Data Driven Lung Cancer Staging and Prediction

A PROJECT REPORT

18CSE422T – INTRODUCTION TO MACHINE LEARNING

(2018 Regulation)

III Year/ VI Semester

Academic Year: 2023 -2024

Submitted by

SANJAY J. K. [RA2111003010441]
YOHAN VERGIS VINU [RA2111003010417]
AKKINENI NAGA CHARITESH [RA2111003010364]

Under the Guidance of

Dr.A.Revathi
Associate Professor
Department of Computational Intelligence

in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE ENGINEERING - CORE

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

SCHOOL OF COMPUTING
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR- 603 203

MAY 2024

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that **18CSE422T - INTRODUCTION TO MACHINE LEARNING** project report titled “**Data Driven Lung Cancer Staging and Prediction**” is the bonafide work of “**SANJAY J. K. [RA2111003010441], YOHAN VERGIS VINU [RA2111003010417], AKKINENI NAGA CHARITESH [RA2111003010364]”** who carried out the task of completing the project within the allotted time.

SIGNATURE
Dr.A.Revathi
Course Faculty
Associate Professor
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur

SIGNATURE
Dr.M.Pushpalatha
Head of the Department
Professor
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur

ABSTRACT

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, necessitating the development of accurate staging methods for effective treatment planning and patient prognosis. This project aims to integrate clinical parameters and molecular insights to improve lung cancer staging accuracy using machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM).

The project leverages a comprehensive dataset comprising clinical data such as patient demographics, tumor characteristics, and treatment history, alongside molecular data obtained from genomic profiling and biomarker analysis. Through feature engineering and selection techniques, relevant clinical and molecular features are identified and incorporated into the predictive models.

The Logistic Regression model provides a baseline for predicting lung cancer staging based on clinical parameters, while KNN utilizes the similarity of patient features to classify staging. Naive Bayes employs probabilistic reasoning to estimate the likelihood of staging given clinical and molecular features. SVM, with its ability to handle high-dimensional data, aims to optimize the decision boundary between different staging categories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

This chapter sets the stage for our exploration of data-driven approaches for lung cancer prediction and staging. We begin by outlining the current limitations of traditional methods and the critical role of early detection and accurate staging in successful patient outcomes.

Following this, we present a well-defined problem statement that pinpoints the specific challenges addressed by this project. Finally, we establish clear objectives that guide our investigation into data-driven techniques for lung cancer prediction and staging. This chapter paves the way for our journey into exploring how data analysis can revolutionize the fight against lung cancer.

INTRODUCTION

Lung cancer continues to cast a long shadow, claiming countless lives globally each year. Early detection and accurate staging are paramount for successful treatment strategies and improved patient outcomes. Traditionally, the diagnostic process for lung cancer has relied on invasive procedures such as biopsies and chest X-rays. These methods, while valuable, can be time-consuming and present limitations in terms of sensitivity and specificity.

This project ventures into the frontier of data-driven approaches for lung cancer prediction and staging. By harnessing the transformative power of machine learning and artificial intelligence, we aim to revolutionize the landscape of lung cancer diagnosis. Our approach integrates cutting-edge technical advancements with a deep understanding of the medical domain.

The technical cornerstone of this project lies in advanced image analysis techniques, specifically radiomics. This revolutionary approach allows us to extract a wealth of quantitative features from medical images, particularly CT scans. These features, often imperceptible to the human eye, hold the key to unlocking hidden insights into the characteristics of tumors. Machine learning algorithms, meticulously trained on vast datasets of labeled images, can then analyze these features and identify subtle patterns that differentiate benign from malignant lung nodules. Additionally, we will explore the immense potential of deep learning algorithms. Inspired by the intricate structure and function of the human brain, these algorithms possess the remarkable ability to learn complex, non-linear relationships within medical images. This capability could lead to a paradigm shift in diagnostic accuracy, enabling us to detect lung cancer at its earliest stages.

From a medical standpoint, this project holds immense promise. By facilitating faster and more precise detection of lung cancer, we can pave the way for earlier intervention. This translates to a wider range of treatment options, potentially including curative approaches that offer patients a significantly better chance of survival. Furthermore, accurate staging, which determines the extent of cancer spread, allows physicians to tailor treatment plans with laser-like precision. This may encompass surgery, radiation therapy, chemotherapy, or a combination of these modalities, depending on the specific stage of the disease.

By bridging the gap between cutting-edge technical tools and the critical needs of the medical domain, this project has the potential to revolutionize the fight against lung cancer. Through the powerful lens of data analysis, we aim to usher in a future where lung cancer is detected and managed more effectively, ultimately leading to improved patient prognoses and enhanced survival rates.

Traditionally, predicting and staging lung cancer involved a multi-step process with inherent limitations. Doctors evaluated patients for symptoms like persistent cough, weight loss, and chest pain, but these can be misleading as they occur in various respiratory illnesses. Chest X-rays, while widely available, often miss early-stage lung cancer due to their lower sensitivity for detecting small nodules. Low-dose CT scans, though more sensitive, can lead to false positives,subjecting patients to unnecessary biopsies. Definitive diagnosis relied heavily on biopsies obtained through bronchoscopy or needle aspiration, which can be uncomfortable or even risky for some patients, potentially delaying diagnosis or discouraging them from undergoing the procedure. Staging, which determines treatment options, traditionally involves imaging techniques and sometimes even exploratory surgery. However, these methods can be subjective in interpretation by radiologists, leading to potential misdiagnosis. Additionally, surgery adds another layer of invasiveness to the diagnostic process. These limitations in traditional methods highlight the need for improved approaches. Data-driven methods, as explored in this project, have the potential to address these shortcomings by offering a potentially more objective, less invasive, and hopefully more accurate approach to lung cancer prediction and staging, ultimately leading to earlier diagnoses and improved patient outcomes.

Despite advances in treatment modalities, lung cancer remains a formidable challenge in modern medicine. Its high mortality rate is often attributed to late-stage diagnoses, where treatment options are limited and prognosis is poor. The traditional reliance on symptomatic presentation and imaging techniques alone has proven insufficient for early detection and accurate staging. Consequently, there is an urgent need for novel approaches that not only enhance diagnostic precision but also streamline the diagnostic journey for patients, reducing discomfort and minimizing unnecessary procedures.

This project represents a convergence of interdisciplinary expertise, bringing together professionals from the fields of radiology, oncology, computer science, and biomedical engineering. By fostering collaboration between these diverse disciplines, we can leverage the unique strengths of each to tackle the complexities of lung cancer diagnosis comprehensively. Moreover, this collaborative framework encourages the exchange of ideas and methodologies, fostering innovation and pushing the boundaries of what is achievable in the realm of medical diagnostics.

