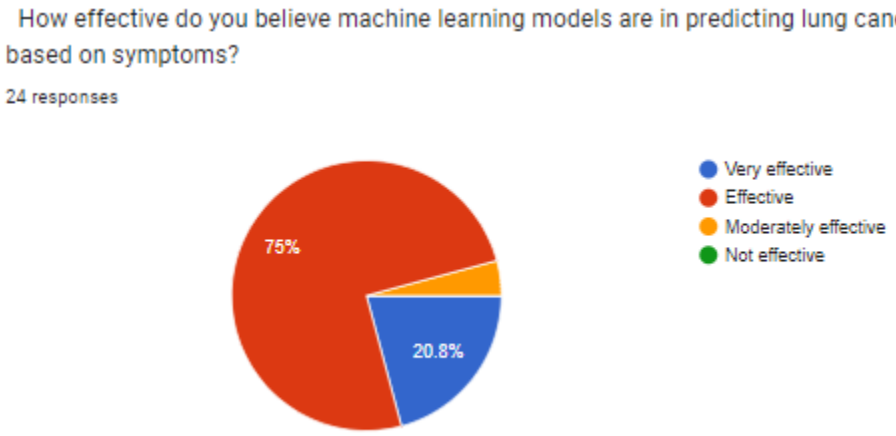


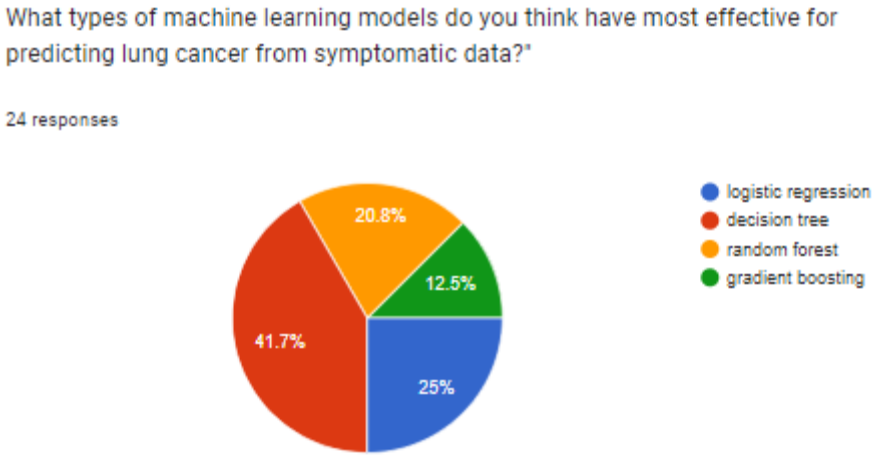
The most common symptoms were shortness of breath and coughing, followed by chest pain and smoking history. Medical professionals identified lung cancer by shortness of breath, coughing, and chest



Given 60% of respondents indicating they "always" encounter patients with a history of smoking and 30% reporting "often," revealed most patients encountered by medical professionals were smokers. This highlights the significant correlation between smoking and lung cancer and the need of smoking history as a critical symptomatic indicator in diagnosis of the lung cancer.

Within the data scientists, 79.2% have either considered or used machine learning models in medical diagnosis. Especially logistic regression (25%) and decision trees (41.7%), they showed great trust in the capacity of machine learning models to predict lung cancer based on symptoms.





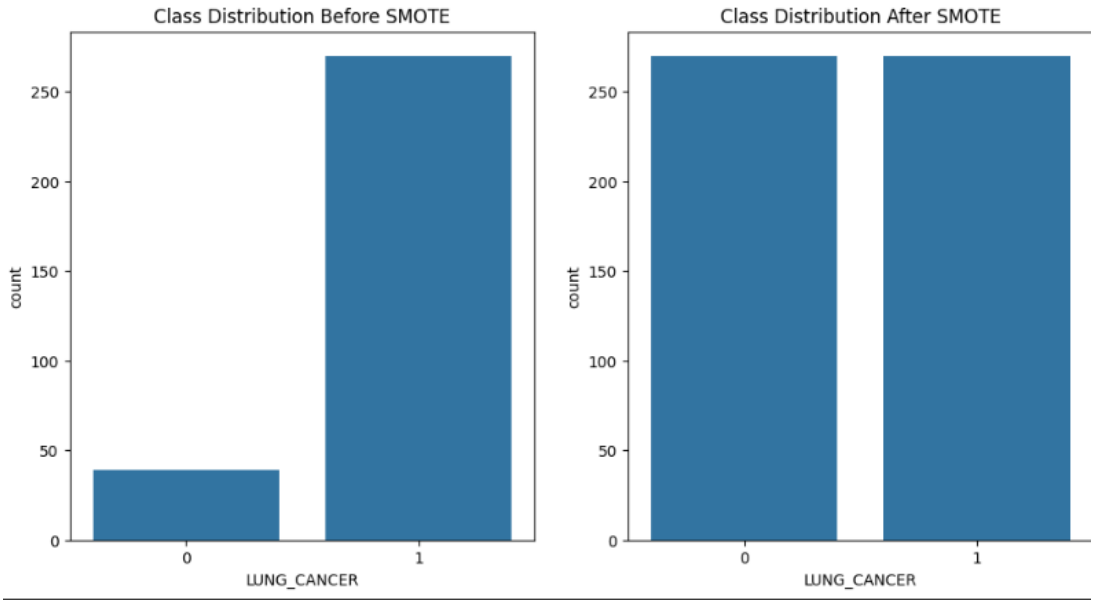
Responding to the "What improvements in data collecting or model training techniques do you think could enhance the accuracy of lung cancer predictions?" highlighted some essential recommendation:

