Lungs Disease Detection Using Machine Learning

Mohammad Sameer* Shambhavi Kumari* Shahzeb Khan*

*Department of Computer Science & Applications, Sharda School of Computing Science & Engineering, Sharda University, Greater Noida, India

Abstract

This is paper presents a machine learning approach for the detection of lung diseases. Various algorithms are explored to improve detection accuracy, utilizing datasets and image processing techniques.Lungs disease refers to many disorders affecting the lungs such as Pneumonia, Tuberculosis, Lungs cancer, and many other breathing problems. COVID-19 as a prominent lung disease is prioritized over others. Imaging modalities including ,X-rays, Ct scan,MRIs are primarily employed in medical assessment because they provide computed data that can be utilize as input datasets for computer assisted diagnostic systems Datasets Used are chest x-ray datasets, Ct-scan datasets, Clinical datasets , NSCLC Radiomics datasets. COPD(chronic obstructive pulmonary disease) is a long term inflammatory lung condition that causes airflow obstruction in the lungs. Nowadays Lung diseases are becoming a Significant problem. Machine learning (ML) with feature selection techniques play a significant role in the medical field by making diseases diagnoses accurate and early.Many classification algorithms are used to predict lungs problem Logistic Regression, Random Forest, and Bayesian Networks. This paper explores the application of machine learning in lung disease detection,providing an overview of technique, datasets, challenges, and future direction.

Keywords: Lung Disease, Machine Learning, Medical Imaging, AI in Healthcare

1. Introduction

Detection of lung disease is an important area in medical diagnosis. Machine learning has demonstrated promising consequences in analyzing medical images for early detection of diseases such as pneumonia and lung cancer. Lung cancer is still one of the most common and deadly diseases in the world, and its high mortality is the result of late diagnosis. Preliminary identification and accurate risk evaluation is required to improve patient results. Traditional clinical approaches, including imaging and clinical assessment, are usually less in detecting high risk people in an early stage. However, new methods have been made possible for screening of lung cancer, survival analysis and prediction by development in Artificial Intelligence (AI) and machine learning (ML). Recent studies have shown that a variety of machine learning (mL) approaches, such as theory component analysis, supervised learning, regression models and machine learning, can improve the diagnosis of lung cancer and diagnosis. These studies check a variety of data sources to improve future stating accuracy, including calculated tomography (CT) scan, standard blood test results, electronic health records (EHRS), and patient lifestyle factors.*(9)*

Medical diagnostics have thanked a revolution for machine learning (ML), which provides state -of -the -art technology for early diagnosis and risk evaluation of serious diseases such as lung cancer. Due to most of the late phase diagnosis, lung cancer has a significant mortality rate and cancer -related deaths continue to rank in the major causes of the world. Traditional screening techniques, such as symptoms-based evaluation and low-khurak calculation tomography (LDCT), often decreases in identifying high-risk individuals. Propecting accuracy, automating risk evaluation, and improving the patient's results is made possible by the inclusion of machine learning models in the healthcare system. Improvement in clinical results brought about progress in diagnosis and treatment, the 5-year survival rate is only (22%), nd most lungs are also found after cancer.[1] Demographic, lifestyle and therapy symptoms characteristics that are relevant to the diagnosis of lung cancer are included in the dataset used in this investigation. Important characteristics used as future factors for machine learning models include age, alcohol intake, smoking history, wheezing, cough and chest pain. This study attempts to assess model performance and determine the best strategy for predicting lung cancer using a variety of machine learning approaches. To guarantee strong prediction performance, feature selection and model optimization techniques are used, including cross-validation and hyperpieme tuning. This research paper examines the effectiveness of various ML algorithms in predicting the risk of lung cancer using structured clinical data.

The patient may be given a proper treatment, allowing the execution of proper preventive measures, if the cancer is discovered within a specific time period for treatment and many risk factors exist for further diagnosis. Many computer techniques have been developed to detect or forecast lung cancer, which helps the medical professionals to determine the forecast of patients after the most effective course of treatment and diagnosis. Medical scientists have used machine learning and soft computing techniques to correctly identify many types of cancer types in their early stages using classification techniques.*(17)*

The study intends to develop AI-operated initial identifica-

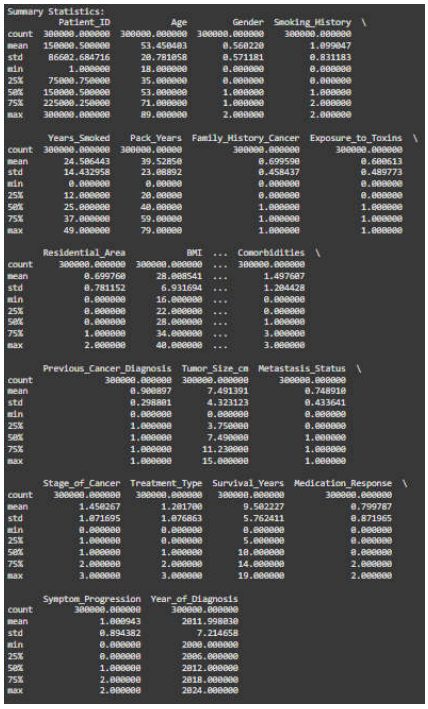
tion systems by assessing several models using performance criteria such as accuracy, accuracy and recall. Early detection improvement, reducing mortality, and streamlining medical decision making processes are all possible consequences of effective applications of ML models in the diagnosis of lung cancer.

In the end, this research provides feedback for the following questions: 1. In the research sector, which is the risk factor for most cases of lung cancer?

