Lung cancer prediction using ML algorithms

Sambhav Dey   
Department of Electronics and communication engineering   
SRMISTKattankulathur, Tamil Nadu, 603203  
sd8900@srmist.edu.in

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

Atharv Mathur  
Department of Electronics and communication engineering   
SRMISTKattankulathur, Tamil Nadu, 603203  
am6299@srmist.edu.in

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

Eshan Tushar Joshi  
Department of Electronics and communication engineering   
SRMISTKattankulathur, Tamil Nadu, 603203  
ej2305@srmist.edu.in

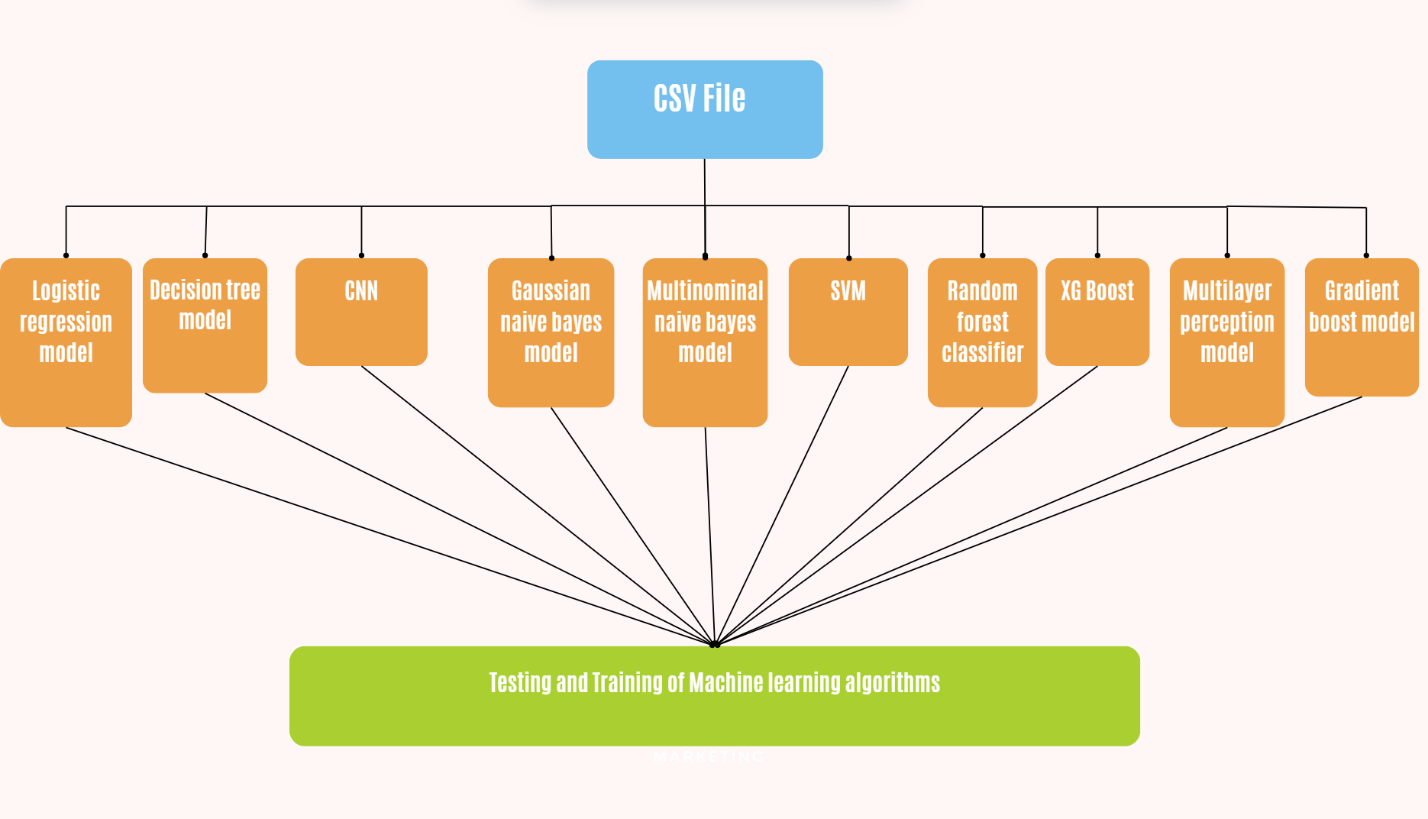
line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or

**Abstract**

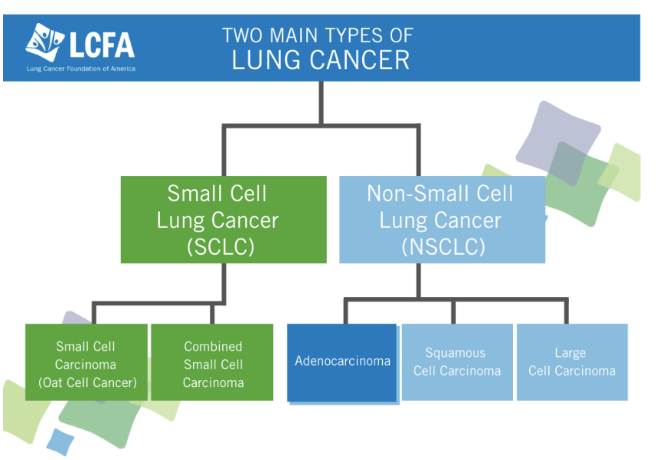
As one of the primary causes of cancer-related death globally, lung cancer underscores the critical need for efficient prediction models to boost early identification and optimize patient outcomes. This work investigates the use of machine learning (ML) approaches to predict the risk of lung cancer using a variety of datasets that include imaging, demographic, and clinical variables. To evaluate their prediction performance, we used a variety of algorithms, such as deep learning models, decision trees, random forests, and logistic regression. We used a feature selection process to find the most important predictors, and then we rigorously cross-validated the model to make sure it was resilient. Accordingly, the Random Forest model has the highest accuracy (94.6%) according to the Stratified K-Fold cross validation. Other models with almost equal accuracy include Gradient Boost, Support Vector Classifier, and XGBoost, while the Multinomial Naive Bayes model has the lowest accuracy (75.7%).Future research aiming at incorporating these predictive models into clinical processes is made possible by this study, which highlights the potential of machine learning to revolutionize lung cancer screening techniques. In the end, our research supports an early detection of lung cancer that is data-driven, as this could greatly lower death rates and improve patient quality of life.