- Balancing datasets to reduce bias.
- Increasing the quantity and quality of medical data labels.
- Incorporating a variety of data types (imaging, genetic information, health records).
- Using advanced machine learning methods and transfer learning.
- Ensuring continuous data collection over time.

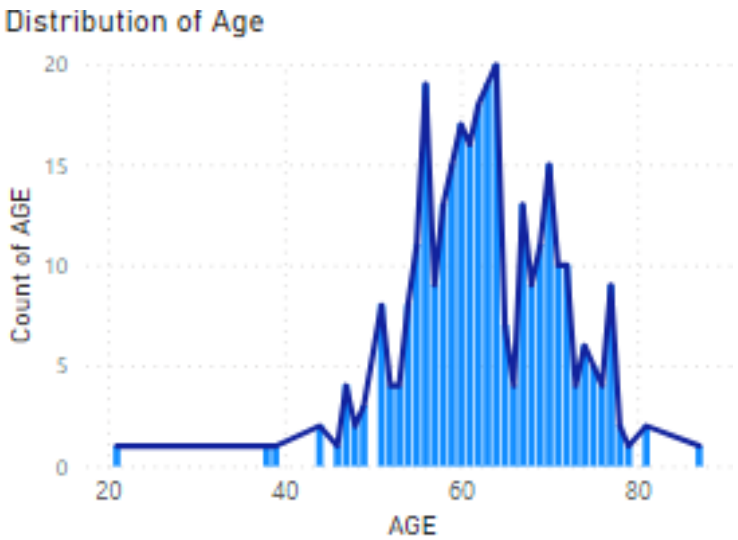
5.2 Secondary Data Analysis:

The dataset used in this research provides a comprehensive collection of patient data, including symptoms indicators for lung cancer, and shows an interesting imbalance with 270 diagnosed cases compared to 39 undiagnosed. This imbalance is critical for the analytical methods as it considerably affects the training and accuracy of predictive models.

in the dataset preprocessing for lung cancer prediction, major measures were performed to refine data their suitability for predictive modeling. Important transformations included transforming category variables like 'SMOKING', 'YELLOW FINGERS', and 'ANXIETY' from numerical to binary formats to remove any determined ordinal relationships. The essential 'LUNG_CANCER' target variable was also converted from textual to binary format to match classification algorithm criteria.

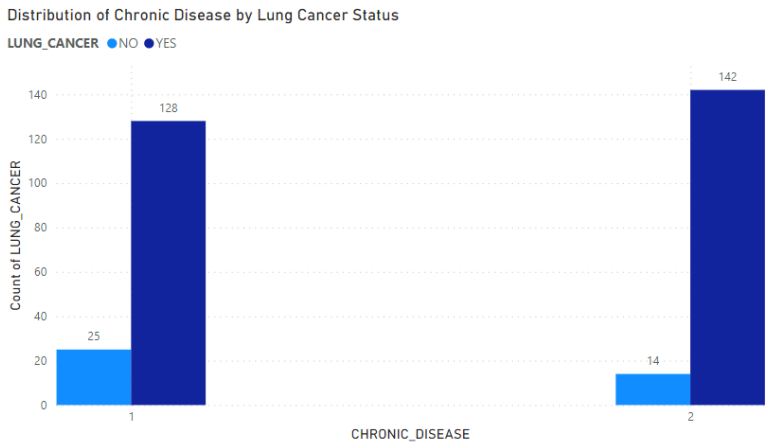


To rectify the initial class imbalance identified in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized. This strategy increased the minority class by synthesizing fresh samples, thus balancing the dataset and boosting model training efficacy. Visualizations of class distributions before and after using SMOTE indicated a significant rectification of class skewness, necessary for lowering model bias and boosting the robustness of lung cancer forecasts. These preprocessing processes are crucial for constructing a solid diagnostic model that operates efficiently across various patient data.