2. How can a machine learning model be used to determine the severity of lung cancer?

The full body of the work of this study has been structured as follows: resources and techniques used to meet the goal of the study are included in metodology. The findings and future instructions of this research are eventually included in the future scope.[2]

2. Data Description



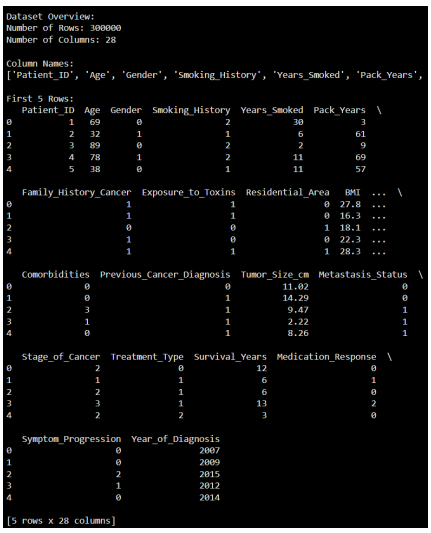
```
Summary Statistics:
          Patient_ID            Age         Gender  Smoking_History  \
count  300000.000000  300000.000000  300000.000000    300000.000000
mean   150000.500000      53.450403       0.560220         1.099047
std     86602.684716      20.781058       0.571181         0.831183
min         1.000000      18.000000       0.000000         0.000000
25%     75000.750000      35.000000       0.000000         0.000000
50%    150000.500000      53.000000       1.000000         1.000000
75%    225000.250000      71.000000       1.000000         2.000000
max    300000.000000      89.000000       2.000000         2.000000

       Years_Smoked    Pack_Years  Family_History_Cancer  Exposure_to_Toxins  \
count  300000.000000  300000.00000          300000.000000       300000.000000
mean       24.506443      39.52850               0.699590            0.600613
std        14.432958      23.08892               0.458437            0.489773
min         0.000000       0.00000               0.000000            0.000000
25%        12.000000      20.00000               0.000000            0.000000
50%        25.000000      40.00000               1.000000            1.000000
75%        37.000000      59.00000               1.000000            1.000000
max        49.000000      79.00000               1.000000            1.000000

       Residential_Area            BMI  ...  Comorbidities  \
count     300000.000000  300000.000000  ...  300000.000000
mean           0.699760      28.808541  ...       1.497607
std            0.781152       6.931694  ...       1.284428
min            0.000000      16.000000  ...       0.000000
25%            0.000000      22.000000  ...       0.000000
50%            0.000000      28.000000  ...       1.000000
75%            1.000000      34.000000  ...       3.000000
max            2.000000      40.000000  ...       3.000000

       Previous_Cancer_Diagnosis  Tumor_Size_cm  Metastasis_Status  \
count              300000.000000  300000.000000      300000.000000
mean                   0.900897       7.491391           0.748910
std                    0.298801       4.321123           0.433641
min                    0.000000       0.000000           0.000000
25%                    1.000000       3.750000           0.000000
50%                    1.000000       7.490000           1.000000
75%                    1.000000      11.230000           1.000000
max                    1.000000      15.000000           1.000000

       Stage_of_Cancer  Treatment_Type  Survival_Years  Medication_Response  \
count    300000.000000   300000.000000   300000.000000        300000.000000
mean          1.450267        1.201700        9.502227             0.799787
std           1.071695        1.076863        5.762411             0.871965
min           0.000000        0.000000        0.000000             0.000000
25%           1.000000        0.000000        5.000000             0.000000
50%           1.000000        1.000000       10.000000             0.000000
75%           2.000000        2.000000       14.000000             2.000000
max           3.000000        3.000000       19.000000             2.000000

       Symptom_Progression  Year_of_Diagnosis
count        300000.000000      300000.000000
mean              1.000943        2011.998830
std               0.894382           7.214658
min               0.000000        2000.000000
25%               0.000000        2006.000000
50%               1.000000        2012.000000
75%               2.000000        2018.000000
max               2.000000        2024.000000
```

Figure 1: Summary of the dataset

This dataset is a source from Kaggle used in this study. It consists of health parameters of more than 3 lakh patient diagnosed with lung cancer. A total of more than 10 features were taken into consideration for this study like,Age, gender, smoking habits, BMI, chest function, chest pain, tumor size and stage of cancer,etc were among the characteristics. Depending on a labeled data model, metastasis and other health status indicators are calculated directly from dataset to provide an detective output that shows whether a disease exists. *(11)*

This article summarizes the conclusions with ten studies examining the methods of machine learning (ML) for diagnosis and prediction of lung cancer. The functioning used in this synthesis collects, analyzes and valid data from several studies in an organized manner.

A. Recap of dataset



```
Dataset Overview:
Number of Rows: 300000
Number of Columns: 28

Column Names:
['Patient_ID', 'Age', 'Gender', 'Smoking_History', 'Years_Smoked', 'Pack_Years',

First 5 Rows:
   Patient_ID  Age  Gender  Smoking_History  Years_Smoked  Pack_Years  \
0           1   69       0                2            30           3
1           2   32       1                1             6          61
2           3   89       0                2             2           9
3           4   78       1                2            11          69
4           5   38       0                1            11          57

   Family_History_Cancer  Exposure_to_Toxins  Residential_Area   BMI  ...  \
0                      1                   1                 0  27.8  ...
1                      1                   1                 0  16.3  ...
2                      0                   0                 1  18.1  ...
3                      1                   0                 0  22.3  ...
4                      1                   1                 1  28.3  ...

   Comorbidities  Previous_Cancer_Diagnosis  Tumor_Size_cm  Metastasis_Status  \
0              0                          0          11.02                  0
1              0                          1          14.29                  0
2              3                          1           9.47                  1
3              1                          1           2.22                  1
4              0                          1           8.26                  1

   Stage_of_Cancer  Treatment_Type  Survival_Years  Medication_Response  \
0                2               0              12                    0
1                1               1               6                    1
2                2               1               6                    0
3                3               1              13                    2
4                2               2               3                    0

   Symptom_Progression  Year_of_Diagnosis
0                    0               2007
1                    0               2009
2                    2               2015
3                    1               2012
4                    0               2014

[5 rows x 28 columns]
```

Figure 2: Missing Values in Dataset

This dataset includes 3,00,000 rows and 15 columns. All columns are complete, as illustrated in the picture (see Figure 2). Each column has the same data type, which is integer.

observation 1: graph indicates that the high proportion of patients is diagnosed with advanced phase compared to another category. This emphasizes the need for progress in early detection technology. In it (see Figure 3) represents the age of X-Xis patent and represents the phase of Y-Axis cancer, which is divided into four stages. Cancer is likely to be between 35 and 70 ages. *(10)*