**INTRODUCTION**

At about 18% of all cancer-related fatalities, lung cancer is a serious worldwide health concern. A major obstacle to early detection is its persistence, which frequently results in late-stage diagnoses and dismal prognoses despite improvements in awareness and treatment. Creative methods for early detection and intervention are required due to the complexity of lung cancer pathology and risk factors like smoking, environmental exposures, and genetic predisposition.New directions in disease outcome prediction and improved diagnostic accuracy have been made possible by recent advances in machine learning (ML). Massive datasets from imaging studies, genomic profiles, and electronic health records are used by machine learning (ML) algorithms to find patterns and correlations that conventional statistical techniques might miss. These algorithms show promise for helping with risk stratification and individualized treatment planning in a number of medical specialties, including oncology. This research attempts to investigate how well different machine learning methods predict the risk of lung cancer by analyzing imaging data in conjunction with a thorough examination of clinical and demographic factors. In addition to improving early detection capabilities, our goal is to create strong predictive models that shed light on the variables that influence the onset of lung cancer. In order to improve patient outcomes in lung cancer management, this research aims to establish a framework for integrating machine learning into clinical practice by methodically evaluating various algorithms and incorporating feature selection methodologies. We hope to add important knowledge to the ongoing efforts in precision medicine and cancer epidemiology through this investigation.



In this scenario, machine learning-based prediction of lung cancer is about using many different algorithms to analyze the patient data be it demographic and medical history, genetic data, or even results from imaging in order to estimate the chance of cancer. XGBoost stands out as one of the greatest and most popular ensemble methods available, famous for their efficiency, scalability, and flexibility. It generates several decision trees one after another. The successive decision tree tries to correct the errors of the previous decision tree. Therefore, it is very effective on complex datasets. It works even in cases with missing or highly imbalanced data which often occurs in a medical dataset. However, it is computationally heavy and calls for careful hyperparameter tuning.  
  
The second kind is ensembling with the Random Forest Classifier, which constructs many decision trees in the course of training and then aggregates the result of those to maximize accuracy while avoiding overfitting. It helps in lung cancer prediction because it can handle high-dimensional data plus feature importance, but it would require much more memory, especially in computational resources, given the number of trees it is building.  
  
Decision Trees are simple, interpretable models that split the data on the basis of the feature values. They are easy to understand and interpret but they overfit a lot on small datasets unless techniques such as pruning or ensemble methods, like Random Forest or XGBoost, are used.  
  
K-Nearest Neighbors, or KNN, is a non-parametric algorithm, based upon the majority class of its nearest neighbors in feature space, which will classify a new sample. Although easy to implement, KNN does not perform well in high-dimensional space; most real-world datasets are indeed in high-dimensional space, especially in the medical domain. Moreover, if there are many samples in the dataset, then KNN is bound to be computationally expensive.  
  
Naive Bayes classifiers are made up of the Gaussian Naive Bayes, which assumes a normal distribution of features, and the Multinomial Naive Bayes, which is more suited to discrete data. These are probabilistic models, based on Bayes' theorem that are very fast and work well in small datasets but, for example, feature independence is generally assumed, very far from reality in most medical applications.  
  
Supporting Vector Classifier, or SVC for short, is a classifier which constructs a hyper-plane in a high-dimensional space to classify data points. This method, therefore, is particularly suited for non-linear decision boundaries because the kernel trick can be applied to map data from an original input space to a higher dimension. However, its efficiency deteriorates considerably as the dataset size increases because it may become too slow and too memory-intensive.  
  
Multi-Layer Perceptron, an artificial neural network model could describe complex, non-linear interrelations in the data. It will be composed of the layers with neurons of more than one level. Such a multi-layer neural network will be able to detect even the most subtle patterns within medical data. Despite its power, MLPs require large amounts of data to avoid overfitting and serious computational resources for training.  
  
Gradient Boosting Machines, such as the popular XGBoost algorithm, are sequential models where they construct models along with focusing on errors produced by the previously constructed models. These models are useful for predicting lung cancer since they can pick up complex patterns within the data, but they are computationally expensive to train and sensitive to choices in hyperparameters.  
  
Though very simple and easy to interpret, logistic regression commonly acts as a baseline model for most problems in binary classification scenarios, such as whether a patient has cancer. Thus, it is assumed that there is some linear relationship between the input features, and those output probabilities. However, such an assumption may not be correct for complex medical datasets. Yet, logistic regression can prove to be really efficient and simple to use with satisfactory insight into the importance of each feature in the given classification.  
  
Each of these models has strengths and weaknesses in predicting lung cancer, and the performance may change with respect to the quality and size of a dataset, feature engineering, and model tuning.

 There are two principal types of lung cancers that are classified according to the microscopic appearance of the cancer cells. The two types have other subtypes with different biological characteristics, courses of treatment, and prognoses.  
There are five subcategories of lung cancers. They are separated according to the histological characteristics of the tumors. The five categories include;  
1.Squamous Cell Carcinoma.  
2.Adenocarcinoma.  
3.Large Cell Carcinoma.  
4.Small Cell Carcinoma.  
5.Bronchioalveolar carcinoma.  
Non-Small Cell Lung Cancer (NSCLC) is a common type of lung cancer that accounts for 85-90% of all lung cancer cases. It is further categorized into three major subtypes. The most common and the first one is Adenocarcinoma, which arises from the mucus-producing lining cells of the outer regions of the lungs. As a matter of fact, this lung cancer type is common in both smokers and nonsmokers and tends to grow relatively slower than other types of lung cancer. It often happens in the peripheral parts of the lungs and can be diagnosed earlier, even before symptoms are apparent.  
  
There is another type of NSCLC known as Squamous Cell Carcinoma; this develops within the squamous cells, which make the inner airways of the lungs. They tend to develop in the central parts of the lungs near the bronchi. Squamous cell carcinoma has been found to have a close association with smoking and is more common among males than females. It tends to grow more slowly than other types of lung cancer, but it might also produce symptoms like coughing and chest pain earlier, simply because of the location that is close to the airways.  
  
The third major type of NSCLC is Large Cell Carcinoma. It is less prevalent in comparison with adenocarcinoma and squamous cell carcinoma. This type can occur anywhere in the lung and grows much faster and may spread much quicker, making it more challenging for treatment. This is termed "large cell" because the cancer cells are larger under a microscope than other lung cancers. Large cell carcinoma is more aggressive in their behavior, often leading to early metastasis to other organs.  
  
On the other hand, SCLC forms only smaller 10-15% of lung cancer but has the characteristic of aggressive behavior. SCLC is highly associated with smoking, and tends to grow rapidly before spreading quickly to other areas of the body before it can be detected, often hence begins in the central part of the lungs near the bronchi. Due to its aggressive nature, SCLC is often diagnosed in the later stages, and treatment is mainly through chemotherapy and radiation rather than surgery. Although it initially responds well to treatment, it usually recurs quickly and eventually becomes harder to control in the long term.  
  
In addition to these are rare forms of lung cancer. For instance, lung carcinoid tumors make less than 5% of cases of lung cancer. Carcinoid tumors develop slowly and usually affect younger patients. They can occur in both the lungs and the digestive system. Although they are rare, they can still be malignant and spread through the body requiring treatment.  
Critically, there are various types of lung cancer, and understanding this is valuable because it directly influences the treatment that's provided and the patient's expected prognosis.