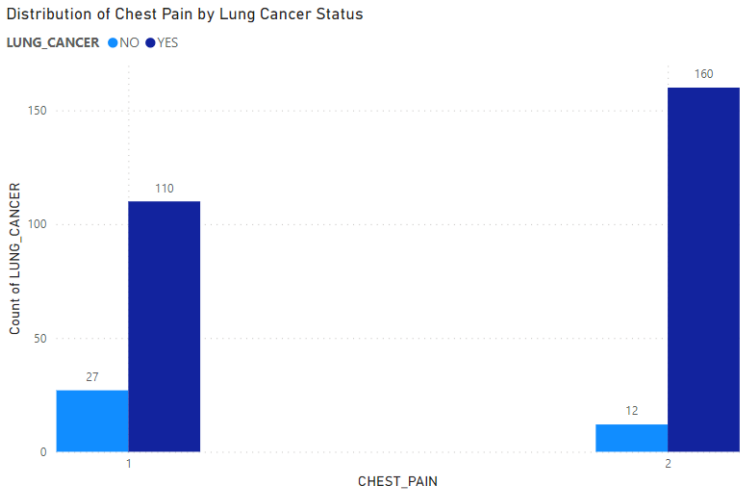


There were more people in the 50–70 age range but the age distribution ranged from roughly 20 to 80. Lung cancer was more common among patients with chronic diseases,

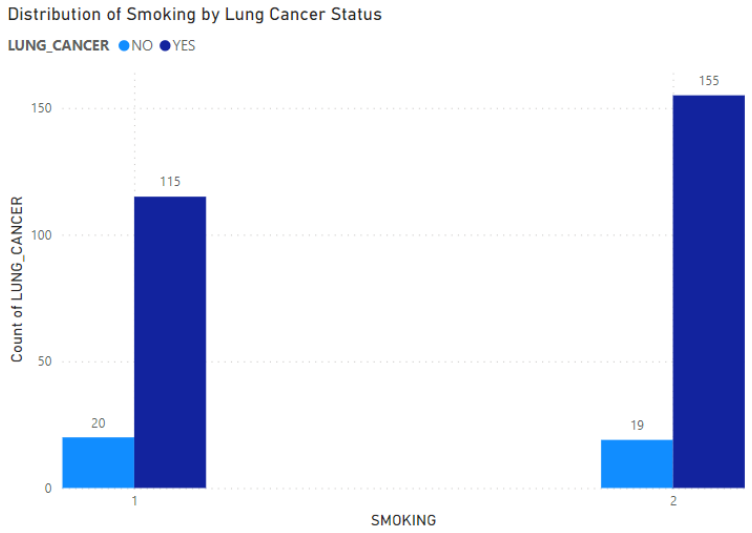
chest pain, and shortness of breath according to the analysis medical indicators. Furthermore, strongly linked to lung cancer was smoking status; smokers have a noticeably higher prevalence of lung cancer patients.



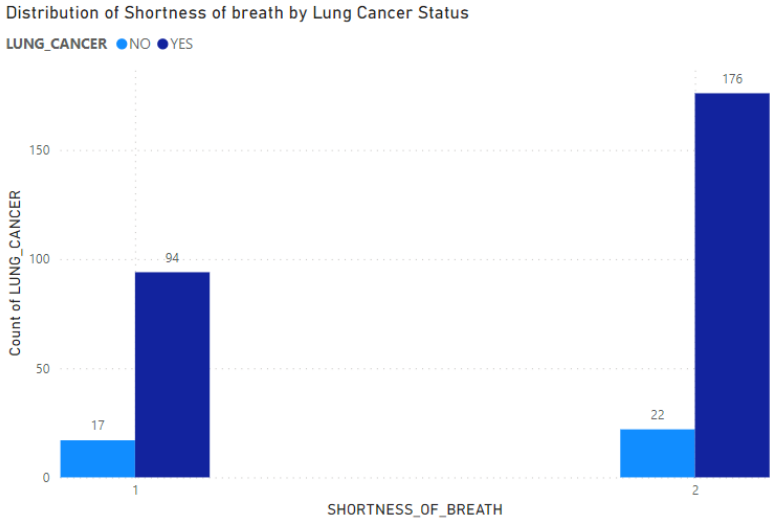
The presence of chronic diseases also shows a notable correlation with lung cancer. Patients with chronic diseases have a higher incidence of lung cancer (142 cases) compared to those without chronic diseases (14 cases).



