Lung Cancer Prediction using Random Forest and Support Vector Machine Algorithm

BUSHRA FATIMA KHAN[1], SHIFA ANSARI[2], MAIMUNA AHMED[3]

[1]*UG Student, Gurunanak College of Pharmacy, India*

[2]*Department of Electrical Engineering, Shri Ramdeobaba College of Engineering and Management, India*

[3]*Department of Information Technology, Yashwantrao Chavan College of Engineering, India*

Abstract— Lung cancer is a major global health concern, accounting for a significant proportion of cancer-related deaths worldwide. It is characterized by the uncontrolled growth of abnormal cells in the lungs, which can spread to other parts of the body if not detected and treated early. Therefore, early diagnosis of lung cancer is critical for improving patient outcomes and survival rates. Machine learning techniques have shown significant promise in disease prediction by identifying lifestyle patterns and existing health concerns. In this paper, we deployed predictive machine learning models like Random Forest (RF) and Support Vector Machines (SVM) for early diagnosis of lung cancer using survey lung cancer dataset. Additionally, we also built data visualizations to identify patterns and utilized evaluation metrics like precision, recall, f-1 score and support to check accuracy our models.

Index Terms— Lung Cancer, Machine Learning, Random Forest, Support Vector Machine

I. INTRODUCTION

According to Arthur Samuel Machine learning (ML) is a subset of artificial intelligence (AI) that allows machines to learn and improve from experience without being explicitly programmed. ML uses algorithms to analyze large amounts of data, identify patterns and correlations, to make decisions and predictions. In recent years, ML Algorithms have gained wide recognition for early diagnosis of certain live threatening conditions like Cancer, Parkinson's disease, Asthma, Alzheimer's disease and many more.

Lung cancer is an aggressive disease with high morbidity worldwide, with an estimated 2.2 million new cases and 1.8 million deaths in 2020. Globally, it is the leading cause of cancer in men and the second most deadly form of cancer in women after breast cancer. The prevalence of cancer and mortality rate is almost twice as high in men as in women, although the incidence and mortality rates are different between men and women in different regions of the world.[1] Research studies have shown that smoking is the leading of lung cancer in 80% of the patients worldwide, Risk factors such as exposure to radiation, carbon products, sulfur dioxide and asbestos are highly reversible with smoking cessation, occupational safety and evidence-based are the preventive measures that can be implemented to reduce the severity of the disease.[2] Lung cancer kills people because it is frequently discovered when the disease has progressed to an advanced stage. Effective early detection, a thorough etiology, and the right medications all contribute to lung cancer treatment. Consequently, it is critical to diagnose lung cancer as soon as possible, particularly when screening high-risk groups (smokers, exposure to fumes, oil fields, hazardous workplaces, etc.), and there is a pressing need to find new biomarkers.[3] A number of conventional methods, such as cytology sputum, serum test, urine test, computerized tomography (CT), positron emission tomography (PET), and chest X-ray (CXR), can be utilized in the early identification of lung cancer.[4] Machine learning in recent years has gained immense popularity for both disease diagnosis and prognosis. Several ML models like Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, XGboost, and Gradient Boosting (GB) can be implemented for early diagnosis of lung cancer.

II. LITERATURE REVIEW

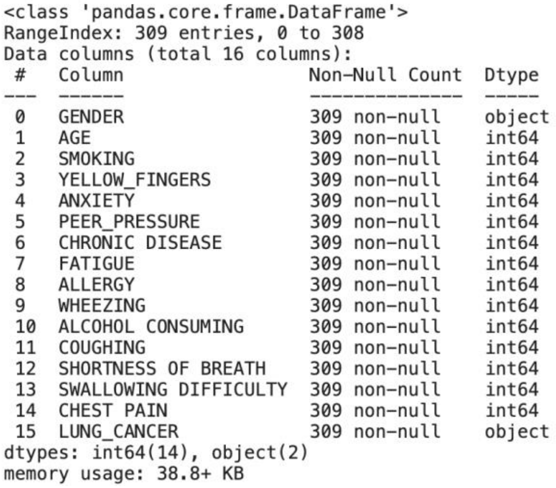
1. Radhanath Patra et al [5] [2020] concluded that Radial Basis Function Support Vector

Machine (RBF SVM) classifier with an accuracy score of be 81.25% was the most effective SVM classifier for predicting lung cancer on UCI machine learning dataset.