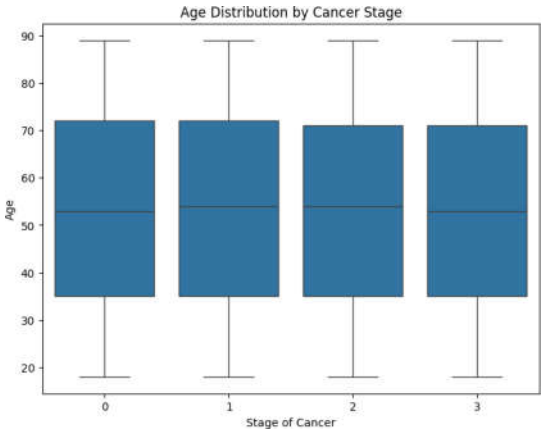


Figure 3: Age Distribution by Cancer Stage

Observation 2: In the figure (see Figure 4)top ten factor that affects the prediction of lung cancer is displayed in the graph. Tumor size, lung function, air quality, smoking history, body mass index and age are important factor. Since the patient ID is not a reliable prophet, it should be disregarded. Conclusions support the importance of smoking, environment variable and cancer phase in smoking. *(19)*

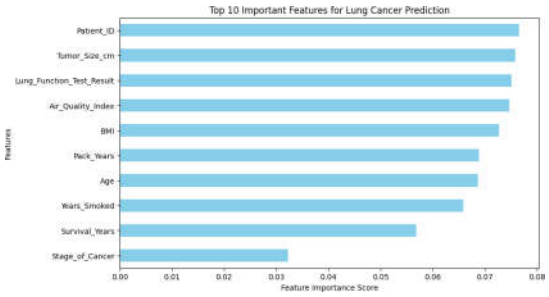


Figure 4: Top 10 Important Features For Lungs Cancer Prediction

B. Limitations of the dataset

This dataset is really spacious and contains a lot of features and information that are not necessary for prediction. A large number of irrelevant features reduce the performance of the model. To make the dataset more accurate, we have abolished a lot of useless features.

```
Missing Values:
Patient_ID                    0
Age                           0
Gender                        0
Smoking_History               0
Years_Smoked                  0
Pack_Years                    0
Family_History_Cancer         0
Exposure_to_Toxins            0
Residential_Area              0
BMI                           0
Lung_Function_Test_Result     0
Chest_Pain_Symptoms           0
Shortness_of_Breath           0
Chronic_Cough                 0
Weight_Loss                   0
Physical_Activity_Level       0
Dietary_Habits                0
Air_Quality_Index             0
Comorbidities                 0
Previous_Cancer_Diagnosis     0
Tumor_Size_cm                 0
Metastasis_Status             0
Stage_of_Cancer               0
Treatment_Type                0
Survival_Years                0
Medication_Response           0
Symptom_Progression           0
Year_of_Diagnosis             0
dtype: int64
```

Figure 5: Missing Values in Dataset

3. LITERATURE REVIEW

Recent advancements in deep learning and medical imaging have significantly improved prognosis, disease detection, and individualized treatment for multiple malignancies, including non-small cell lung cancer (NSCLC).*(5)* Deep neural networks, radiomics and big data sources such as ImageNet have advanced diagnostic accuracy and understanding of tumor features.Many studies now show deep learning models can accurately interpret multidimensional medical imaging data. For example, volumetric CT images were processed with 3D convolutional neural networks (CNNs), allowing for extraction of important factors associated with patients' outcomes. Furthermore, combining clinical data with imaging features indicates that there is potential benefit to a more complete understanding of the path of the disease.In summary, although useful, conventional survival models often fail to incorporate or take advantage of multimodal datasets. [3]*(5)*

As lung cancer remains one of the leading causes of cancer death worldwide, there is a need for improved strategies for early detection and personalized treatment. Recent advances in machine learning (ML) have demonstrated promise to provide better prognostic assessment and diagnostic accuracy in oncology, and specifically in lung cancer.Researchers have examined the use of ML-based algorithms to predict lung cancer based on clinical data, patient symptomatology and lifestyle characteristics. Essentially, these methods create data-driven models to identify important risk factors and improve early detection, both critical for effective treatment outcomes.Twelve ML algorithms were tested in a comparative study, using a dataset taken from 310 patient instances with 16 parameters, including demographic parameters (gender, age), behaviors (alcohol, smoking) and symptoms (yellow fingers, cough, chest pain, allergy). The results showed that lung cancer was prevalent most in males who were between the ages of 55 and 75. Symptoms like, yellow fingers, cough, chest pain and allergy were good predictors for lung cancer and very noticeable. The K-Nearest Neighbor (KNN) and Bernoulli Naive Bayes models were the best as early lung cancer predictors of the algorithms studied, and therefore they can be directly applied for clinical situational settings.The results were in line with another study on small clinical datasets which found that KNN and other simpler models had the best performance (92.86% accuracy), followed by Naive Bayes variants (91%). This suggests that when data are limited, it is possible that models which assume the independence of features and simple categorization rules perform better. These studies do also show the need for larger, diverse datasets and research into alternative machine learning methods, as well as limitations in feature scope and dataset size[4].*(13)*

With the possibility of early diagnosis and improved quality of care, machine learning (ML) is increasingly utilized as a method for analyzing risk factors for lung cancer and predicting severity. Data preparation, feature selection, building models, and evaluation are some important aspects that are discussed in the literature. Data Preparation and Missing Data,When aspects of medical data including that pertaining to lung cancer have missing values, the absence of good imputation would bias models. There has been a vast array of imputation methods explored, from simple techniques such as mode imputation for categorical variables to much more complex model-based and multiple imputation procedures. Considering the prevalence of categorical features in lung cancer datasets, the mode imputation method is frequently used since it is a basic way to keep

data without significant bias Extreme Gradient Boosting (XGBoost) has been determined to predict severity of lung cancer with the highest accuracy (up to 99%) and the lowest error rates than any other machine learning algorithms. .[5].*(15)*

The research paper's reviews several Machine Learning techniques for lung cancer classification. It focuses on the morphological implementations of the various systems using deep learning for lung cancer prognosis and early diagnosis. For segmentation and recognition processes, feature extraction, GLCM, brightness estimates, and visual quality evaluation take place and in previous literature, it was reported that if the search algorithm is employed instead of exiting image that it lucidly improves the early stage of lung cancer analysis since imaging position changes can cause trivial differences in the quality of image products, which can be improved overall.*(16)* Again, classifiers will have previously-sourced datasets indicating that they detected successfully with a generalized error rate amelioration in early stage lung cancer studies within the tests.[6]*(7)*