1.1 PROBLEM STATEMENT

Lung cancer poses a significant global health challenge, with its high mortality rates underscoring the urgent need for improved diagnostic and prognostic tools. Traditional staging methods based solely on clinical parameters often lack the granularity needed to guide personalized treatment decisions effectively. Additionally, the emergence of molecular profiling technologies has opened new avenues for understanding the underlying biological mechanisms driving lung cancer progression. Integrating clinical and molecular data through advanced machine learning techniques presents a promising opportunity to enhance lung cancer staging accuracy and refine patient stratification strategies. By leveraging the wealth of information embedded within patient demographics, tumor characteristics, and genomic signatures, this project aims to address the critical gap in current staging approaches and contribute to the advancement of precision oncology. Ultimately, the motivation behind this research lies in its potential to revolutionize lung cancer care by empowering clinicians with comprehensive tools for tailored treatment planning and improving patient outcomes.

1.2 OBJECTIVE

The primary objective of this project is to develop a robust and accurate lung cancer staging system by integrating clinical parameters and molecular insights using machine learning algorithms. Specifically, the project aims to:

1. Collect and preprocess comprehensive datasets containing clinical data such as patient demographics, tumor characteristics, and treatment history, alongside molecular data obtained from genomic profiling and biomarker analysis.
2. Explore feature engineering and selection techniques to identify relevant clinical and molecular features that contribute to lung cancer staging.
3. Implement and evaluate multiple machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM), to predict lung cancer staging based on the integrated dataset.

CHAPTER 2

This chapter serves as a comprehensive literature survey, examining the current landscape of data-driven techniques employed for lung cancer diagnosis.
We will explore various research efforts, critically analyzing the methodologies used for detecting lung cancer and determining its stage using data analysis tools. The focus will be on identifying advancements in areas like image analysis, machine learning algorithms, and their applications in lung cancer diagnosis. By surveying the existing knowledge base, we aim to establish a solid foundation for our own investigation into data-driven lung cancer prediction and staging methods.

LITERATURE SURVEY

Machine learning algorithms are showing promise in aiding lung cancer detection. Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved high accuracy in classifying lung cancer based on medical images [1]. Other approaches utilize machine learning models on various data sources, including clinical data and patient-reported symptoms. These methods offer advantages like efficiency and the ability to leverage readily available data, but accuracy might be lower compared to image-based approaches [2, 3]. Additionally, research is exploring the potential of combining clinical data with genetic information for improved risk prediction, although this raises privacy concerns [4]. Overall, machine learning offers a valuable toolkit for lung cancer detection, with the most suitable approach depending on the data type and desired balance between accuracy and accessibility.Beyond the high accuracy achieved by deep learning with Convolutional Neural Networks (CNNs) for lung cancer classification [1], ensemble learning approaches are emerging as even more powerful tools. Singh et al. (2023) demonstrated a 95% accuracy rate in lung nodule detection using an ensemble of 2D CNNs applied to chest radiographs [5]. Additionally, Zhao et al. (2023) proposed LungNet, a machine learning model that incorporates multi-view image registration and fusion, achieving high accuracy with a low false positive rate for lung cancer detection [6]. These advancements highlight the potential for even more robust and reliable lung cancer diagnosis. For a broader perspective, Yu et al. (2023) provide a comprehensive review of deep learning techniques in lung cancer screening and diagnosis using CT scans. Their analysis emphasizes the effectiveness of CNNs in surpassing human assessment in some cases for identifying and classifying lung nodules [7]. Finally, Singh et al. (2020) offer a valuable resource with their systematic review of various machine learning models employed for lung cancer prediction. Their exploration of models using clinical data, genetic data, and combinations of both provides insights into the diverse approaches within this rapidly evolving field [8].

[16]. Despite advances in surgery and chemotherapy, the survival rate for lung cancer is still low. The increasing comorbidity of obesity among lung cancer patients means that there are new problems that need to be addressed when dealing with this category of patients. A significant area of concern is whether obesity influences outcomes from surgery or chemotherapy for lung cancer, this is occasioned by the growing observation of survival advantages among overweight or obese

individuals. This enigmatic inverse connection between obesity and death attributable to lung cancer, known as the obesity paradox, remains poorly understood. In recent years a range of insights into obesity phenotype heterogeneity and associations with biological processes implicated in lung cancer's evolution have come up which may explain some of these odd seeming relationships; furthermore, well-designed clinical studies investigating causality linked to molecules involved in obesity are needed. Therefore, we explore possible biological mechanisms underlying the protective effects of adiposity on LC. Herein we emphasize the importance of resolving clinical implications for developing updated intervention strategies for the management of these two coexisting diseases: LC and Obesity.

[19]. The range of lung cancers can be reduced by physical activities based on several studies. The risk of development is 20 to 50% lower in the most active versus the least active individuals, according to these investigations. Lung cancer has been noted to be associated with both malnourishment and obesity, having a non-linear and inverted U-shaped relationship. However, since people who smoke are usually inactive and slim compared to non-smokers, it is highly probable that associations between obesity or physical activity and lung cancers are affected greatly by smoking. Nevertheless, few examinations have investigated smoking status stratified associations which involve correlations amid exercise and obesity and lung cancer (Liu et al., 2016).

Using data from the American Cancer Society Cancer Prevention Study-II Nutrition Cohort, we evaluated relationships between waist circumference (measured in 1997; sex-specific quartiles) body mass index (BMI; weight (kg)/height (m2); <18.5, 18.5–22.0 (reference), 22.1–24.9, 25.0–29.9, 30.0+ kg/m2), baseline recreational physical activity (MET-hours per week; none, 0.1 to <8.75 (reference), 8.75–17.4, 17/5+ MET h/wk) with respect to lung cancer risks stratified by smoking status among former smokers in years since quitting (<10 years but quit before study period as never smoker), current smoker ≤19 or >20/pack-years or ≥20 years). This was a retrospective cohort study conducted between the years 1992-2003 for men enrolled into ACS CPS II Nutrition cohort.