## BACKGROUND : LUNG CANCER

Lung cancer is still one of the major killers of human beings in the world today due to its high mortality rate. The main reason for this is late stage diagnosis. Lung cancer, in the early stages, is an issue of great urgency as there is an added chance of survival and the effectiveness of their treatments. Conventional detection techniques like imaging (CT scans and X-rays) and biopsies are expensive, might not be sufficient for a preliminary lung cancer diagnosis, and prove invasive at times. This has opened up exploration into the use of algorithms of machine learning in predicting lung cancer in a better way compared to earlier times, thereby leading to earlier diagnosis and more tailored treatment strategies.  
  
Machine learning algorithms are particularly useful in medical applications such as cancer prediction because they are able to analyze large complex datasets. Such algorithms can process various types of data for the discovery of patterns that suggest the existence of lung cancer, such as patient demographics, genetic data, imaging data, and clinical reports. In this regard, ML application to lung cancer prediction includes the training of models on historical data that will eventually predict the probability in terms of the possibility that a new case may be cancerous or not. By learning from past patient data, the ML algorithms can contribute to classifying at-risk patients, finding potential biomarkers, and even predicting tumor growth or response to treatment.  
Some machine learning models are considered for lung cancer prediction, which include strengths and weaknesses of each. Majority of them applied ensemble methods namely Random Forest and XGBoost because they have an advantage of combining several decision trees to raise the predictive accuracy. These models can handle high-dimensional data, making them suitable for complex datasets. For example, genomic analyses or images taken from medical analyses can be handled by these models. SVMs and Logistic Regression are the other two models, commonly used, because they are quite effective in classification tasks that have to do with binary-type selections, which is common for cancer diagnosis, either cancerous or not.  
  
Deep learning methods especially CNNs have promising results in the analysis of medical images such as CT scans in identifying lung nodules, considered precursors for lung cancer. ANNs and MLPs can model non-linear relations among the variables of interest, thereby providing a better capture of complex interactions between different features in the system, which in this case includes genetic markers and environmental risk factors. These models are able to outperform more classical statistical methods in some situations but usually need great amounts of data and computational resources.  
  
Naive Bayes classification, especially Gaussian Naive Bayes, have been used where probabilistic classification is of interest, while KNN, although simpler, still holds utility in smaller datasets and can focus on the proximity of points in feature space.  
  
There is much promise in the application of machine learning to predict lung cancer, but it does pose significant challenges. The quality of predictions will largely depend on the quality of the input data in terms of accuracy, completeness, and representativeness. A class-imbalanced issue, which is often relevant to the medical data sets, where cancer-positive cases make a minute number, while cancer-negative cases constitute a significant portion, may result in biased models unless proper steps such as resampling or appropriate evaluation metrics are undertaken.  
  
The long-term objective is to predict lung cancer much earlier before it eventually develops; in fact, the best predictions will potentially be made through early detection and personalized healthcare delivery. With advancing research in this area, the combination of ML algorithms with clinical expertise will enhance the accuracy of diagnostic results, reduce false-positive results, and improve patient outcomes. However, despite so much promise, there are many challenges that need to be overcome, including data quality issues, model interpretability issues, and integration into clinical workflows before any widespread adoption can occur.