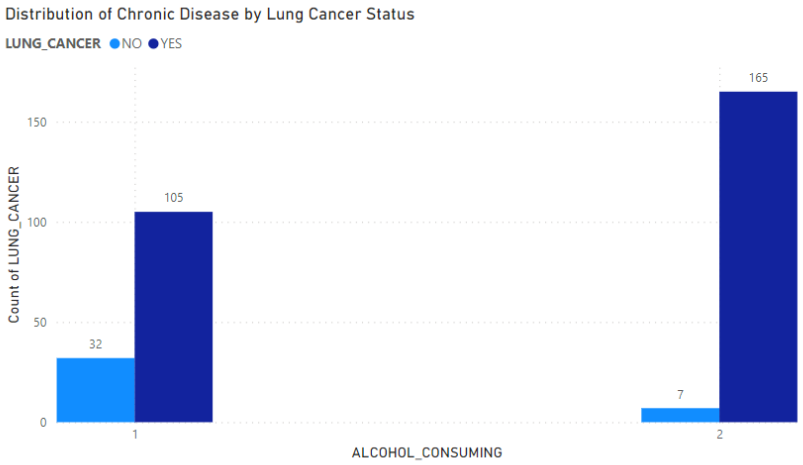
Chest pain is significantly associated with lung cancer. Patients experiencing chest pain are much more likely to have lung cancer (160 cases) compared to those without chest pain (12 cases).



Smoking history is a well-known risk factor for lung cancer. Patients with a history of smoking have a significantly higher incidence of lung cancer (155 cases) compared to non-smokers (19 cases).