[17]. Inclusive literature review was done in PubMed for the purpose of identifying articles that were eligible. In-hospital and long-term survival outcomes were synthesized using odds ratios (OR) and hazard ratios (HR) with their corresponding 95% confidence intervals (CI). The level of heterogeneity as well as publication bias across studies was also assessed. There are also in this review, 25 cohort studies involving 78143 people. By and large, higher BMI patients showed a significantly better long term survival rate whereas increased BMI did not show significant benefits for in-hospital morbidity according to pooled analysis. Additionally, obese patients recorded a decrease on overall morbidity (OR: 0.84; 95% CI: 0.73–0.98; P = 0.025) and in-hospital mortality (OR: 0.78; 95% CI: 0.63–0.98; P = 0.031). Obesity might be viewed as an influential factor associated with lung cancer patients' favorable prognosis at their end stages of life (HR: 0.69; 95% CI: 0.56–0.86; P = 0.001). High robustness rates were observed from these pooled estimates though there was no publication bias detected in this study's reports.

To sum up, obesity can improve the condition of patients who undergo surgical procedures in hospital setting and help them survive longer if they have lung cancer.

The possibility of having an 'obesity paradox' during lung cancer surgery is supported by this statement.

[18]. Several large prospective studies have shown that there is significantly higher risk of some cancers for obese people, the International Agency for Research on Cancer has given evidence of a causal link between obesity and cancer using words such as sufficient for colon, female breast (postmenopausal), endometrium, kidney (renal cell) and esophagus (adenocarcinoma). Given these facts, plus rising trends of obesity worldwide, this amounts to massive overeating being the most important preventable cause of cancer in non-smokers. However there are not many examples of successful long-term weight loss among the fat types and thus no direct evidence exists about how reducing weight can reduce risk of cancer. Assuming total causality for the association between obesity and mortality from malignancy could mean we now conservatively estimate that 1 in 7 cancer deaths among males and 1 in 5 females occur because they were overweight or obese in America today.

CHAPTER 3

Chapter 3 focuses on the specific techniques employed in this project for lung cancer prediction and staging. We will explore six machine learning models: K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Decision Tree,Random Forest, and Support Vector Machine (SVM). These models will be evaluated with and without the application of Principal Component Analysis (PCA) for dimensionality reduction. It also goes into the dataset description, architecture diagram, exploratory data analysis, and the experimental set up.

PROPOSED METHODOLOGY

3.1 MACHINE LEARNING MODELS

3.1.1 Logistic Regression:

[9] Logistic Regression is a supervised binary classification algorithm where it predicts the desired outcome based on the features used and using the Sigmoid Function, the model would transform the selected features into a probability, which will be easy for it to predict the outcome and the relationship between the dependent and the independent features. In our problem, Logistic Regression acts as a base model for predicting the stages of lung cancer, even though being used in binary classification techniques. Its functionality can give a deep insight into the importance of the features used and its aid in predicting the right outcome and other key factors.

3.1.2 K-Nearest Neighbors:

[10] K-Nearest Neighbours is a supervised classification algorithm and also a non-parametric algorithm, where it does make any assumptions about the features used. It generates the outcome based on the number of neighbors we choose. Cross Validation Techniques are used to find the optimal number of neighbors required to classify different classes. Distance metrics such as Euclidean, Manhattan are used along with the number of neighbors to distinguish a class. In our problem, K-Nearest Neighbours is used to explore the relationships among the data points and clustering those points with the help of 13 neighbors chosen to predict the level of Lung Cancer.

3.1.3 Naive Bayes Classifier:

[11] Naive Bayes is a supervised classification algorithm which follows the principles of Bayes Theorem used in solving probability problems. The reason it got the name 'Naive' is due to the fact that it assumes all the features to be independent of each other and having no relationship between them. In our problem, Naive Bayes is used to generate the level of Lung Cancer based on the computational probability distributions given the data which includes, which state does this cancer belong to.

3.1.4 Decision Tree:

[12] Decision Tree is a supervised classification algorithm. Decision trees are a popular method for predictive modeling because they are easy to interpret and visualize. The algorithm works by partitioning the data set into subsets based on the input features. The partitions are made in such a way that the samples within each subset are as similar as possible with respect to the target variable. This process is repeated recursively, resulting in a tree-like structure. In our problem, the Decision tree algorithm is used in predicting the level of lung cancer by analyzing various input features such as Coughing of blood, smoking history, Dust Allergy and other relevant factors. The algorithm works by creating a tree-like structure that partitions the data based on these features and identifies key patterns and relationships. By recursively making these partitions and analyzing the similarity of samples within each subset with respect to the target variable (in this case, the level of lung cancer), the decision tree algorithm can accurately predict the likelihood and severity of lung cancer for individuals based on their specific characteristics.

3.1.5 Random Forest:

[13] Random Forest is a supervised and ensemble classification algorithm. Its working is based on combining a power of multiple decision trees and with the help of this, the model learns from the mistakes the older model did, so that it tries to mitigate the process the older model had to go through. It classifies the data points based on the voting system of all models together. The data point gets denoted to a certain class based on the highest number of votes that class got based on that data. In our problem, our model gets benefited from Random Forest by the outcome of multiple decision trees. This process can mitigate the chance of overfitting and the complexity of the model. It also gives a clear idea of which stage does the cancer belong to.

3.1.6 Support Vector Machines:

[14] Support Vector Machines is a supervised classification algorithm. Its sole purpose is to build a hyperplane in a dimensional space consisting of various data points, in such a way that it separates the distinct classes. The hyperplane is specifically located in a position in such a way that it maximizes the margin between any distinct classes and to mitigate clashes between those classes for the datapoint the dimensionality space In our problem, Support Vector Machines helps in predicting the level of lung cancer by the help of the hyperplane. The functionality of SVM's makes the process of class distinguishing a lot easier. Its ability to predict the correct outcome for unseen data, using kernel functions to capture relationships between the independent and dependent features.

3.1.7 Principal Component Analysis:

[15] Principal Component Analysis is a dimensionality reduction technique, which is used to reduce the dimensions of a given model to a certain lower number without affecting its performance or originality. Based on the number of features we want to reduce, that amount of EigenVectors are produced in a dimensional space and those vectors propagate in a direction where high variance is produced. More the variance, the higher the information is captured

with reduction. These vectors represent the features and the vector with highest variance, the Principal Component is selected and used to train the model. In our model, we tried to implement the above procedure to reduce the complexity and overfitting of the model. Even though five features are used to train the model, we wanted to explore how the model works with even reduction from those 5 features to 3.