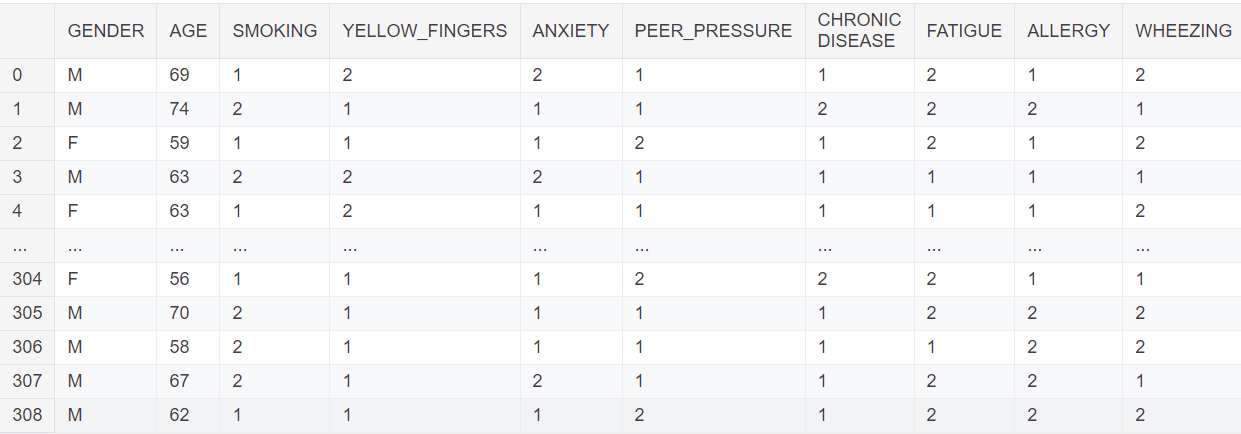
## UCI machine learning repository

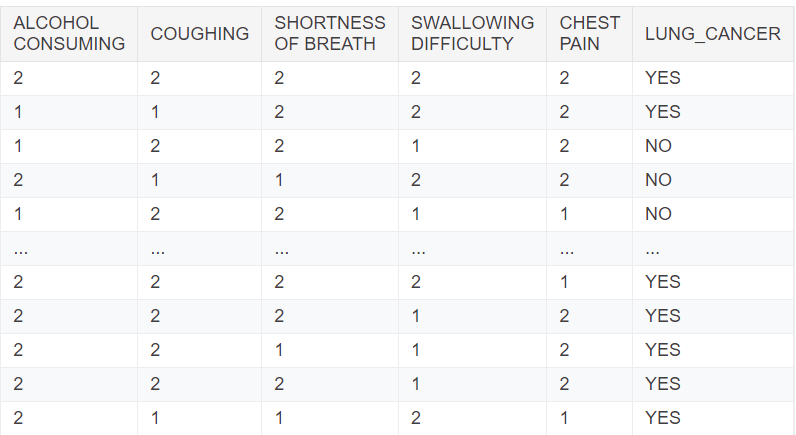
* 1. **Database**

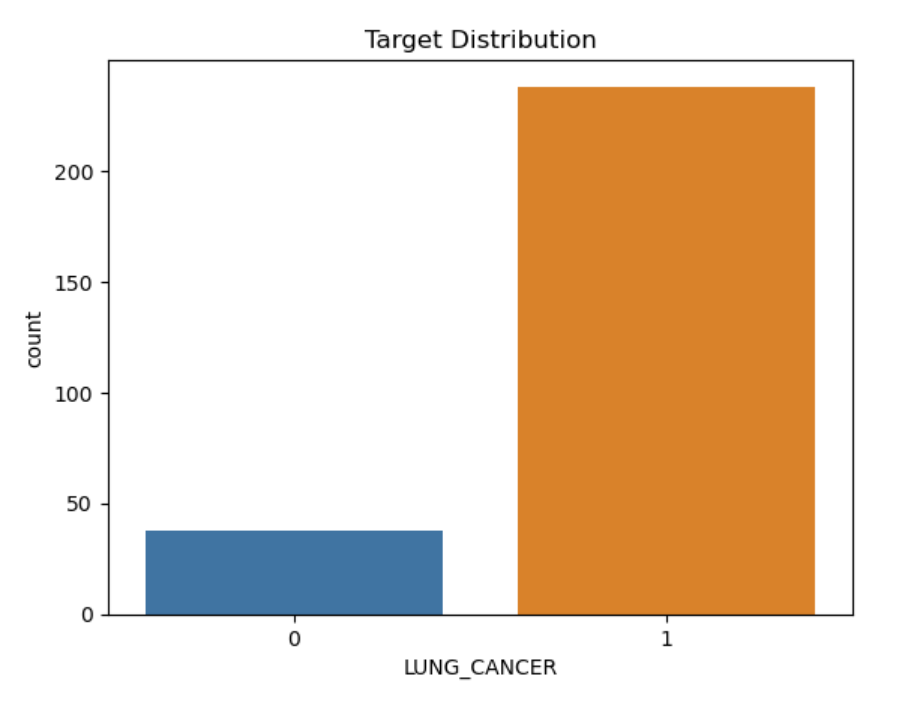
Lung Cancer Prediction Using 10 Machine Learning Classification Models via Scikit-learn Library in Python: Development of a predictive model for the detection of lung cancer in patients. It uses 10 different machine learning algorithms: logistic regression, decision tree, k-nearest neighbor, Gaussian naive Bayes, multinomial naive Bayes, support vector classifier, random forest, XGBoost, multi-layer perceptron, and gradient boosting classifier, to predict the likelihood of having lung cancer based on a range of variables. The data set used in the code has varied columns of gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallowing difficulty, chest pain, and lung cancer. Through such variables' analysis along with the predictive pattern and correlation established using machine learning algorithms, it can help perform accurate risk estimations of a patient for the risk to be developed about lung cancer.

## METHODOLOGY

The methodology to be followed for a prediction of lung cancer using machine learning models comprises data collection, preprocessing, model selection, training, evaluation, and validation. In this context, the purpose is to prepare predictive models that could classify patients appropriately: whether they have lung cancer or not.  
  
The first step is relevant data collection that is collected from medical sources, such as public lung cancer datasets, clinical data repositories, or hospital records. Such datasets usually include features such as patient demographics like age, gender, and smoking history; clinical parameters like tumor size, biomarkers; imaging data like CT scans, X-rays; and genomic data. It is very important to ensure that the dataset collected is of a high quality and representative samples are maintained for good performance of the model.  
  
Data pre-processing involves cleaning and preparing the data for analysis. This stage includes any type of processing which involves missing values, normalization or scaling continuous features, encoding categorical variables, and so on. Techniques applied also feature selection and/or dimensionality reduction, for example Principal Component Analysis retaining only the most relevant ones that can improve the models' efficiency and performance. Note that in the case of lung cancer data, which is usually imbalanced-there are more negatives than positives-the technique could be deployed to handle imbalances with the application of either SMOTE, which stands for Synthentic Minority Over-sampling Technique, or undersampling.  
  
The selected models from the above list of machine learning models then are applied in the prepared data set. This makes XGBoost useful for imbalanced data handling as it uses an ensemble approach via decision trees, which are trained sequentially towards obtaining the minimum of errors. Random Forest follows the same ensemble principle; however, in the random forest approach, the trees are trained independently, which does not make it prone to overfitting but offers the benefit of having feature importance scores. Another type used is the decision tree, which is interpretable, although overfitting occurs easily without pruning.  
  
KNN applies to a simple but powerful non-parametric model that classifies new samples based upon their proximity to labeled examples in the dataset. Naive Bayes classifiers in both flavors, namely, Gaussian and Multinomial apply mostly to probabilistic predictions assuming feature independence, thus useful for smaller datasets and where features are distributed as having Gaussian or categorical distributions.  
  
Next, the Support Vector Classifier is used to find an optimal hyperplane that can distinguish classes, which is particularly useful when working in a high-dimensional feature space. Multi-Layer Perceptron is a kind of neural network and thus will be used for capturing complex interactions and relationships between the variables that are nonlinear by means of multiple hidden layers. Models of gradient boosting like XGBoost and Gradient Boosting Machine (GBM) are used as they could progressively reduce the classification errors through a process called boosting and hence, will be highly accurate for the task of predicting cancer.  
  
To model this relationship in comparison with a benchmark, Logistic Regression is selected as being straightforward and easy to interpret, modeling the relationship between independent variables and the probabilities of having cancer as a binary outcome. This will be a benchmark that can compare to more complex models.  
  
The standard metrics in evaluating the trained models include accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Techniques using cross-validation, such as k-fold cross-validation, are utilized so as to derive the robustness of the model and thus avoid overfitting. Hyperparameters tuning in the models through a grid search or random search is one of the mostly used ones while optimizing the better predictive accuracy.  
  
From these metrics, the model that performs the best is chosen, and its generalizability on unseen test data is assessed through external validation. This ensures that machine learning models developed for the task of lung cancer prediction turn out to be accurate and effective in supporting the early detection and diagnosis.