2. Ying Xie et al [6] [2021] while working with metabolomics and biomarker genes found that Naïve Bayes could be the next exploitable tool for early lung tumor prediction.
3. Sharmila Nageswaran et al [7] [2022] experimented with 83 CT scans from 70 distinct lung cancer patients and deployed classifiers like ANN, KNN and RF. It was reported that Artifical Neural Network or ANN was the best model for image processing.
4. Belal Alsinglaw et al [8] [2022] proposed a predictive Length of Stay (LOS) framework using explainable machine learning techniques like random forest for lung cancer patients in ICU setting.
5. Eali Stephen Neal Joshua et al [9] [2020] in their extensive review found that innovative methods must be used for medical picture noise reduction, particularly for MRI (Magnetic Resonance Imaging) and DiCOM (Digital Imaging and Communications in Medicine) images.
6. Mehdi Amini et al [10] [2021] proposed a framework that facilitated identifying the best time-to-event prognostic methods for predicting survival of NSCLC patients using various PET/CT radiomics, including both single- and multimodality approaches.
7. Shuhei Ishii et al [11] [2022] [11] in their research found gene testing in lung cancer tumors challenging, costly and time consuming, as a preventive solution they developed a gene alteration prediction model for lung cancers by machine learning based on cytologic images.

III. DATASET DESCRIPTION

We acquired this open-source lung cancer survey dataset from Kaggle (A Data Science competition platform and online community for data scientists and machine learning practitioners under Google LLC). This dataset contains 309 rows and 16 columns. The dataset can be obtained from: https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64 
 2   SMOKING                309 non-null    int64 
 3   YELLOW_FINGERS         309 non-null    int64 
 4   ANXIETY                309 non-null    int64 
 5   PEER_PRESSURE          309 non-null    int64 
 6   CHRONIC DISEASE        309 non-null    int64 
 7   FATIGUE                309 non-null    int64 
 8   ALLERGY                309 non-null    int64 
 9   WHEEZING               309 non-null    int64 
 10  ALCOHOL CONSUMING      309 non-null    int64 
 11  COUGHING               309 non-null    int64 
 12  SHORTNESS OF BREATH    309 non-null    int64 
 13  SWALLOWING DIFFICULTY  309 non-null    int64 
 14  CHEST PAIN             309 non-null    int64 
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

Fig 1. Attributes of Dataset

IV.METHODOLOGY

The architecture of proposed lung cancer prediction model is depicted in Fig 2. After importing the survey lung cancer dataset in jupyter notebook, we performed iterative processes like data preprocessing and feature engineering to identify patterns, anomalies and build predictions. Additionally, we also plotted data visualizations like heatmap, box plot and histograms to depict relationship between different features. After successful data pre-processing, the dataset is split into training and testing. 80% of the dataset is used for training, while 20% is used for testing. We deployed machine learning algorithms like random forest and support vector machine. Lastly, for evaluating performance of both the models we utilized metrices like precision, recall, f-1 score and accuracy and chose the best one. The architecture of our proposed model is shown in Fig 2.

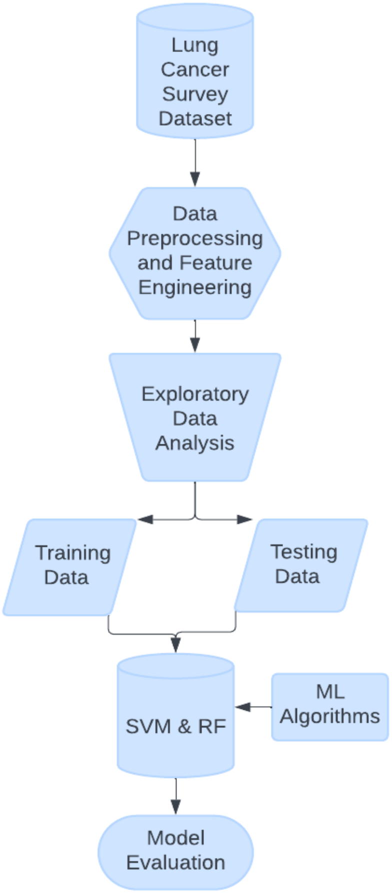

Lung Cancer Survey Dataset
Data Preprocessing and Feature Engineering
Exploratory Data Analysis
Training Data
Testing Data
SVM & RF
ML Algorithms
Model Evaluation