3.2 DATASET DESCRIPTION

Table 3.2.1

VARIABLE NAME	**DESCRIPTION**
Age	The age of the patient. (Numeric)
Gender	The gender of the patient. (Categorical)
Air Pollution	The level of air pollution exposure of the patient. (Categorical)
Alcohol use	The level of alcohol use of the patient. (Categorical)
Dust Allergy	The level of dust allergy of the patient. (Categorical)
OccuPational Hazards	The level of occupational hazards of the patient. (Categorical)
Genetic Risk	The level of genetic risk of the patient. (Categorical)
chronic Lung Disease	The level of chronic lung disease of the patient. (Categorical)
Balanced Diet	The level of a balanced diet of the patient. (Categorical)
Obesity	The level of obesity of the patient. (Categorical)
Smoking	The level of smoking of the patient. (Categorical)
Passive Smoker	The level of passive smoking of the patient. (Categorical)
Chest Pain	The level of chest pain of the patient. (Categorical)

Coughing of Blood	The level of coughing of blood of the patient. (Categorical)
Fatigue	The level of fatigue of the patient. (Categorical)
Weight Loss	The level of weight loss of the patient. (Categorical)
Shortness of Breath	The level of shortness of breath of the patient. (Categorical)
Wheezing	The level of wheezing of the patient. (Categorical)
Swallowing Difficulty	The level of swallowing difficulty of the patient. (Categorical)
Clubbing of Finger Nails	The level of clubbing of finger nails of the patient. (Categorical)

3.3 ARCHITECTURE DIAGRAM (FLOW CHART)

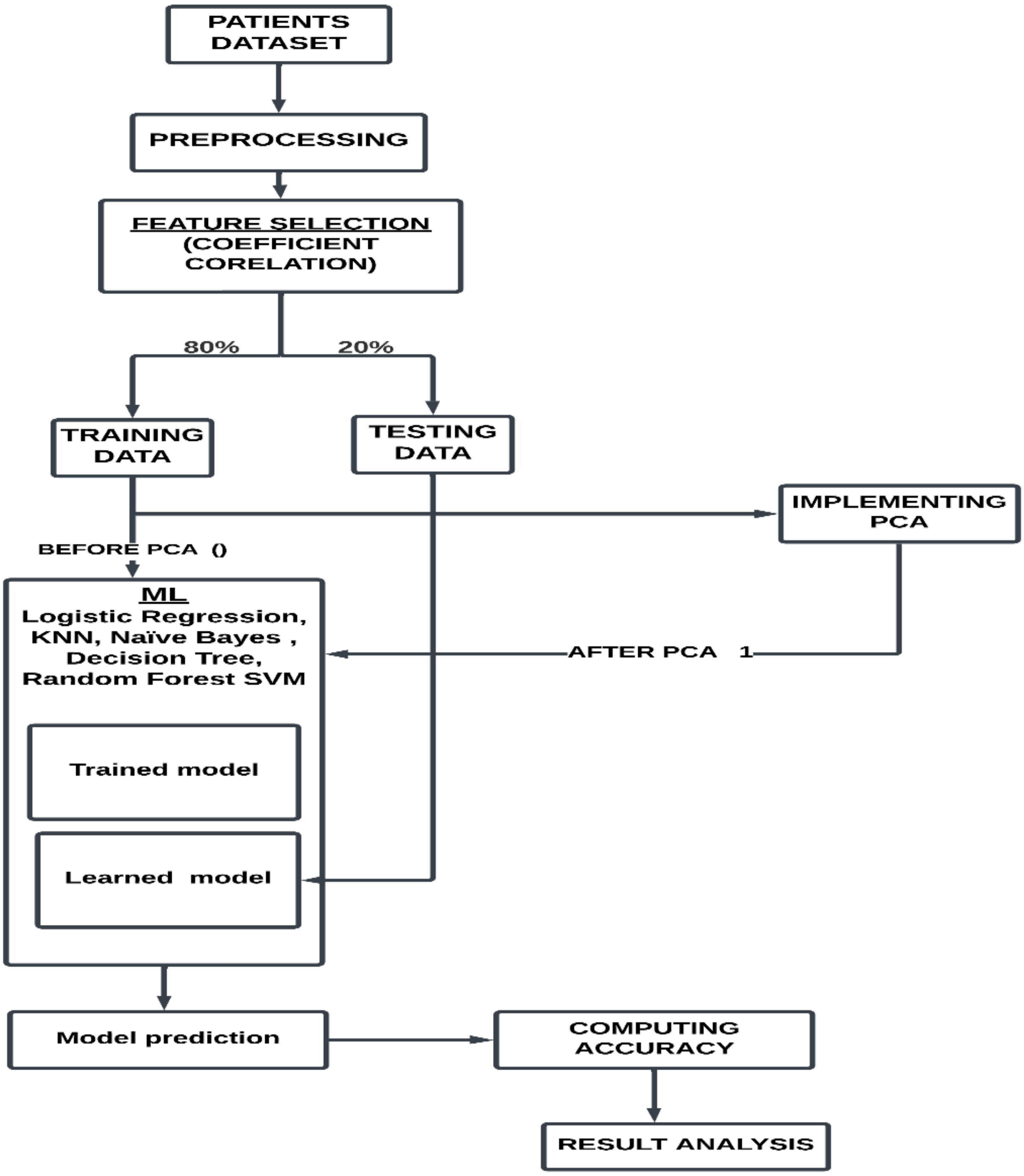


Fig. 3.3.1

- **Data Collection:** The process starts with acquiring a dataset of patients labeled as having lung cancer or not (healthy).
- **Data Preprocessing:** The data is cleaned and formatted for the machine learning model. This might involve removing errors, inconsistencies, scaling, or normalization.
- **Data Splitting:** The data is divided into two sets: training data used to train the model and testing data used to evaluate its performance on unseen data.
- **Model Training:** A machine learning algorithm is trained using the training data. This allows the algorithm to learn patterns for classifying patients.
- **Model Evaluation:** The model's performance is assessed on the testing data. Predictions are

compared to actual labels to determine its accuracy on unseen data.

• **Decision Point:** Based on the evaluation results: * If the model performs well, it can be deployed for real-world use (implementation). * If the model performs poorly, revisit previous steps and adjust (e.g., modify data preprocessing, try a different algorithm)

3.4 EXPERIMENTAL SET UP

- Utilized Jupyter Notebook for initial data exploration, preprocessing, model development, and result visualization.
- Employed Streamlit for building an interactive web application to facilitate easy input of patient data and display prediction results in real-time.
- Leveraged scikit-learn, a machine learning library in Python, for implementing various classification algorithms including K-Nearest Neighbors (KNN) for lung cancer prediction.
- Utilized pandas for data manipulation, cleaning, and preprocessing tasks, facilitating efficient handling of structured data.
- Utilized NumPy for numerical computations and array operations, enabling efficient handling of multidimensional arrays and matrices.
- Utilized Matplotlib for data visualization, generating various plots such as histograms, scatter plots, and confusion matrices to analyze model performance and interpret results.
- Employed Pickle for serializing trained machine learning models and StandardScaler objects, allowing seamless storage and retrieval of model state and preprocessing parameters.