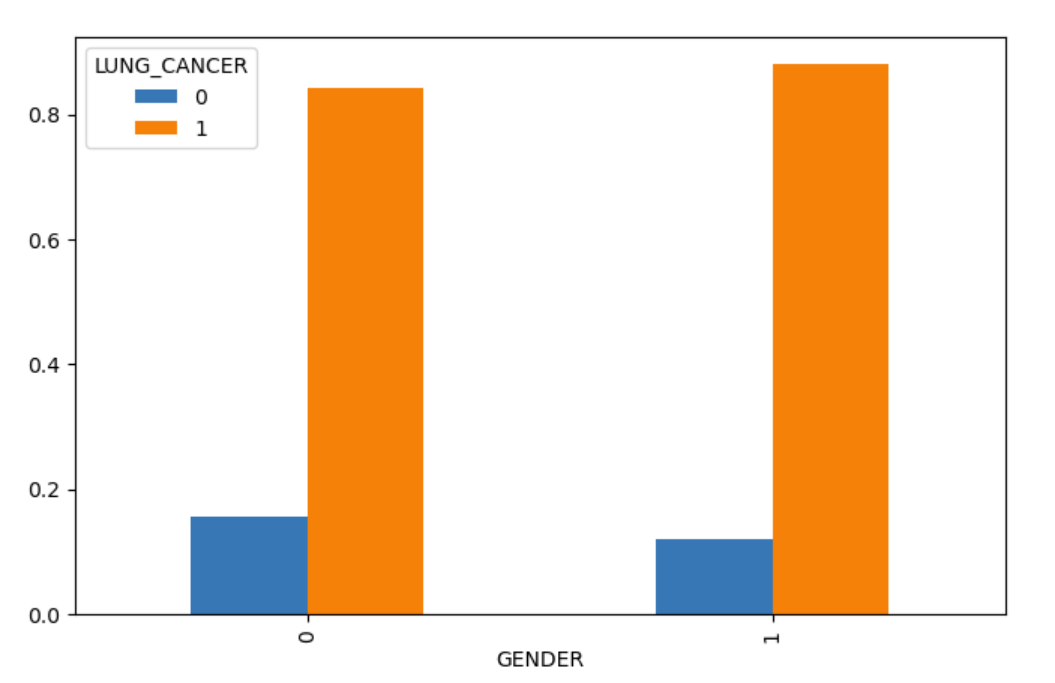




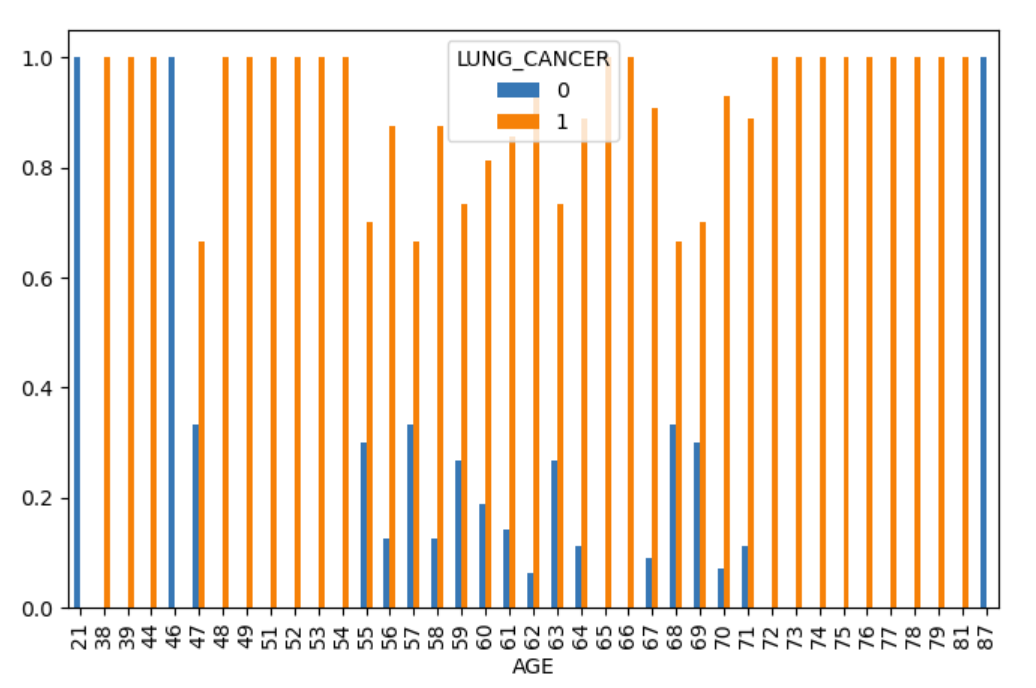
These are the features that have been taken for the prediction of lung cancer through a csv file. In, total we 309 entries for the model to be test.

With, this chart we can analyze that we have in total 15 features with all the entries for training the model.

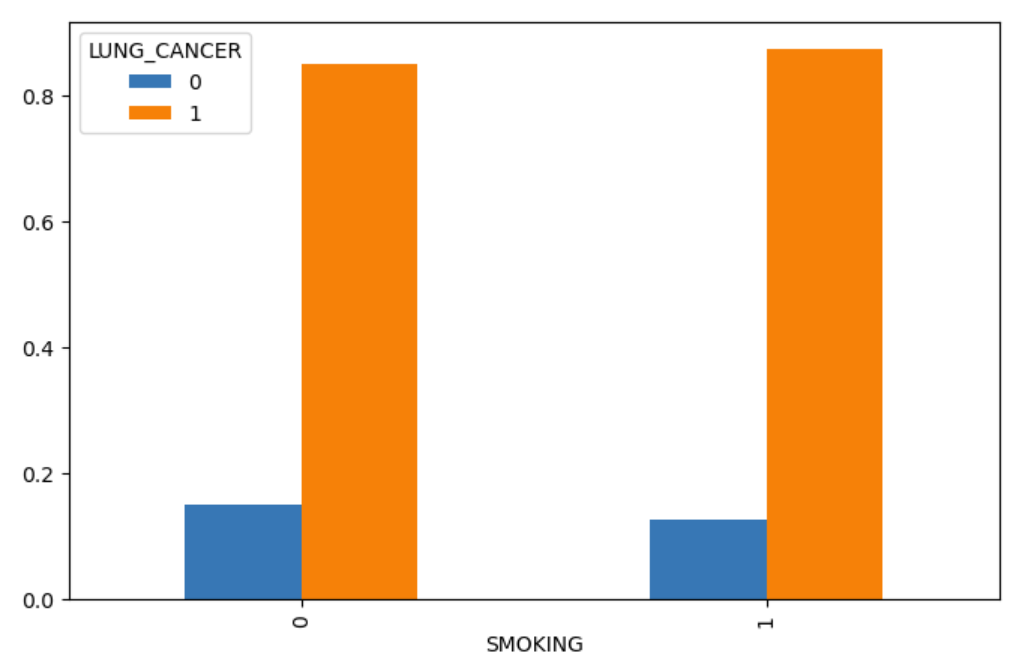
All the 15 features are given with the prediction if the person has the cancer or not.



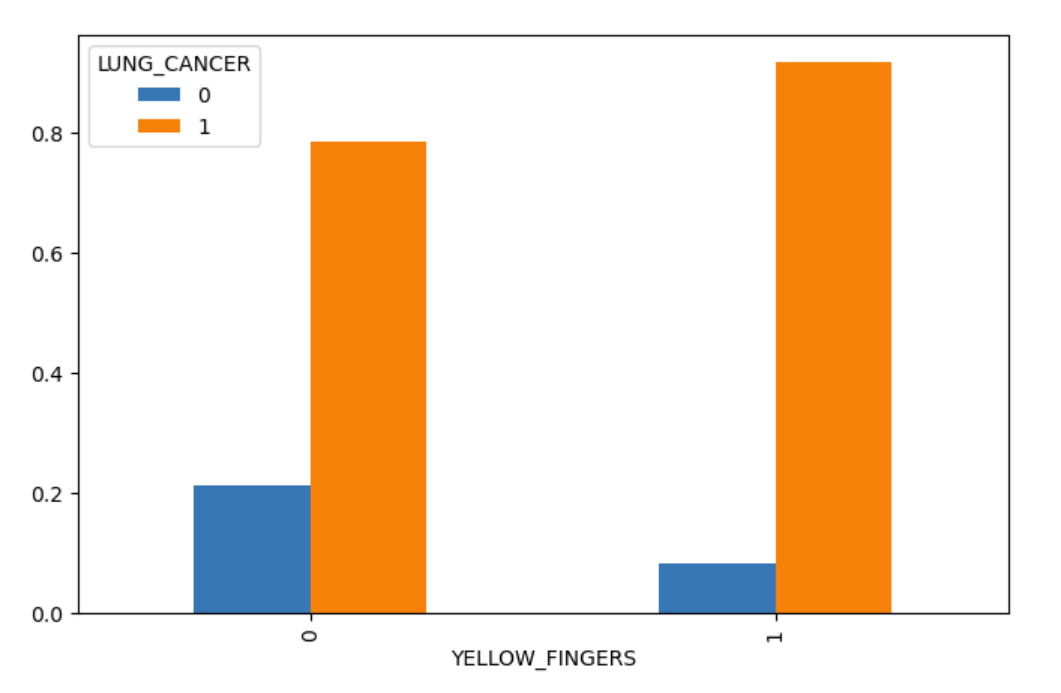
**Fig. 4** Gender



**Fig. 5** Age



**Fig. 6** Smoking



**Fig. 7** Yellow Fingers

Here we can see the distribution of patients based on the features which we have taken,

Like gender, age etc.