In cancer care, the field has been revolutionized in recent years by advancements in deep learning and medical imaging, providing more accurate illness detection, prognostic predictors, and individualized treatment plans. Deep neural networks and radiomics have been employed to decipher complex medical images to better understand tumor characteristics and patient outcomes. With an increase in big datasets like ImageNet facilitating the development of reliable algorithms that can pull significant features from high-dimensional data. In traditional survival analyses, the variability in tumor biology and patient characteristics is typically done by using single-modality data with little consideration for the true range of tumor biology and clinical factors.*(2)* Multimodal approaches have aimed to address this limitation by creating deep learning frameworks that incorporate complementary input from multiple modalities.[7].Lung cancer continues to be the predominant cause of cancer mortality, so improved risk stratification is indicated. Older recommendations such as USPSTF are based predominantly on age and smoking history, failing to catch many at risk.AI algorithms such as XGBoost now incorporate larger clinical and behavioral information, producing excellent performance (ROC-AUC 0.82 for PLCO, 0.70 for NLST) and appropriate calibration (Brier score 0.044).*(5)* These algorithms enhance personalized risk estimation and have improved recall and discrimination compared with USPSTF.Tools for interpretability such as SHAP increase user trust, while web applications such as MyLungRisk allow these models to be easily available.[8]. Fast track clinics have been initiated in Denmark for speeding up the diagnosis of LC, but although these diagnostic procedures are applied to a good number of such patients, there are still patients diagnosed with LC who are not exposed to these procedures, among them lung cancer patients in Denmark. As compared to non-LC patients, LC patients were older, female, and had greater histories of smoking, current or past. Statistically significant results for certain lab tests are present, as alkaline phosphatase, bilirubin, CRP, LDH, leucocytes, neutrophils, and basophils were all elevated and albumin, eosinophils, hemoglobin, INR, sodium, and platelets seemed to be reduced. These statistically significant (p ¡ 0.001) results have the potential to define prognostic features of LC from routine blood monitoring.[9]vWith over 236,000 new cases and 130,000 deaths each year, lung cancer is the leading cause of cancer-related deaths in the US (Study Summary). Behavioral, genetic, environmental, and psychosocial factors all contribute to its development.it continues to be the primary risk factor, accounting for more than 80% of cases. Risk is greatly increased by heavy smoking and a family history of lung cancer, especially in first-degree relatives. According to behavioral trends, people with lung cancer tend to use electronic cigarettes more and make more attempts to stop smoking. Patients with lung cancer, particularly those with secondary uases, are more likely to suffer from psychiatric disorders such as anxiety, depression, substance use disorders, and sleeplessness.The study emphasizes the need for more genetic research and longitudinal approaches to improve risk prediction using data from the All of Us Research Program [10].DeepCAD-NLM-L has demonstrated these capabilities with nodule detection, malignancy prediction, and clinical data across multiple time periods. DeepCAD-NLM-L has been validated in large datasets, including the National Lung Screening Trial (NLST)(16) and the SUMMIT study cohort(2), and achieves state-of-the-art performance (AUC 88%), which is comparable to or better than the performance of radiologists and existing traditional risk models (e.g., Brock model).*(18)* This demonstrates that it is indeed feasible to build fully automated, longitudinal, multi-modal deep learning systems that can be embedded into clinical workflows and improve early lung cancer detection.[11]Application of some machine learning (ML) approaches, including logistic regression, support vector machines (SVM), and decision trees, for stage classification of cancer and prediction of disease progression has been explored widely.*(3)* It was shown that combination techniques, such as voting classifiers, enhance performance, with some accuracy levels up to around 94% being achieved in classification of lung cancer stages. It emphasizes how some models can perform so well and be combined with high predictive accuracy and robustness.Handling medical data of high dimensions is still very challenging. To reduce dimensionality, eliminate noise, and identify the most relevant features, techniques such as Principal Component Analysis (PCA) and correlation-based feature selection are commonly applied. With a view to achieve maximum model performance and ensure precise predictions, data preprocessing is indispensable, such as normalization strategies like Min-Max scaling.[12]Mangukiya's breast cancer detection model used a range of machine learning methods to detect breast cancer.*(8)* Analysis of the Wisconsin breast cancer data was the main objective of the research. Even though they used a range of algorithms, XGBoost had the highest accuracy of all at 98.24%.[13]

4. PROPOSED METHODOLOGY

All these will facilitate effective communication Conclusion for your stakeholders in the model. Thus, The functioning is clearly documentation and imagining it It is transparent and copyable for itself, and therefore, it is transparent. Encourage further research on models and reforms.

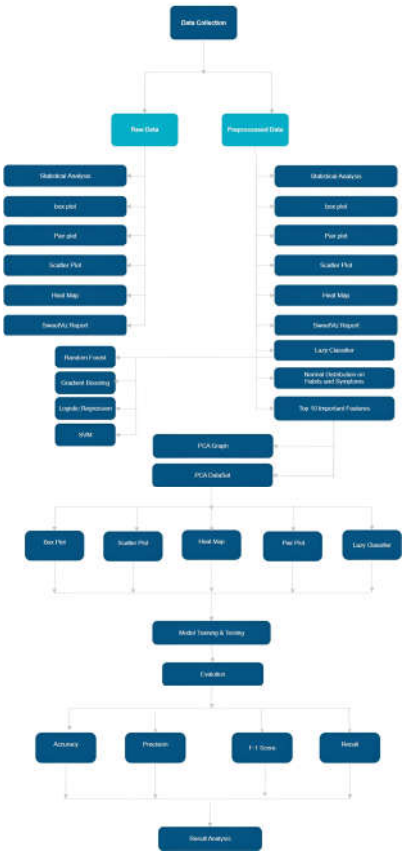


Figure 6: Flow Chart

- A flowchart related to machine learning-based disease identification and diagnosis is shown in Figure 6. this flowchart is a summary of our whole work how we collected data ,how we pre-process data and do traning and testing on the data and make a model.This is the sequential process:

- i)Collect data: The first step in the process is to compile the necessary medical information to diagnose diseases.ii)Data Preprocessing: This simply means that the data must be cleaned before being used for analysis. iii)Data Removing: This phase eliminates any mistakes, missing numbers or unnecessary information that can reduce the quality of data. To explain the model reactions in a meaningful way, it normalizes the scales of the variable in the data.iv)Feature selection is suitable of features that reduces noise and model performs performance. v)Feature extraction: Additional, improved features can be obtained or modified from the features in this challenge. vi)Data Separation for Testing and Training: The dataset is divided into two sections: one for training the model and the other for assessing its performance.vii)Choose a Model: Select the machine learning model for the disease diagnosis from among the several that are offered. Model Training: The training data will have been used to train the chosen model.