3.5 EXPLORATORY DATA ANALYSIS

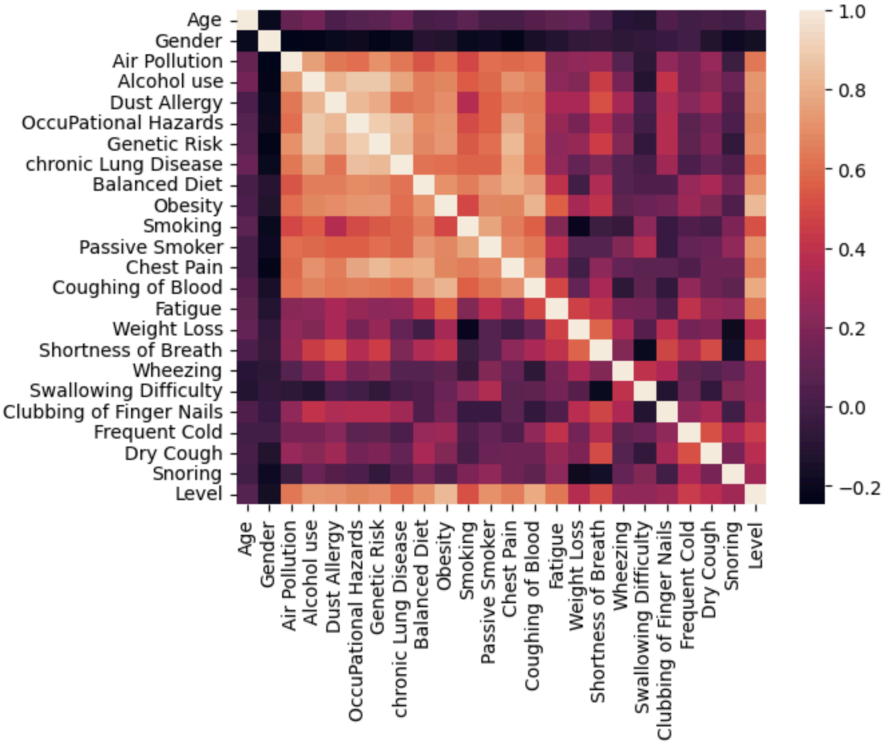


Fig. 3.4.1

The above heatmap describes the correlation of each attribute in the dataset with each other. The constant of proportion is indicated by the shades of color in the legend. From the above heatmap, it can be observed that

- Obesity and the level of cancer are strongly correlated
- The gender of the person is not related to the level of cancer
- A patient with dust allergy is more susceptible to higher levels of cancer

With such insights, feature selection can take place and the supervised learning models can be trained.

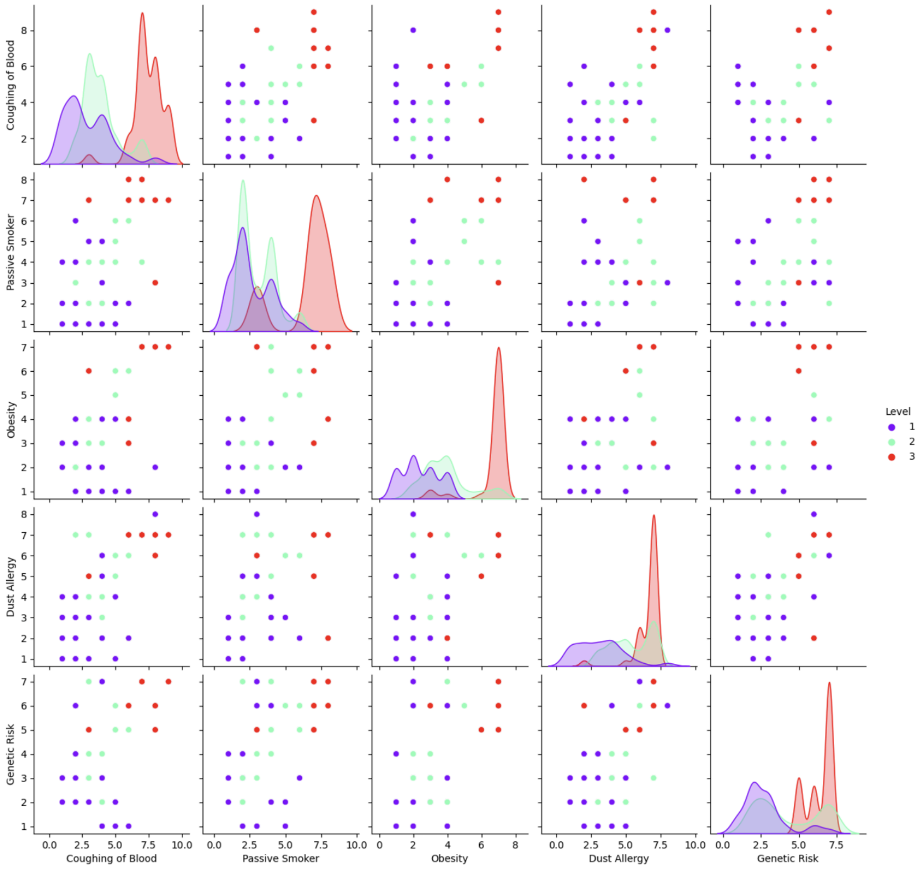


Fig 3.4.2

The above pair plot describes how each of the selected attributes vary with each other. Each datapoint is indicated by a hue to denote which level of cancer occurs at that datapoint. For example, in the pair plot between genetic risk and passive smoker, we can see that as both attributes get higher, the level and chance of characteristic increases and vice versa.

CHAPTER 4

Chapter 4 delves into the results of all the models, and a comparison of each of their metrics. The models are also ranked in order of accuracy. Each model's metrics are included with and without principal component analysis.