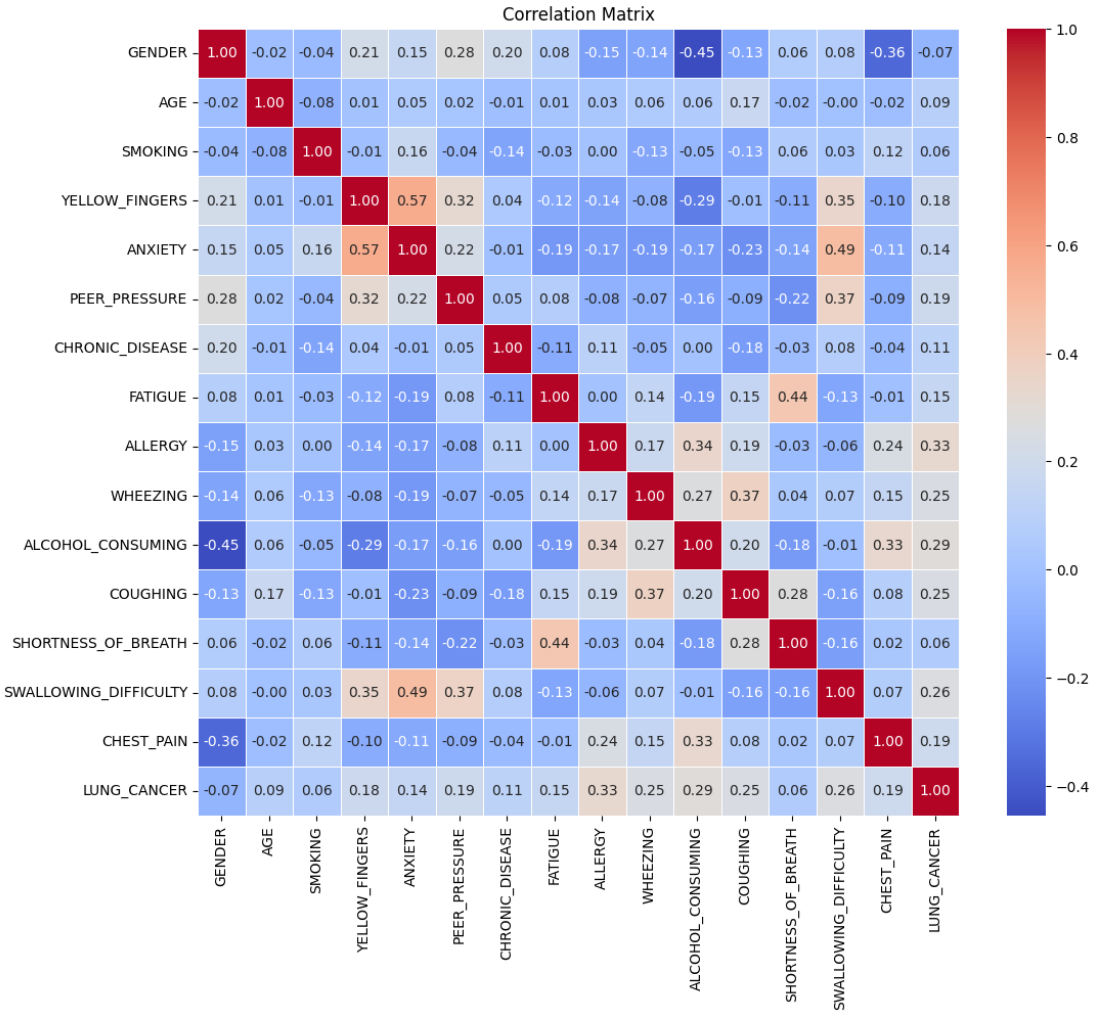


Shortness of breath is another significant indicator. Patients who experience shortness of breath have a much higher incidence of lung cancer (176 cases) compared to those who do not (22 cases).



The chart shows a strong correlation between alcohol consumption and lung cancer. Patients who consume alcohol have a much higher incidence of lung cancer (165 cases) compared to non-drinkers (7 cases).

Validating the results of the primary data, the secondary data analysis reveals the significance of some symptoms in lung cancer prediction. Important indicators that should be included in the diagnosis and prognosis of lung cancer are the five essential elements exposed by this study: consumption of alcohol, chest pain, chronic disease, shortness of breath, and smoking history. These features not only match the evaluations of the medical experts but also provide a strong foundation for developing prediction models applying machine learning.



used a correlation matrix in the research conducted to help better understand the relationships between certain symptoms and lung cancer. The matrix shows the strength and direction of the interactions between lung cancer and the symptoms. With a 0.33 high positive association, coughing became the most significant predictor, therefore confirming the survey findings stressing coughing as a crucial warning of lung cancer. Likewise, consistent with the core data results, alcohol intake (0.29), wheezing (0.25), and swallowing difficulty (0.26) all showed high correlations. Though with less correlation values, chest pain (0.15) and shortness of breath (0.19) highlight even more their relevance as markers. Therefore, the correlation matrix confirms our initial data analysis and shows the necessary symptoms that should be included in lung cancer prediction and diagnosis.

Model Development:

Developed and reviewed three predictive models: Random Forest, Decision Tree, and Logistic Regression. The following table summarizes the performance measures for these models:

Model	Mean CV Accuracy	Std CV Accuracy	Test Accuracy	F1 Score
Logistic Regression	0.9655	0.0328	0.9537	0.9573
Decision Tree	0.9216	0.0491	0.9444	0.9483
Random Forest	0.9585	0.0266	0.9537	0.9573

The findings of this research fit very well with the aims and questions of the research. The main goal was to use machine learning models to investigate, using several symptomatic signs, the predictive power for lung cancer. Survey data and secondary dataset analysis support the results, which demonstrate that coughing, shortness of breath, and chest pain are important symptoms of lung cancer.

The deployment of SMOTE in particular proved essential in fixing the class imbalance in the dataset, which enhanced the predictive model performance and dependability.

5.3 How the Results Meet the Research Question and Objectives

The research question involves using machine learning techniques to project lung cancer potential based on symptomatic indicators. Detection of important indicators, building predictive models, and improving early diagnosis accuracy are among the objectives.

The most significant symptoms reported in the primary data analysis were coughing, shortness of breath, and chest pain. The secondary data analysis confirmed these results with modestly favorable correlations for coughing (0.33), alcohol consumption (0.29), wheeze (0.25), and swallowing difficulty (0.26). These relationships highlight the importance of these symptoms in lung cancer prediction since they line up with the primary conclusions

Three prediction models Logistic Regression, Decision Tree, and Random Forest were effectively developed in this research. These models proved highly solid and accurate for predicting lung cancer. Considered by the respondents as successful, logistic regression and random forest displayed outstanding test accuracy and F1 values. although the Decision Tree model also performed well. These findings help to justify using these models for predictive diagnosis.

Strong model training and enhanced early lung cancer diagnosis rely critically on the preprocessing procedures, especially the deployment of SMOTE to fix the class imbalance. SMOTE greatly improved the applying and reliability of the algorithms by balancing the dataset, therefore predicting lung cancer diagnosis. The positive feedback of data scientists on the performance of machine learning models assist in highlighting the objective of the research of improving early detection. Their suggestions on better data collecting and model training approaches highlight the possibility for always improving diagnosis accuracy.