- Let's start with breif explaination ,As we have futher discuss in data description about the data collection.we have two type of dataset one is Raw dataset and another is numerical dataset.We have pre-process both dataset and perform statistical analysis,EDA,Corellation matrix,lazzy classifier and lastly we have applied normal distibution on specific features.we have collected top 10 features from the dataset and perform Pca(Principle Component Analysis) on it.we have made pca graph as well as dataset.after this we have perfom statistical analysis on it.*(14)*

- After the altitude, the data will be separated in training and testing the most and important characteristics were found using PCA. The performance of the model will be evaluated using matrix such as accuracy, accurate, recall and F1-score, and its advantages and disadvantages will be exposed in light of conclusions.

- The purpose of this study is the evaluation of performance and contrast of the model. With the ultimate goal of identifying any association present in the dataset, pre–processing to fill the missing values and normalizes the features, and collecting data, the Exploratory Data Analysis (EDA) aims to integrate all the ideological structure eventually.After all these we are ready to predict the lung cancer.Here is the step-by-step explanation.

***A.STATISTICAL ANALYSIS ON RAW DATA*:**

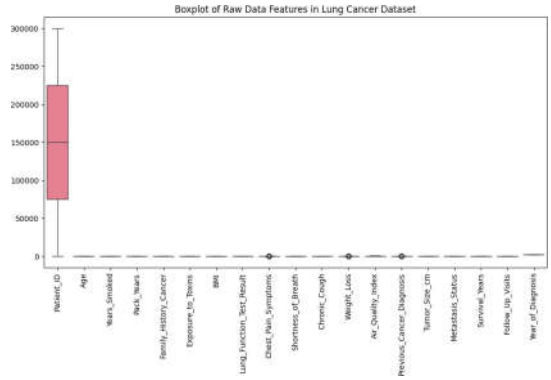


Figure 7: Box Plot

- **BOXPLOT ON RAW DATA**:(See fig :7)This box-plot (see Figure 6) shows the values and contributions of features in the data. Although box plots are unbalanced, they will be balanced after the data is preprocessed. The disparity results from intriguing facts that show a far wider variability in some traits than others, such as smoking history and habitat. This variance is a result of the dataset's excessive size.

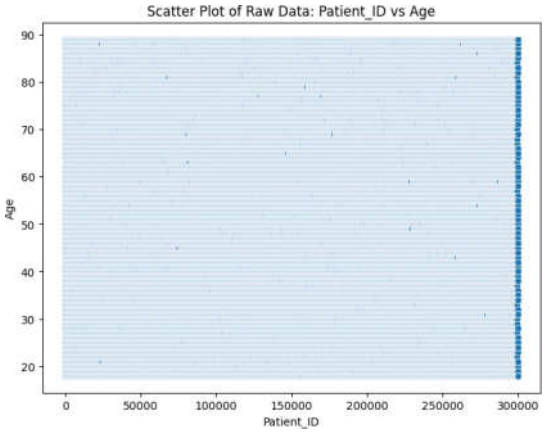


Figure 8: Scatter Plot

· **SCATTER PLOT ON RAW DATA**:(see fig 8)The relationship between numeric variables is displayed in this graph using statistics and data visualization. With each plot point dataset matches a pair of values. we have uses (25%)of dataset because the dataset is too large to use the(100%)dataset.X-Axis and Y-Axis represent the relationship between features.

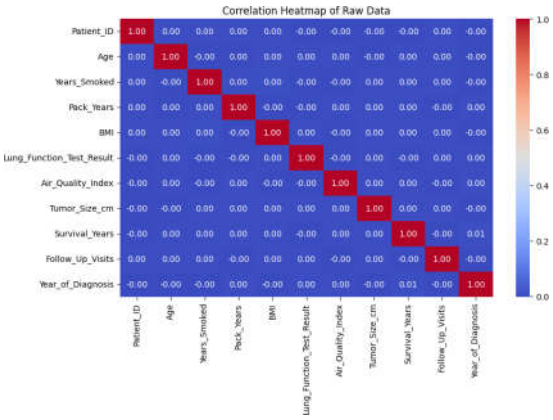


Figure 9: HeatMap

· **HEATMAP OF RAW DATA**:(see fig 9)It is a correlation matrix among the features.the technique that uses colors to indicate values in . It is frequently used to highlight patterns, correlations, and density in large datasets.It is clear that a large number of features have very little link with each other.

(6)

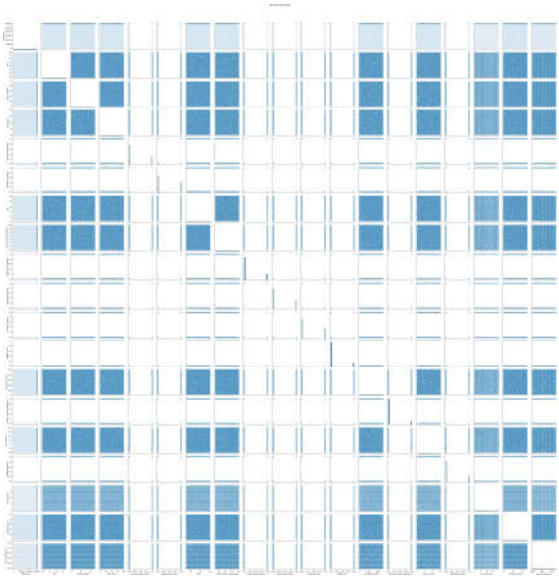


Figure 10: Pair Plot

· **PAIR PLOT OF RAW DATA**:In the given (fig 10) a dataset, the partner correlation between several numeric variables can be seen using a pair plot, also known as scatterplot matrix. Searching data analysis (EDA) often uses it to understand distribution, trends and correlations among the features.