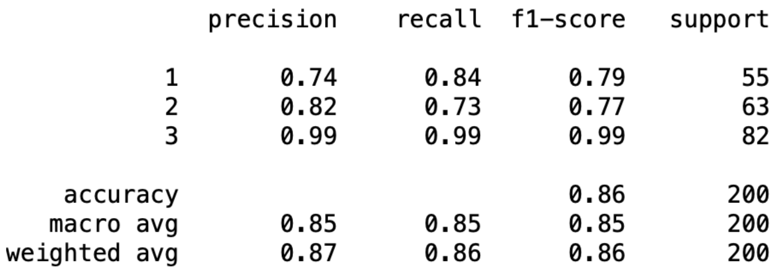
RESULTS AND DISCUSSION

The dataset was used to train 6 different models with and without PCA (Principal Component Analysis) - Logistic Regression, KNN, Naive Bayes, Decision Tree, Random Forest, SVM (Support Vector Machine). The results are as follows:

4.1 SUPERVISED LEARNING MODEL OUTCOMES

LOGISTIC REGRESSION:

Without PCA:

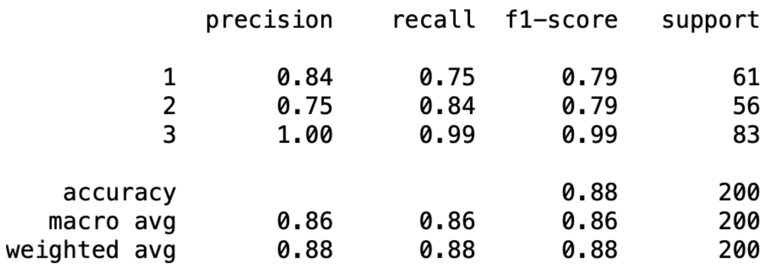


```
              precision    recall  f1-score   support

           1       0.74      0.84      0.79        55
           2       0.82      0.73      0.77        63
           3       0.99      0.99      0.99        82

    accuracy                           0.86       200
   macro avg       0.85      0.85      0.85       200
weighted avg       0.87      0.86      0.86       200
```

With PCA:



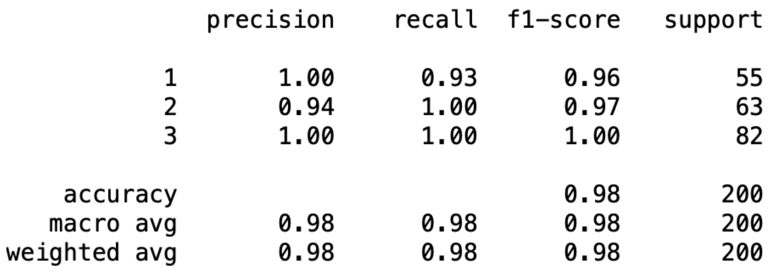
```
              precision    recall  f1-score   support

           1       0.84      0.75      0.79        61
           2       0.75      0.84      0.79        56
           3       1.00      0.99      0.99        83

    accuracy                           0.88       200
   macro avg       0.86      0.86      0.86       200
weighted avg       0.88      0.88      0.88       200
```

K-NEAREST NEIGHBORS:

Without PCA:

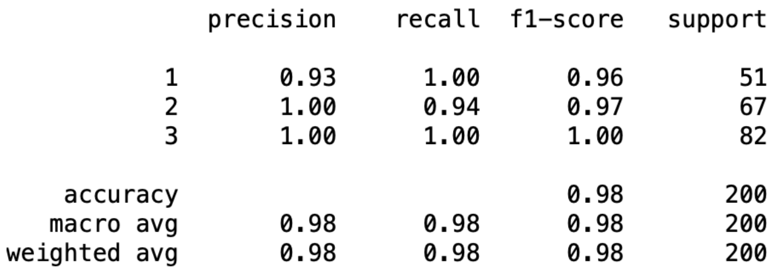


```
              precision    recall  f1-score   support

           1       1.00      0.93      0.96        55
           2       0.94      1.00      0.97        63
           3       1.00      1.00      1.00        82

    accuracy                           0.98       200
   macro avg       0.98      0.98      0.98       200
weighted avg       0.98      0.98      0.98       200
```

With PCA:



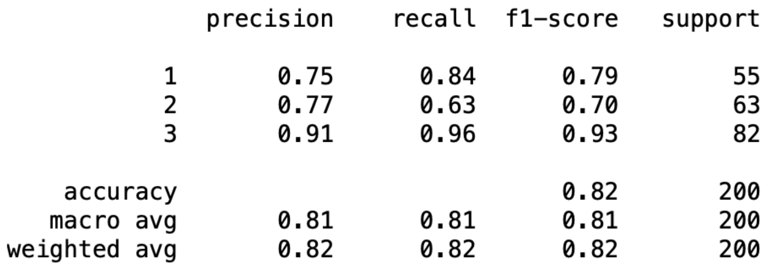
```
              precision    recall  f1-score   support

           1       0.93      1.00      0.96        51
           2       1.00      0.94      0.97        67
           3       1.00      1.00      1.00        82

    accuracy                           0.98       200
   macro avg       0.98      0.98      0.98       200
weighted avg       0.98      0.98      0.98       200
```

NAIVE BAYES:

Without PCA:

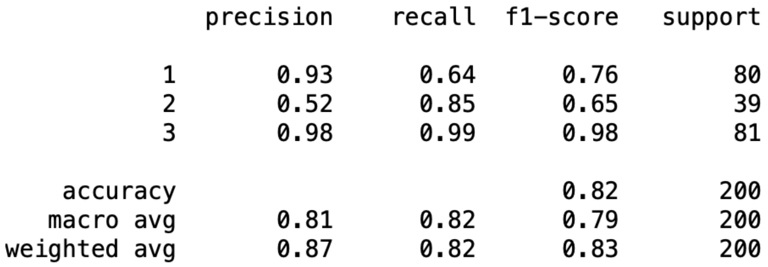


```
              precision    recall  f1-score   support

           1       0.75      0.84      0.79        55
           2       0.77      0.63      0.70        63
           3       0.91      0.96      0.93        82

    accuracy                           0.82       200
   macro avg       0.81      0.81      0.81       200
weighted avg       0.82      0.82      0.82       200
```

With PCA:



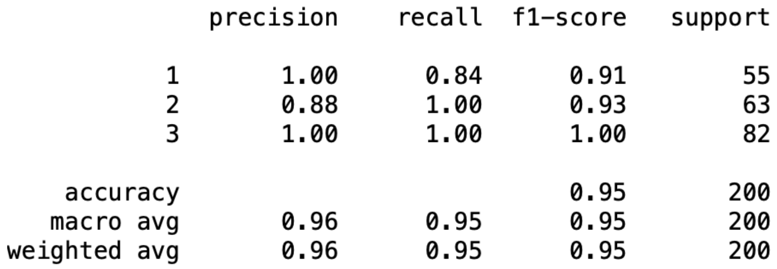
```
              precision    recall  f1-score   support

           1       0.93      0.64      0.76        80
           2       0.52      0.85      0.65        39
           3       0.98      0.99      0.98        81

    accuracy                           0.82       200
   macro avg       0.81      0.82      0.79       200
weighted avg       0.87      0.82      0.83       200
```