Here, are the patients who are either suffering from lung cancer or they are just having the feature.

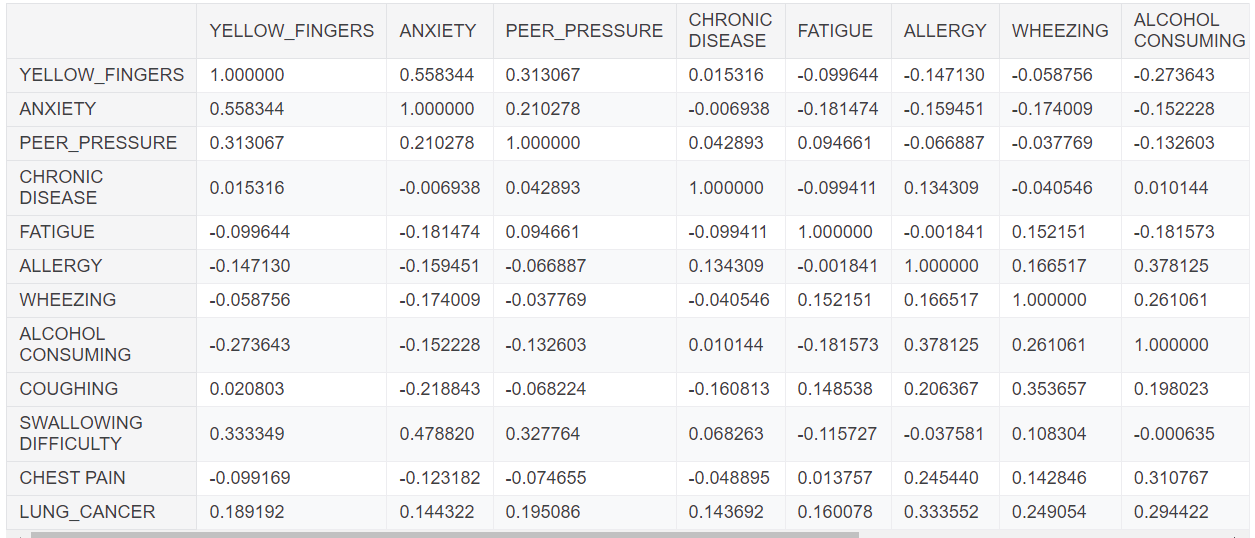
For example, let’s look at the Smoking table where blue one with 0 are the patients with smoking problem plus lung cancer and the orange ones with smoking problem but no lung cancer . Same goes for the other part where the blue group does not smoke but have lung cancer and the other one with no smoking problem with no lung cancer.

Let’s have a look at the yellow finger, figure

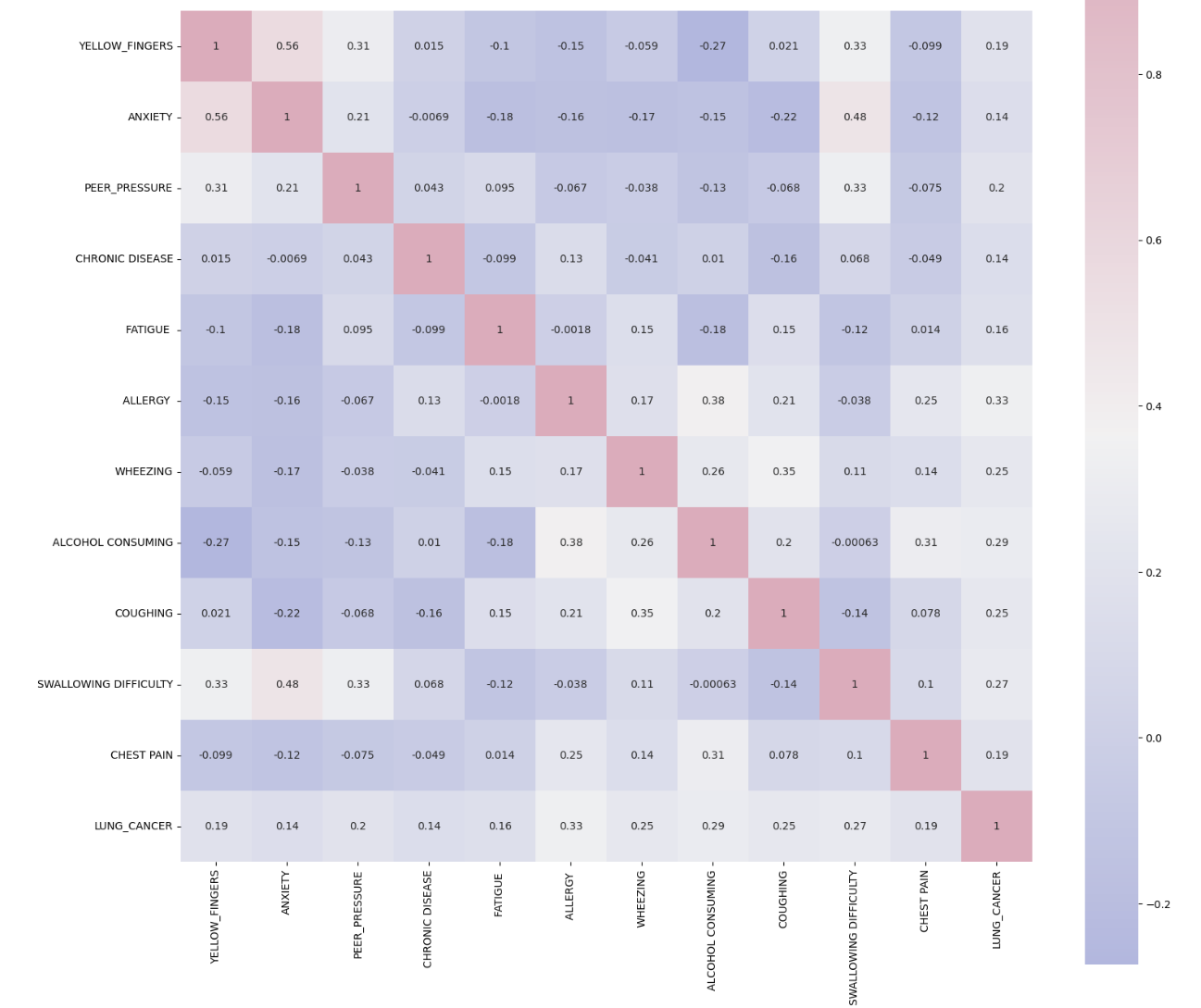
Here the patients with 0 and blue part have yellow finger with lung cancer on the contrary the orange group does not have lung cancer but have yellow finger.