To conclude, combining primary and secondary data along with using advanced machine learning methods led to an efficient and reliable method of lung cancer prediction.

5.4 Merits and limits:

Merits:

- Integration of primary and secondary data gave a strong basis for the research, therefore guaranteeing a diverse range of outcomes.
- Using SMOTE greatly improved the balance of the dataset, which enhanced the model's performance.
- Particularly the Logistic Regression and Random Forest models, the created models show great reliability and accuracy.
- High Predictive Accuracy: The developed models demonstrated high accuracy and F1 scores, indicating reliable lung cancer prediction.

Limits:

- Sample bias: The main demographic bias of the data for younger and male respondents and data scientist might limit the application of the results.
- Class Imbalance: The original dataset imbalance could still affect the bias of the model even with SMOTE applied.
- Subjectivity in Responses: The survey answers could reflect personal biases and opinions of the medical experts.
- Small Sample Size: The main poll attracted just 34 responders, which could not accurately represent the larger medical and data science community.
- Limited Scope of Symptoms: The restrictions of the dataset might have caused additional possibly relevant symptoms to be missed even as notable ones were found.

6 Conclusion and Recommendations

This section summarizes the research results. It recommends future study subjects to help to enhance early diagnosis and patient outcomes and provides strategies for enhancing data collecting and machine learning techniques.

6.1 Conclusion

The research was aimed primarily at using machine learning methods to predict lung cancer potential based on symptomatic indicators such coughing, smoking, and chest discomfort...etc. to help with early and accurate lung cancer diagnosis, the study aimed to identify significant relationships within symptom data, which could enhance patient outcomes. Results of primary and secondary data analysis confirmed the importance of particular symptoms for lung cancer prediction. The most important signs of lung cancer, according to primary data gathered from a survey of experts in oncology, radiology, pulmonology, and data science, coughing, shortness of breath, and chest pain. Secondary evidence confirmed this by showing a strong correlation between these symptoms and lung cancer cases. Using machine learning techniques mostly logistic regression and random forest great accuracy in lung cancer prediction based on these symptoms was shown. Important in improving the performance and reliability of these models were the preprocessing phases. The application of the SMOTE was crucial in addressing class imbalance, efficiently enhancing the robustness of the models. The research mostly achieved its goals by proving that

enhanced preprocessing methods can increase model performance and that machine learning models can consistently predict lung cancer depending on symptomatic indicators.

6.2 Recommendations:

Based on the findings there's several recommendations that can be made to improve lung cancer prediction and early diagnosis such as:

- Enhancing data collection requires a more accurate and comprehensive dataset for using different types of demographic profiles and medical indicators helps to extend the type of data obtained. Including complete instances could help to enhance the dataset by adding more diversity of features and might provide new ideas on lung cancer indicators.
- Use modern machine learning methods Training with fresh data and regular updates help to keep predictive models relevant and accurate. Moreover, combining several data sources such as imaging, genetic data, and electronic health records allows an overall picture of every patient's health, which enhances the prediction ability of the models. Using complex relationships between several symptoms and lung cancer mostly depends on applying advanced feature engineering techniques.
- Early diagnosis and effective treatment require a solid knowledge of how lung cancer develops, which can be shown by long-term study tracking of symptom change over time. Moreover, helping to adjust methods of treatment in early phases is knowledge about the dynamics of the disease obtained by comprehensive long-term studies.

6.3 Future Work

The primary goal of future work should be expanding the dataset by including more diversity of symptoms and patient profiles. Examining the combination of devices that collect real-time data and health monitoring systems could potentially provide significant study of symptom development and improve early identification capacity. Working together, medical professionals and data scientists should continuously improve and review prediction models to ensure their practical applicability in medical applications. Further research might also include building and testing more complex models able to integrate multi-modal data inputs, for aggregating genetic data with imaging and symptom data. This approach could greatly increase the accuracy and reliability of lung cancer predictions. Maintaining the efficacy of the prediction models in clinical practice will rely on constant improvements and changes depending on new data and technological advancements.

7 Reflections

This section evaluates alternative approaches, discusses their merits and limitations, and offers steps for next research to enhance the efficacy of the study based on the utilized research methodology.

7.1 Selected Research Methodology

Although challenging, choosing and implementing machine learning lung cancer prediction techniques has been successful. Examining the research process, a quantitative, mono-method approach enhanced data analysis. This was absolutely essential to guarantee statistical validity and reliability.

Significantly important for predictive modeling, this method could manage large datasets and execute strong statistical analysis. The technique has disadvantages, although. Qualitative data might have uncovered subtleties missing from numerical data. SMOTE drawbacks were highlighted when it was used to fix biases in imbalanced datasets.