DECISION TREE:

Without PCA:

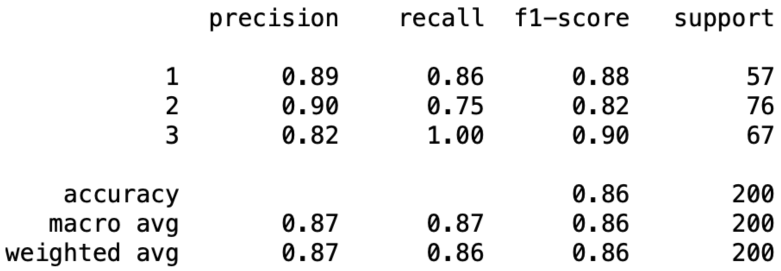


```
              precision    recall  f1-score   support

           1       1.00      0.84      0.91        55
           2       0.88      1.00      0.93        63
           3       1.00      1.00      1.00        82

    accuracy                           0.95       200
   macro avg       0.96      0.95      0.95       200
weighted avg       0.96      0.95      0.95       200
```

With PCA:



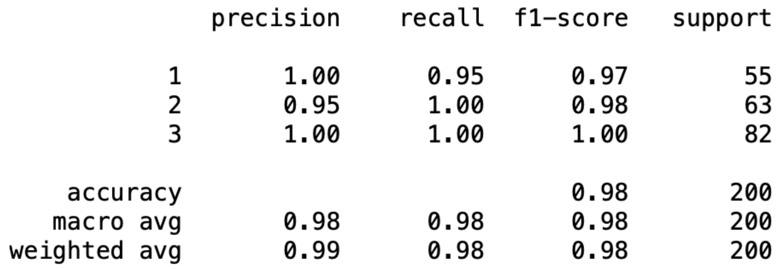
```
              precision    recall  f1-score   support

           1       0.89      0.86      0.88        57
           2       0.90      0.75      0.82        76
           3       0.82      1.00      0.90        67

    accuracy                           0.86       200
   macro avg       0.87      0.87      0.86       200
weighted avg       0.87      0.86      0.86       200
```

RANDOM FOREST:

Without PCA:

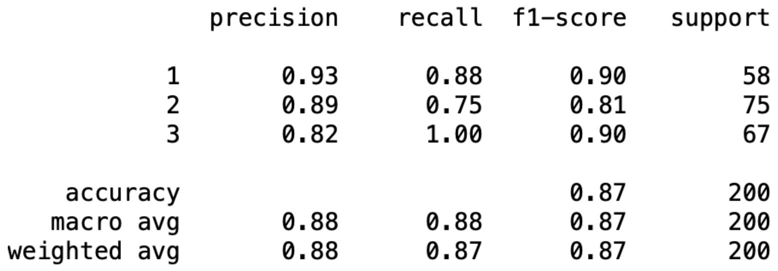


```
              precision    recall  f1-score   support

           1       1.00      0.95      0.97        55
           2       0.95      1.00      0.98        63
           3       1.00      1.00      1.00        82

    accuracy                           0.98       200
   macro avg       0.98      0.98      0.98       200
weighted avg       0.99      0.98      0.98       200
```

With PCA:



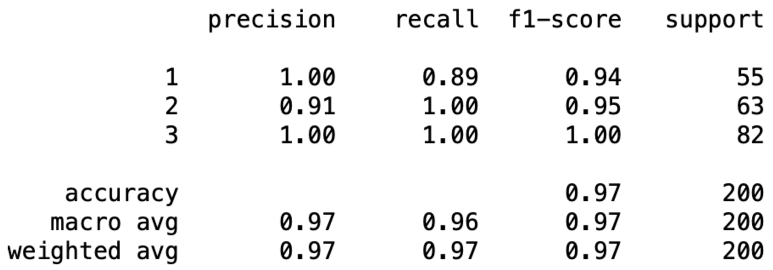
```
              precision    recall  f1-score   support

           1       0.93      0.88      0.90        58
           2       0.89      0.75      0.81        75
           3       0.82      1.00      0.90        67

    accuracy                           0.87       200
   macro avg       0.88      0.88      0.87       200
weighted avg       0.88      0.87      0.87       200
```

SUPPORT VECTOR MACHINE (SVM):

Without PCA:

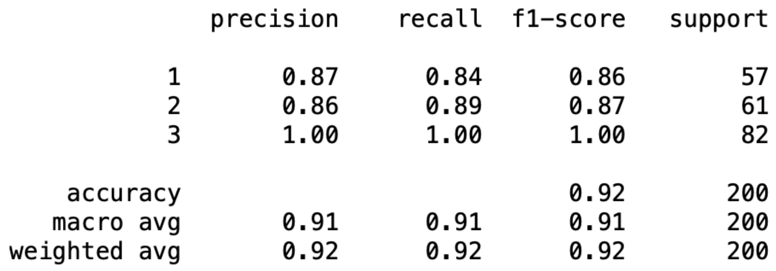


```
              precision    recall  f1-score   support

           1       1.00      0.89      0.94        55
           2       0.91      1.00      0.95        63
           3       1.00      1.00      1.00        82

    accuracy                           0.97       200
   macro avg       0.97      0.96      0.97       200
weighted avg       0.97      0.97      0.97       200
```

With PCA:



```
              precision    recall  f1-score   support

           1       0.87      0.84      0.86        57
           2       0.86      0.89      0.87        61
           3       1.00      1.00      1.00        82

    accuracy                           0.92       200
   macro avg       0.91      0.91      0.91       200
weighted avg       0.92      0.92      0.92       200
```

4.2 COMPARISON OF MODELS:

Table 4.2.1 - Before PCA

MODEL	TRAINING ACCURACY	TESTING ACCURACY
Logistic Regression	0.79	0.865
KNN	0.99	0.98
Naïve Bayes Classifier	0.765	0.825
Decision Tree	0.91	0.955
Random Forest	0.952	0.985
SVM	0.93	0.97

Table 4.2.2 - After PCA

MODEL	TRAINING ACCURACY	TESTING ACCURACY
Logistic Regression	0.80	0.875
KNN	0.992	0.98
Naïve Bayes Classifier	0.80	0.82
Decision Tree	0.98	0.865
Random Forest	0.99	0.87
SVM	0.915	0.92