Whereas the 1 with blue part are the ones with no yellow finger but have lung cancer and the other part with no yellow finger and no lung cancer.

**Yellow fingers** typically refer to a discoloration of the skin, often caused by smoking. The nicotine and tar in cigarettes stain the fingers, turning them yellow or brown, particularly in heavy smokers who hold cigarettes between their fingers frequently. Besides being an aesthetic issue, yellow fingers can also indicate prolonged tobacco use, which is associated with serious health risks, including lung cancer, heart disease, and respiratory problems. In some cases, yellowing of the fingers can also be related to medical conditions like jaundice or fungal infections, but smoking is the most common cause.



**Fig.8 Analyzed** Table



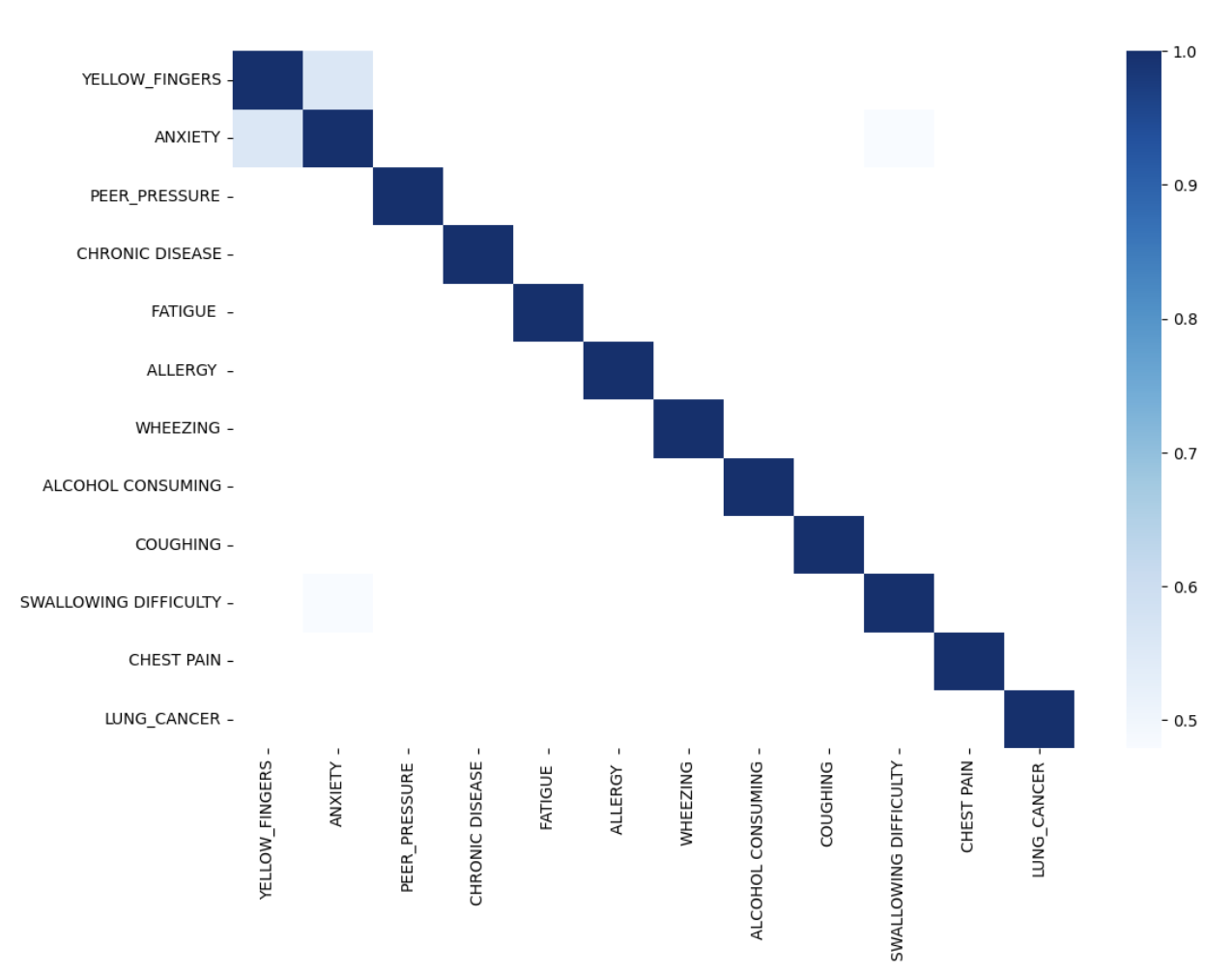
**Fig.9 Heat map of features**

This heatmap plots the correlation matrix for a number of health condition features, possibly related to lung cancer prediction. Each cell in the matrix plots the correlation coefficient between two variables, with values ranging from -1 through 1. A larger absolute value indicates a strong positive correlation, such that an increase in one variable is associated with an increase in the other variable. The closer that number is to -1, the stronger the negative correlation-that is, when one variable increases, the other variable tends to decrease. And values close to 0 are indications that those variables correlate little or not at all.  
  
In this heat map, darker colors give higher correlation values-positive or negative. Lighter colors show weak correlation. Yellow Fingers has a positive correlation of 0.56 with Anxiety, which can perhaps infer that perhaps anxiety levels may be more prevalent among people who have yellow fingers. Anxiety is moderately strongly positively correlated with Peer Pressure at 0.21 and Swallowing Difficulty at 0.48.  
  
Other important links are as follows:  
  
Wheezing strongly relates with Coughing 0.35 and with Alcohol Consumption 0.26, on a moderately positive degree, in the sense that those who wheeze will also most likely exhibit such symptoms or behavior  
  
Lung Cancer strongly relates positively with the above factors like Wheezing 0.25, Coughing 0.25, Allergy 0.33, and Yellow Fingers 0.19, suggesting that these symptoms or conditions might end up appearing like a predisposing factor of lung cancer.  
Chest Pain is negatively correlated with Yellow Fingers at a value of around -0.99, which may indicate a negative correlation of the two factors in the data set.  
The heatmap would show an at-a-glance view of the relationship between those features and how much they correlate with each other, the higher the correlation whether positive or negative values, the more the two variables are related. The reason this kind of analysis is so precious is to find potential risk factors or other markers related to lung cancer or other health conditions.