Fig 2. Architecture of proposed lung cancer prediction model

A. Data Pre-Processing and Feature Engineering

After importing the raw survey lung cancer dataset, we performed data pre-processing and feature engineering to remove all the null, duplicate and missing values. We also converted numeric data into categoric data for identifying our features easily.

B. Exploratory Data Analysis (EDA) and visualizations

EDA or Exploratory Data Analysis is a technique used to analyze datasets in order to detect patterns, relationships, and anomalies. We plotted box plot, histograms and correlation heatmap to identify patterns in our data.

C. Machine Learning Model Building

After successful data pre-processing and visualizations, the data set is split into training and testing. 80% of the dataset is used for training, while 20% is used for testing using train test split. We experimented with the following machine learning algorithms:

1. Support Vector Machine (SVM)

Support Vector machines or SVMs are systems that utilize a hypothesis space consisting of linear functions in a high-dimensional feature space. They are trained using a learning algorithm from optimization theory that incorporates a learning bias based on statistical learning theory.[12] For a hyperplane equation SVM is represented as:

$$f(x)=sign(w\cdot x+b)$$

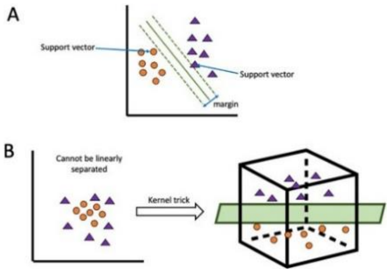


Fig 3. SVM Model

2. Random Forest

A random forest is a classifier consisting of a collection of tree structured classifiers {h(x,Θk), k=1, ...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. [13]

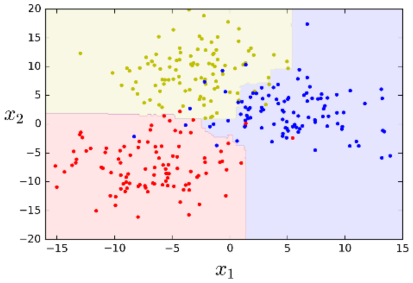


Fig 4. Random Forest Classifier

D. Model Evaluation

Model evaluation in machine learning refers to the process of assessing how well a machine learning model performs on a given task. The goal is to determine the model’s accuracy, generalizability, and suitability for solving the problem it was designed for.

We utilized metrics like accuracy, precision, recall, f-1 score and support to evaluate accuracy of both the algorithms.

V. EXPERIMENTAL RESULTS

The proposed machine learning approach for lung cancer prediction and visualizations were implemented in the Python programming language. Correlation Heatmap depicts relationship between different features like age, smoking, yellow fingers, chronic disease, coughing etc. As we can see in Fig 5. there is multicollinearity between different features.

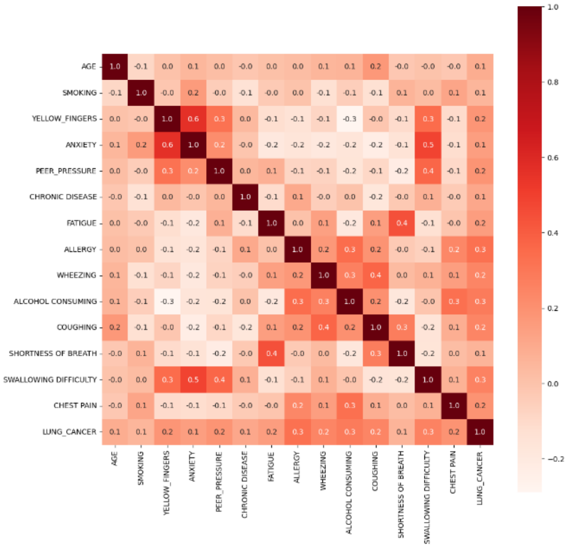


Fig 5. Correlation Heatmap that depicts relationship between different features

Box Plot shows the relationship between age and gender. We found that the interquartile range (the box itself) is similar for both genders, indicating similar variability in age distribution among the middle 50% of the samples for both genders and the median age for both male and female appears almost same.