B.STATISTICAL ANALYSIS ON NUMERICAL DATA:

We have discuss before about dataset (see in fig/;1).these figure dicuss about the summary of the datadset. We have a normal distribution on the featurers of the dataset to see the difference according to the data.Here are some normal distribution of the following:

· **NORMAL DISTRIBUTION IN NUMERICAL DATA:**

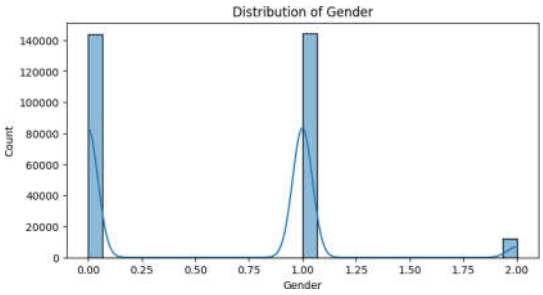


Figure 11: distribution of gender graph

Normal distribution of gender shown in (Fig 11). This graph shows that the male denoted with (0) and the female denoted with (1) and the transgender denoted with (2). Male and female have a high chance of having lung cancer whereas transgender has a low chance of having lung caner.*(12)*

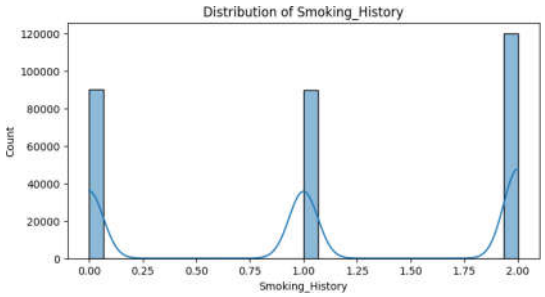


Figure 12: distribution of smoking history graph

Normal distribution based on smoking history shown in(fig 12)this graph shows that the current is indicated with (0), the former is indicated with (1) and never with (2). Those who currently and were used to smoke have a lower chance of having lung cancer compared to those who never smoke.

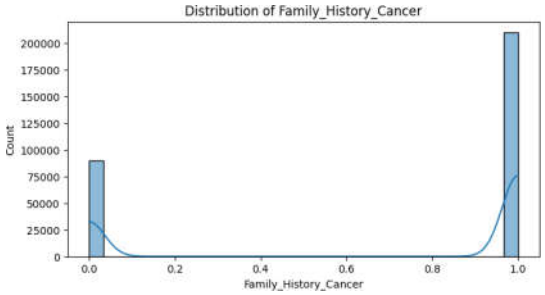


Figure 13: distribution of family history graph

Normal Distribution based on family history: It looks like it does not matter that if a person belongs to a family of lung cancer exposures or not both people can have cancer,illustrate in the given (fig: 13)

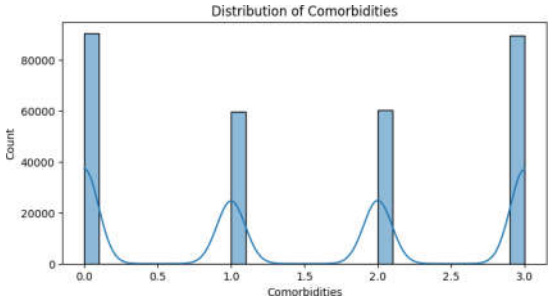


Figure 14: distribution of comorbidities graph

Normal distribution based on comorbidities shown in (fig 14)this graph shows that the Hypertension is indicated with (0) COPD is indicated with(1), Diabetes is indicated with (2) and none is indicated with (3). The graph shows that COPD and a diabetic person have low chance to have cancer, while the person having hypertension and none has more likely to have cancer.

After normal distibution we have perform several operations on the numerical dataset:

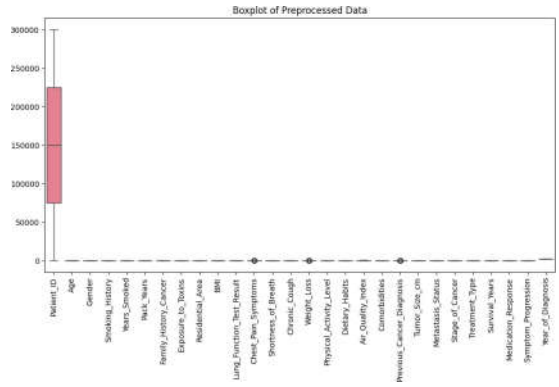


Figure 15: Box Plot

- **BOX PLOT ON NUMERICAL DATA:**(see fig 15)shows the values and contributions of features in the data. This box plots are balanced, they were balanced after the data is preprocessed.After preprocessing data the plot distributed between -1 to 1.There are three features which dominates among all and they are smoking history,residential area and comorbidities.

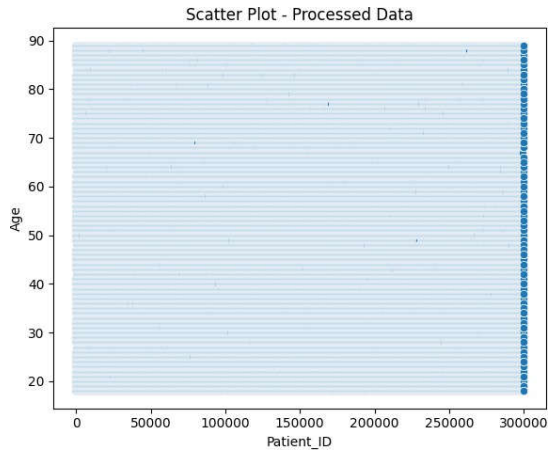


Figure 16: Scatter Plot

- **SCATTER PLOT ON NUMERICAL DATA:** (see fig 16)We have use only 25% of dataset to map scatter plot. It discuss the relationship between the features.Dataset has high dimensionality and low correlation among the features.