7.2 Alternative Research Methodologies

Given the results, different approaches could have offered unique perspectives. For instance, a mixed-method approach integrating qualitative and quantitative data could have yielded more comprehensive research. The qualitative interviews of doctors could point out lung cancer symptoms and diagnosis difficulties.

Future research could benefit from looking more at more advanced machine learning and deep learning techniques as Convolutional Neural Networks (CNNs), for images data, thus improving prediction accuracy even more. Early detection models would be improved and significant new viewpoints on the dynamics of lung cancer development would be supplied by following symptom development over time utilizing long-term investigations. More rapidly the most predictive elements could be found using enhanced feature engineering tools including automated feature selection and extraction approaches. Maintaining relevance and accuracy in models calls for continuous training using new data. Predicting ability could be greatly improved by adding real-time data from smartwatches or health monitoring systems in a procedure for regular updates.

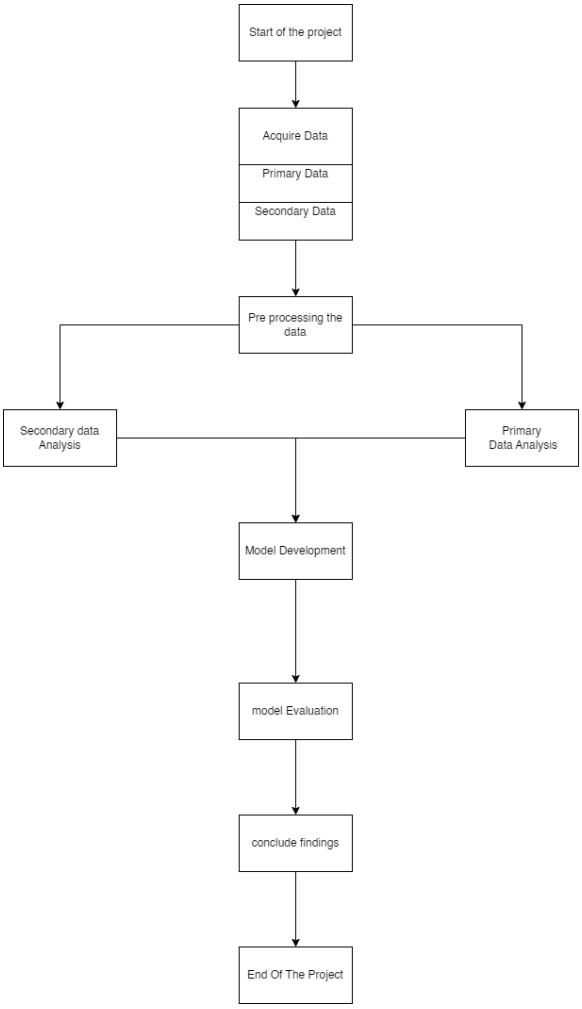
Important lessons from the present method include data diversity and integration. Including qualitative data could have improved the research by providing context for the numerical findings and maybe by emphasizing fresh expected aspects.

7.3 Recommended Actions and Future Considerations

- Increase the dataset to include more varied demographic characteristics and thorough patient histories. The models' generalizability and robustness will increase.
- Add qualitative insights from medical specialists by means of polls and interviews, therefore augmenting quantitative data and offering a more complete knowledge.
- Update models often with fresh data and investigate sophisticated machine learning approaches including deep learning and ensemble methods to improve prediction accuracy.
- Long-term research helps to better understand the illness dynamics and track symptom progression, which enhance guiding early diagnosis and treatment plans.

7.4 Recommended Methodology

The updated technique uses an expanded approach to guarantee strong and accurate models of lung cancer prediction. This revised approach underlines advanced preprocessing methods, diverse data collecting, and ongoing model improvement. The steps included are shown on the flowchart below.



Acquire Data:

Primary Data: Target medical and data science experts for interviews and survey responses to obtain comprehensive understanding of lung cancer symptoms and diagnosis approaches. Combining quantitative and qualitative techniques, this method increases the data's richness.

Secondary Data: Get comprehensive records from medical facilities and public databases. These databases will comprise a wide range of information including extra medical indicators and demographic profiles also such as can combine imaging and genetic data to provide an accurate prediction.

Data Preprocessing:

Handling missing values and outliers assists to guarantee that the dataset is correct and complete. Although using modern techniques, feature engineering finds crucial features for the prediction models. If the Data is imbalanced could be addressed by means of SMOTE or other advanced techniques to handle the imbalanced, therefore balancing the dataset and so minimizing the biased of the model towards the majority class.