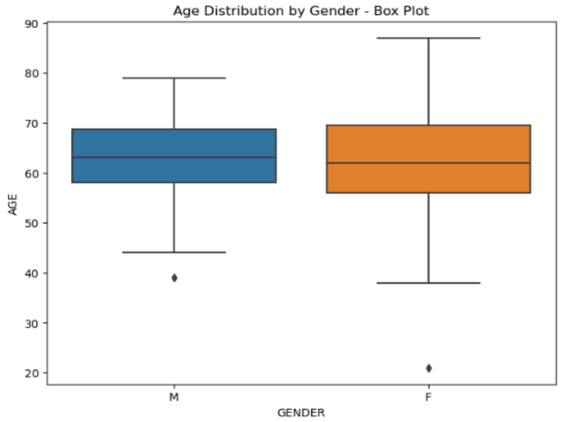


Fig 6. Box plot that shows the relationship between age and gender

Histograms showing specified conditions in terms of 'Yes' and 'No' like Smoking, Allergy, Yellow Fingers, Anxiety, Fatigue, Allergy etc. The majority of conditions seem to have more "No" responses than "Yes" ones, suggesting that a bigger proportion of sample members do not meet these criteria. But 'FATIGUE' and 'SHORTNESS OF BREATH' have a notably greater percentage of 'Yes' answers, indicating that these symptoms are more frequently reported by the participants.

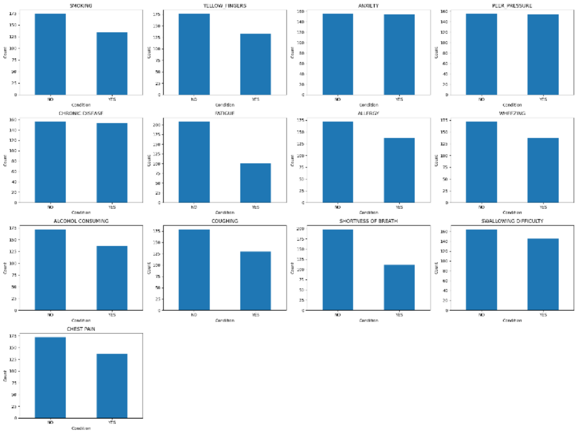


Fig 7. Histograms showing different conditions during survey in terms of 'yes' or 'no'

Circular Histogram shows the frequency of 'YES' response for each condition. This histogram suggests a possible association between the different health disorders by showing a very equal distribution across them. This consistency raises the possibility that the existence of one illness may be linked to the probability of others, all of which may raise the risk of lung cancer

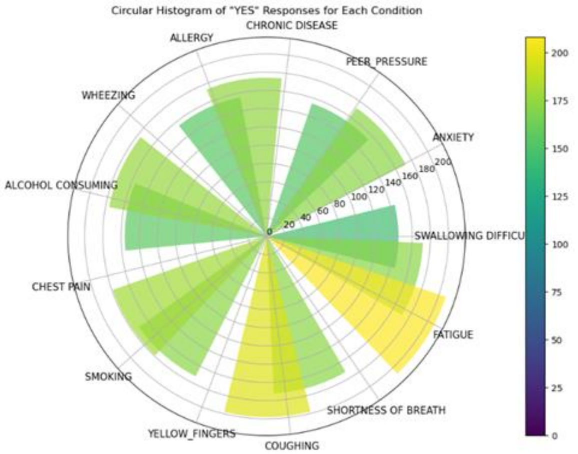


Fig 8. Circular Histogram showing frequency of 'YES' response for each condition

	Precision	Recall	F1-score	Support
0	0.33	0.50	0.40	2
1	0.98	0.97	0.97	60
Accuracy			0.95	62
Macro avg	0.66	0.73	0.69	62
Weighted avg	0.96	0.95	0.96	62

Table 1. Evaluation Metrics of Support Vector Machine Algorithm

	Precision	Recall	F1-score	Support
0	0.50	0.50	0.50	2
1	0.98	0.98	0.98	60
Accuracy			0.97	62
Macro avg	0.74	0.74	0.74	62
Weighted avg	0.97	0.97	0.97	62

Table 2. Evaluation Metrics of Random Forest Classifier.