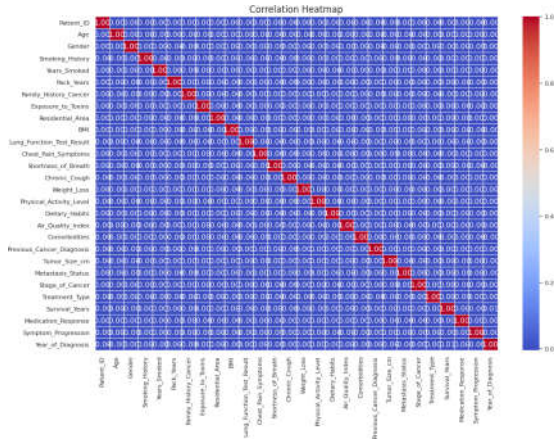


Figure 17: Heatmap

· **HEATMAP ON NUMERICAL DATA:**(see fig 17) it is a correlation matrix which were represented using different colour. Some features were extracted during the preprocessing of the data set.The range is between(0 to 1).

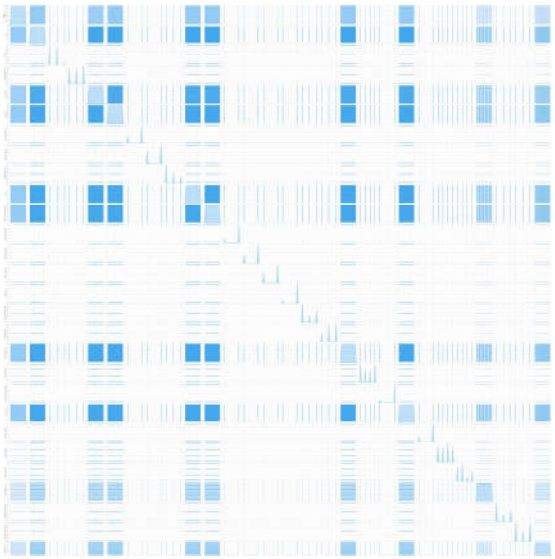


Figure 18: Pair Plot

· **PAIR PLOT ON NUMERICAL DATA:** In the given (fig 19) a dataset, the partner correlation between several numeric variables can be seen using a pair plot, also known as scatterplot matrix. Searching data analysis (EDA) often uses it to understand distribution, trends and correlations among the features.

C.STATISTICAL ANALYSIS ON PCA (Principal Component Analysis) DATA: **DataFrame with PCA results:**

Principal component analysis (PCA) was employed to solve these issues. PCA captures the optimal patterns while shrinking the data's dimensionality according to variance. Even normalized data face problems such as redundancy, overfitting, and difficulty in detecting relationships between data when dealing with high-dimensional data. Scaling and normalization ensure that the data contributes equally to the model.*(4)*

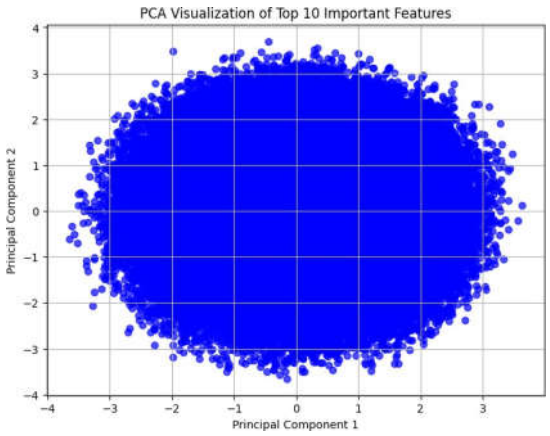


Figure 19: DataFrame with PCA results

SCATTER PLOT OF PCA

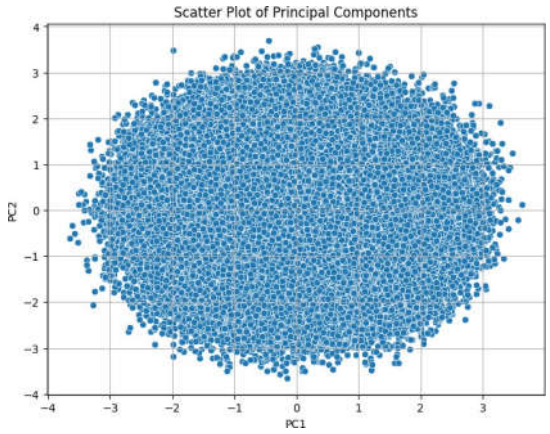


Figure 20: PCA scatter plot

1. Principal Components (PC1 and PC2) are the axes.

The Principal Component 1 (PC1) is the X-axis. The Principal Component 2 (PC2) is the Y-axis. These two components are new PCA-constructed axes that maximize your data's variance.

2. Each Point Is a Sample

One data point (or row) from your original dataset is a single dot that represents it.Every point of data is now plotted based on its coordinates in the PC1–PC2 space following the transformation of the data with PCA.(see in fig:20)

BOX PLOT OF PCA

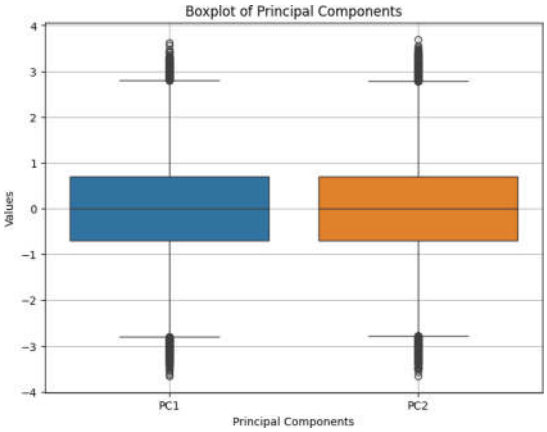


Figure 21: PCA Box Plot

In given fig(see fig 21) Principal Component 1 (PC1) The data is centered if the median is approximately 0.Between Q1 and Q3, the interquartile range (IQR) is between -0.4 and 0.4. Outside the whiskers, there are many outliers, especially near ±3. Principal Component 2 (PC2) is also at 0.IQR and PC1 are very similar, showing the same spread and variance.Outliers are slightly higher or lower than PC1, but overall it's very balanced.