Data analysis:

From the expert survey answers and interview insights, find key patterns and symptomatic markers. This combination method gives more background and helps confirm results from secondary sources. Using Power BI and Python, evaluate relationships, trends, and patterns inside the entire dataset.

Model Development:

selecting appropriate algorithms, such as Deep Learning models, Random Forest, and Logistic Regression. Also integrate Cross-validation methods are used during model training to make sure the models perform well when applied to new data.

Model Evaluation:

To assessing models' performance use metrics including recall, accuracy, F1 score,

and precision. Although training models with fresh data on a regular basis to keep them accurate and relevant.

Conclude key findings:
Research results show that, depending on symptoms, statistical methods are rather effective in predicting lung cancer. The results highlight the importance of the study to the understanding of diagnostic techniques and suggest additional research to create these approaches, so aiming to enhance lung cancer diagnostics outcomes.

In conclusion, by increasing the amount and variety of the dataset, covering a wider range of features, and combining quantitative surveys with qualitative interviews for primary data collecting, the new technique strengthens the old one. These modifications guarantee a more accurate and complete dataset, so allowing a deeper understanding of lung cancer prediction. Advanced preprocessing methods such as SMOTE solve data imbalances, therefore improving the model performance. Integrating multimodal input and regular model upgrades ensure the models stay relevant and accurate over time. Long-term research can help to clarify disease processes and enable early identification and intervention by means of which This comprehensive strategy not only solves current research objectives but also provides a good basis for next lung cancer prediction studies.

8 References

Ahmad, A. S. (2020). A new tool to predict lung cancer based on risk factors. Informatics in Medicine. *Elsevier*.

Ausawalaithong Worawate, T. A. (2018). *Automatic Lung Cancer Prediction from Chest.*

Dritsas, E. &. (2019). Lung Cancer Risk Prediction with Machine Learning Models. *Cancers, 11(11), 1595. MDPI*.

Nemlander, E. R. (2022). Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, former smokers and current smokers. *PLOS ONE*.

Radhika P R, R. V. (2019). A comparative study of lung cancer detection using machine learning algorithms.

Salaken, S. M. (2021). Lung Cancer Classification Using Deep Learned Features on Low Population Dataset. *Canadian Journal of Electrical and Computer Engineering, 44(1), 1-9. IEEE.*

Sowjanya, M. M. (2016, January). Lung Cancer Detection in Chest X - Ray Image. *international Journal of Research and Analytical*, pp. 88-96.

8 Appendix

survey link: Press Here

Predictive Analysis of Lung Cancer Susceptibility Based on Symptomatic Indicators

The purpose of this survey is to get perspectives on the application of symptomatic indicators and machine learning models in lung cancer prediction from experts in the fields of machine learning and lung cancer. The research on enhancing the detection of lung cancer will gain insight from your answers. I appreciate you taking part.

Please take note that all answers will maintained confidential.

يمكنك تسجيل الدخول إلى Google لحفظ مستوى التقدم. مزيد من المعلومات

* تشير إلى أنّ السؤال مطلوب

* Age

إجابتك

* Gender

Male

Female

* What is your area of specialization

Oncology, Radiology, Pulmonology

Data Scientist

(For Medical Specialists)

?How often do you encounter patients with a history of smoking

- Always
- Often
- Sometimes
- Rarely
- Never

* In your experience, which symptoms do you consider are most important indicator of lung cancer
pick the applicable

- [ ] Shortness of breath
- [ ] Chest pain
- [ ] Chronic disease
- [ ] Smoking
- [ ] coughing
- [ ] FATIGUE
- [ ] SWALLOWING DIFFICULTY
- [ ] PEER PRESSURE
- [ ] WHEEZING
- [ ] YELLOW FINGERS
- [ ] ALLERGY
- [ ] ALCOHOL CONSUMING
- [ ] ANXIETY

* Which imaging technique do you primarily use for diagnosing lung cancer

- X-ray,
- CT Scan
- MRI
- اخرى:

Data Scientist

(For Machine Learning/Data Science Specialists)

* Have you ever used or considered using machine learning models in medical ?diagnostics

- Yes
- No

* How effective do you believe machine learning models are in predicting lung ?cancer based on symptoms

- Very effective
- Effective
- Moderately effective
- Not effective

* What types of machine learning models do you think have most effective for "?predicting lung cancer from symptomatic data

- logistic regression
- decision tree
- random forest
- gradient boosting

What improvements in data collection or model training techniques do you think ?could enhance the accuracy of lung cancer predictions

اپنا جواب