Based on evaluation metrics like precision, recall, f-1 score:

- The Accuracy of Support Vector Machine is 95.16 %.
- The Accuracy of Random Forest Classifier is 96.7 7 %.

From the accuracy, sensitivity and specificity values t he medical partitioner can choose the appropriate mo del for predicting lung cancer.

CONCLUSION

To recapitulate, leveraging machine learning algorithms machine for lung cancer prediction showed promising results with high testing accuracies. Random Forest Classifier with an accuracy of 96.77% outperformed Support Vector Machine by a mere 1.61%. Integrating Medical Imaging and Generative Artificial Intelligence (GenAI) in lung cancer detection can improve both diagnosis and prognosis of lung cancer.

REFERENCES

[1] Leiter, A., Veluswamy, R.R. and Wisnivesky, J.P., 2023. The global burden of lung cancer: current status and future trends. *Nature reviews Clinical oncology*, *20*(9), pp.624-639.

[2] Huang, J., Deng, Y., Tin, M.S., Lok, V., Ngai, C.H., Zhang, L., Lucero-Prisno III, D.E., Xu, W., Zheng, Z.J., Elcarte, E. and Withers, M., 2022. Distribution, risk factors, and temporal trends for lung cancer incidence and mortality: a global analysis. *Chest*, *161*(4), pp.1101-1111.

[3] Nooreldeen, R. and Bach, H., 2021. Current and future development in lung cancer diagnosis. *International journal of molecular sciences*, *22*(16), p.8661.

[4] Li, W., Liu, H.Y., Jia, Z.R., Qiao, P.P., Pi, X.T., Chen, J. and Deng, L.H., 2014. Advances in the early detection of lung cancer using analysis of volatile organic compounds: from imaging to sensors. *Asian Pacific Journal of Cancer Prevention*, *15*(11), pp.4377-4384.

[5] Patra, Radhanath. "Prediction of lung cancer using machine learning classifier." In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1*, pp. 132-142. Springer Singapore, 2020.

[6] Xie, Y., Meng, W.Y., Li, R.Z., Wang, Y.W., Qian, X., Chan, C., Yu, Z.F., Fan, X.X., Pan, H.D., Xie, C. and Wu, Q.B., 2021. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, *14*(1), p.100907.

[7] Nageswaran, S., Arunkumar, G., Bisht, A.K., Mewada, S., Kumar, J.S., Jawarneh, M. and

Asenso, E., 2022. [Retracted] Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. *BioMed research international*, *2022*(1), p.1755460.

[8] Alsinglawi, B., Alshari, O., Alorjani, M., Mubin, O., Alnajjar, F., Novoa, M. and Darwish, O., 2022. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific reports*, *12*(1), p.607.

[9] Joshua, E.S.N., Chakkravarthy, M. and Bhattacharyya, D., 2020. An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study. *Revue d'Intelligence Artificielle*, *34*(3).

[10] Amini, M., Hajianfar, G., Avval, A.H., Nazari, M., Deevband, M.R., Oveisi, M., Shiri, I. and Zaidi, H., 2022. Overall survival prognostic modelling of non-small cell lung cancer patients using positron emission tomography/computed tomography harmonised radiomics features: the quest for the optimal machine learning algorithm. *Clinical Oncology*, *34*(2), pp.114-127.

[11] Ishii, S., Takamatsu, M., Ninomiya, H., Inamura, K., Horai, T., Iyoda, A., Honma, N., Hoshi, R., Sugiyama, Y., Yanagitani, N. and Mun, M., 2022. Machine learning-based gene alteration prediction model for primary lung cancer using cytologic images. *Cancer Cytopathology*, *130*(10), pp.812-823.

[12] Jakkula, V., 2006. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*(2.5), p.3.

[13] Breiman, L., Random Forests, Machine Learning 45(1), 5-32, 2001.