Table 4.2.3 - Model Metrics

MODEL/ACCURACY	Recall score	F1 score	Precision score	Accuracy
Logistic Regression	0.85	0.84	0.85	0.865
KNN	0.975	0.977	0.98	0.98
Naïve Bayes Classifier	0.81	0.80	0.81	0.825
Decision Tree	0.94	0.94	0.95	0.955
Random Forest	0.98	0.98	0.98	0.985
SVM	0.96	0.96	0.97	0.97

As seen from the data the accuracy in increasing order is as follows:

1. Naive Bayes without PCA (82% Accuracy)
2. Naive Bayes with PCA (82% Accuracy)
3. Logistic Regression without PCA (86% Accuracy)
4. Decision Tree with PCA (86% Accuracy)
5. Random Forest with PCA (87% Accuracy)
6. Logistic Regression with PCA (88% Accuracy)
7. SVM with PCA (92% Accuracy)
8. Decision Tree without PCA (95% Accuracy)
9. SVM without PCA (97% Accuracy)
10. KNN without PCA (98% Accuracy)
11. Random Forest without PCA (98% Accuracy)
12. KNN with PCA (98% Accuracy)

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENT

Future Enhancements:

- Data Acquisition and Sharing:
    - Develop standardized data collection protocols for lung cancer diagnosis.
    - Foster secure data sharing platforms to facilitate the development of robust and generalizable machine learning models.
- Explainable AI:
    - Integrate explainable AI techniques into deep learning models to improve transparency and trust in their decision-making processes for lung cancer detection.
- Integration with Clinical Workflow:
    - Develop seamless integration of machine learning models into clinical workflows to enhance physician decision-making and personalize patient care.

Conclusion:

Machine learning offers a powerful arsenal for lung cancer detection. With continued advancements in deep learning, ensemble learning, and explainability, these models have the potential to become even more accurate and reliable. By fostering data sharing, integrating with clinical workflows, and prioritizing explainability, machine learning can become a cornerstone in the fight against lung cancer.

In this study, we explored the performance of various machine learning models for the task of lung cancer prediction. Leveraging a dataset rich in patient features and diagnostic indicators, we trained and evaluated six different models: Logistic Regression, K-Nearest Neighbors (KNN), Naïve Bayes Classifier, Decision Tree, Random Forest, and Support Vector Machine (SVM).

The results of our evaluation revealed notable variations in the performance of these models. KNN and Random Forest emerged as the top-performing models, consistently achieving high accuracy, precision, recall, and F1 scores, all critical metrics for assessing predictive performance in medical contexts. Decision Trees and SVM also demonstrated strong performance, while Logistic Regression and Naïve Bayes Classifier exhibited more moderate predictive capabilities.

These findings underscore the importance of selecting an appropriate machine learning algorithm tailored to the complexities of the dataset and the specific requirements of the predictive task. Moreover, the success of models like KNN and Random Forest highlights the potential of machine learning in aiding early detection and diagnosis of lung cancer, offering clinicians valuable decision support tools to improve patient outcomes.

Moving forward, further refinement and validation of these models using larger and more diverse datasets, as well as integration with clinical workflows, will be essential for translating these findings into real-world clinical practice. Additionally, exploring ensemble methods and deep learning architectures could offer avenues for enhancing predictive performance and robustness in future iterations of this research.

In conclusion, this study contributes valuable insights into the application of machine learning for lung cancer prediction, laying the groundwork for continued advancements in early detection and personalized treatment strategies for this devastating disease.

CHAPTER 6

REFERENCES

1. Gupta, S., et al. (2019). Performance Evaluation of Deep Learning Techniques for Lung Cancer Prediction.
2. Nageswaran, A. K., et al. (2022). Lung Cancer Classification and Prediction Using Machine Learning and Image Processing.
3. Nemlander, A., et al. (2022). Lung Cancer Prediction Using Machine Learning on Data from a Symptom E-Questionnaire.
4. Xu, Y., et al. (2020). Lung Cancer Risk Prediction with Machine Learning Models.
5. Singh, S. et al. (2023). Deep learning ensemble 2D CNN approach towards early and automatic identification of Lung nodules in Chest Radiographs. Nature Research.
6. Zhao, J. et al. (2023). Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion. Hindawi Publishing Corporation.
7. Yu, L. et al. (2023). A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. MDPI.
8. Singh, D. et al. (2020). Lung Cancer Prediction Using Machine Learning Models: A Systematic Review.
9. Joanne Peng , Kuk Lida Lee,Gary M.Ingersoll. An Introduction to Logistic Regression Analysis and Reporting. The journal of Educational Research 96(1) : 3-14,2002.
10. Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer. KNN Model-Based Approach in Classification. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Lecture Notes in Computer Science 2888:986-996. (2003)
11. Irina Rish. An Empirical Study of the Naive Bayes Classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence, 3, page 41--46. IBM New York, (2001).
12. Harsh Patel,Purvi Prajapati.Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering 6(10):74-78.2018
13. Gerard Biau.Analysis of a Random Forests Model. Journal of Machine Learning Research 13 (2012) 1063-1095. 2012
14. Thedoros Evgeniou,Massimiliano Pontil.Support Vector Machines:Theory and Applications. pp 249–25. (2001).
15. Ian T.Jolliffe and Jorge Cadmia.Principal Component Analysis: A review and recent developments. Philosophical Transactions A 374(2065):20150202.2016.
16. Xueli Zhang PhD , Yamin Liu PhD , Hua Shao MD, PhD , Xiao Zheng PhD. Obesity Paradox in Lung Cancer Prognosis: Evolving Biological Insights and Clinical Implications. Journal of Thoracic Oncology. 2017 Oct;12(10):1478-1488. 2017
17. Shuangjiang Li, Zhiqiang Wang, Jian Huang, Jun Fan, Heng Du, Lunxu Liu, Guowei Che. Systematic review of prognostic roles of body mass index for patients undergoing lung cancer surgery: does the ‘obesity paradox’ really exist. European Journal of

CARDIO-THORACIC SURGERY. Volume 51, Issue 5, May 2017, Pages 817–828.2017.

18. Eugenia E Calle and Michael J Thun. Obesity and Cancer. Oncogene. 2004 Aug 23;23(38):6365-78.2004.
19. Patel, A.V., Carter, B.D., Stevens, V.L. et al. The relationship between physical activity, obesity, and lung cancer risk by smoking status in a large prospective cohort of US adults. Cancer Causes Control 28, 1357–1368 (2017).