5. RESULT ANALYSIS

We have now successfully trained and tested our model using appropriate performance metrics without any caveats, we can now compare the performance of the base models to our own. The suggested ensemble model is better than any of the three independent base models for Accuracy, Recall, F1-Score, and Precision. Now, we will evaluate the percentage improvement (%imp) to see how much better the proposed model is in comparison to the simple lazy classifier models used.*(1)*

6. CONCLUSION

We thank the contributors and researchers who provided the datasets used in this study.This study has indicated that Machine learning has a considerable potential in determining the phases of lung cancer,where early diagnosis would be prioritized for improved results.With a patient dataset of 3,00,000 and PCA and correlation matrices methods, we were able to determine the most significant improve features and minimize noise in the data. we utilized top four model to achieve better performance.

References

[1] Anonymous. Enhancing lung cancer early detection: A hybrid ensemble model. *Journal of Electrical Systems*, 20(10s), 2024.

[2] Shahab Aslani, Pavan Alluri, Eyjolfur Gudmundsson, Edward Chandy, John McCabe, Anand Devaraj, Carolyn Horst, Sam M Janes, Rahul Chakkara, Daniel C Alexander, et al. Enhancing cancer prediction in challenging screen-detected incident lung nodules using time-series deep learning. *Computerized Medical Imaging and Graphics*, 116:102399, 2024.

[3] Pierre-Louis Benveniste, Julie Alberge, Lei Xing, and Jean-Emmanuel Bibault. Development and external validation of a lung cancer risk estimation tool using gradient-boosting. *arXiv preprint arXiv:2308.12188*, 2023.

[4] Birte Bomhals, Lara Cossement, Alex Maes, Mike Sathekge, Kgomotso M. G. Mokoala, Chabi Sathekge, Katrien Ghysen, and Christophe Van de Wiele. Principal component analysis applied to radiomics data: Added value for separating benign from malignant solitary pulmonary nodules. *Journal of Clinical Medicine*, 12(24):7731, 2023.

[5] Urmila Chandran, Jenna Reps, Robert Yang, Anil Vachani, Fabien Maldonado, and Iftekhar Kalsekar. Machine learning and real-world data to predict lung cancer risk in routine care. *Cancer Epidemiology, Biomarkers & Prevention*, 32(3):337–343, 2023.

[6] V. Deepa Priya, S. Selvabalaji, M. Rithesh, M. Manikandan, V.J. Sanjay Prabhu, and A. Akash. Comprehensive survey on exploratory data analysis and machine learning approaches for lung cancer detection. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 3(3):724–733, 2025.

[7] Demeke Endalie and Wondmagegn Taye Abebe. Analysis of lung cancer risk factors from medical records in ethiopia using machine learning. *PLOS Digital Health*, 2(7):e0000308, 2023.

[8] Ricco Noel Hansen Flyckt, Louise Sjodsholm, Margrethe Høstgaard Bang Henriksen, Claus Lohman Brasen, Ali Ebrahimi, Ole Hilberg, Torben Frøstrup Hansen, Uffe Kock Wiil, Lars Henrik Jensen, and Abdolrahman Peimankar. Pulmonologists-level lung cancer detection based on standard blood test results and smoking status using an explainable machine learning approach. *Scientific Reports*, 14(1):30630, 2024.

[9] Mohammad Hassan, Kiran Bashir, Syed Taqi Abbas, Kashif Arshad, Kaneez Fatima Parveen, Farooq Ahmad, Sidra Bashir, and Muhammad Abbas. Artificial intelligence for early detection of lung cancer: a systematic review. *NPJ Digital Medicine*, 6(1):1–13, 2023.

[10] Y Li, X Wang, L Zhang, W Li, Y Zhang, Y Wang, X Chen, and J Li. The value of artificial intelligence in the diagnosis of lung cancer: A systematic review and meta-analysis. *Frontiers in Oncology*, 13:10035910, 2023.

[11] Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang, and Yuan Luo. Machine learning for lung cancer diagnosis, treatment, and prognosis. *Genomics, Proteomics Bioinformatics*, 20(5):850–866, 2022.

[12] Yifan Lu, Xia Wang, Li Zhang, Wei Li, Yujie Zhang, Yao Wang, Xiaoyan Chen, and Jun Li. Gender disparities in lung cancer incidence in the united states from 2001 to 2019: A population-based study. *Scientific Reports*, 13(1):1–12, 2023.

[13] Satya Prakash Maurya, Pushpendra Singh Sisodia, Rahul Mishra, and Devesh Pratap Singh. Performance of machine learning algorithms for lung cancer prediction: a comparative approach. *Scientific Reports*, 14(1):18562, 2024.

[14] Mehdi Naseriparsa and Mohammad Mansour Riahi Kashani. Combination of pca with smote resampling to boost the prediction rate in lung cancer dataset. *arXiv preprint arXiv:1403.1949*, 2014.

[15] Geetanjali Paliwal and Umashankar Kurmi. A comprehensive analysis of identifying lung cancer via different machine learning approach. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 691–696. IEEE, 2021.

[16] Vikram R Shaw, Jinyoung Byun, Rowland W Pettit, Younghun Han, David A Hsiou, Luke A Nordstrom, and Christopher I Amos. A comprehensive analysis of lung cancer highlighting epidemiological factors and psychiatric comorbidities from the all of us research program. *Scientific Reports*, 13(1):10852, 2023.

[17] Jiao Shen, Wei Chen, Xianliang Zhang, Wei Zhao, and Meng Li. Machine learning-based prediction of lung cancer risk using demographic, lifestyle, and clinical features: a comprehensive study. *BMC Medical Informatics and Decision Making*, 21(1):1–11, 2021.

[18] Yujiao Wu, Jie Ma, Xiaoshui Huang, Sai Ho Ling, and Steven Weidong Su. Deepmmsa: A novel multimodal deep learning method for non-small cell lung cancer survival analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1468–1472. IEEE, 2021.

[19] Y Zhao, S Wang, Y Zhang, Y Wang, Y Wang, and Y Zhang. Deep learning algorithms for diagnosis of lung cancer: A systematic review and meta-analysis. *Cancers*, 14(16):3856, 2022.