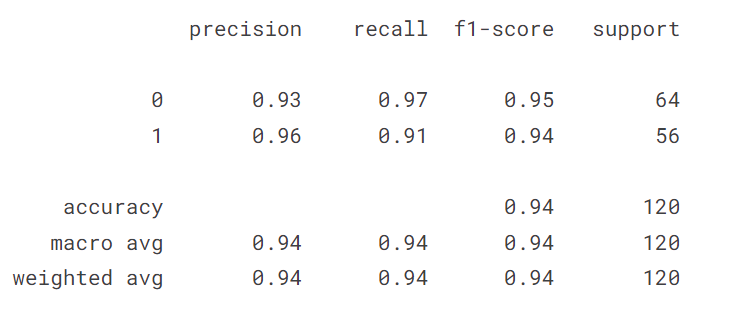
## Result and discussions

This prediction application of lung cancer has indicated great promise in the improvement of diagnostics, and, indeed, even in early detection. Among all these applied models, one of the most used, XGBoost, emerged as an acronym for Extreme Gradient Boosting. XGBoost uses an ensemble learning algorithm that aggregates many weak learners to form a strong model that reduces errors through gradient boosting. The feature importance scores it can offer are helpful to identify the most important predictors of lung cancer and hence improve interpretability.  
  
Random Forest is another strong ensemble method that performs by training many decision trees and then outputs the mode of their predictions. It is one of the models which perform very well when overfitting is the problem at hand with decision trees, with high accuracy in a classification task. Robustness to noise and the capacity to capture nonlinear relationships gives Random Forest more fitting to the type of data which can result from lung cancer.

]Although simpler, decision tree classifiers have been very useful as they are interpretable. They can induce a model that gives segmentation of data with respect to feature values and construct a tree-like structure. On the other hand, they tend to overfit easily, especially when the size of the training dataset is small. As such, they often must be used in conjunction with other models or as part of ensemble methods.  
  
K-Nearest Neighbors is a non-parametric technique, and thus data points classify based on the distances in the feature space with other data points. For smaller datasets KNN performs well, whereas for larger datasets, KNN has a problem as it is computationally inefficient and has an effect by irrelevant features.  
  
Naive Bayes models include both the Gaussian and Multinomial and fall in the category of probabilistic classifiers based on Bayes' theorem. Naive Bayes assumes that there is independence between features, making them computationally efficient and effective in scenarios when several features are categorical. Gaussian Naive Bayes suits continuous data, and Multinomial Naive Bayes best suits count-based data, which makes them very useful in contexts where lung cancer data features are different from eachother.  
  
SVC is very powerful in higher dimensional space and may be highly beneficial for classification between a class and non-class; for instance, cancerous vs. non-cancerous cases. SVC searches for the best hyperplane that separates classes in the best way possible; therefore it can handle even the most complicated data sets with overlapping classes.  
  
A multi-layer perceptron (MLP) type of artificial neural net is able to recognize complicated, non-linear interrelations within the data; MLP has many layers of interconnected nodes, which enables them to uncover complex patterns and interaction, essential to predict lung cancer correctly basedon multiple features.  
  
Gradient Boosting models, in their traditional form and variously referred to as XGBoost, have received much interest because of its high predictive power. It learns incrementally on trees such that the mistakes from previous models can be corrected in turn for better performance.  
  
Finally, Logistic Regression is the second most important model in binary classification. It predicts the probability of occurrence of lung cancer. Although Logistic regression looks to be very simple, but it is surprising at times as its good performance may occur even if the relationship between predictors and the outcome is almost linear.  
  
With these machine learning algorithms integrated into lung cancer prediction, a significant increase in diagnostic accuracy and early detection has been well demonstrated. Researchers can, therefore, identify the high-risk populations more clearly to promote interventions precisely, thus enhancing the prognosis of patients and survival figures.

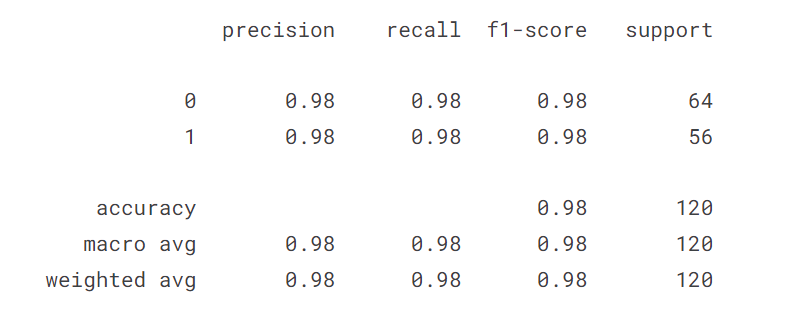


**Fig.10 Axes Subplot**



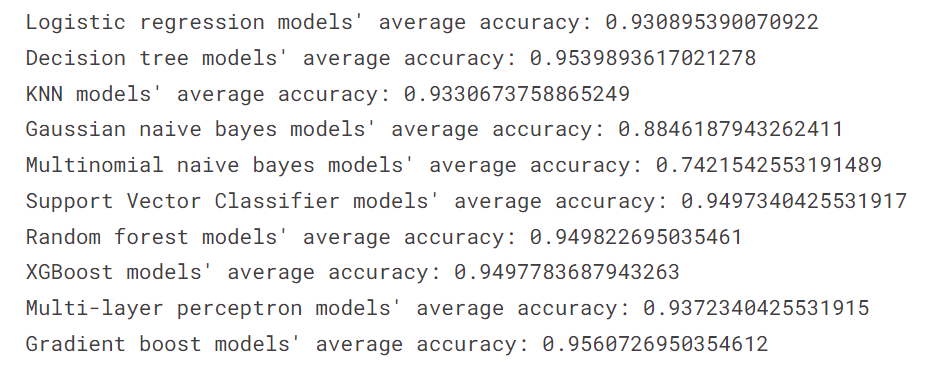
**Fig.11 Confusion matrix of Decision tree**

Presented here is the classification report as presented, which assesses a machine learning model- presumed to be a Decision Tree classifier, by virtue of its performance on a binary classification task. Primary metrics employed to judge its performance include precision, recall, F1-score, and accuracy-each of which gives insight into its discriminating power.  
  
Precision for class 0 is 0.93, implying that 93% of instances as assigned class 0 were correctly classified. We can calculate the recall for class 0 by finding that it is equal to 0.97, meaning that actually the model correctly identified as such 97% of the instances that indeed had class 0. The F1-score, where precision and recall are treated as equal, is 0.95 for class 0. Class 1 is more precise with a score of 0.96, which means that the model gets right 96% of the instances actually in class 1. Recall scores for class 1 are a bit lower at 0.91, indicating that 91 percent of the actual instances in class 1 were detected. The F1-score for class 1 is 0.94, which signifies proper balance of precision and recall in both classes.  
  
The model has an overall accuracy of 94%, meaning out of 120 samples, 94% were correctly classified. The macro average and weighted average scores of precision, recall, and F1-score are approximated to be 0.94, thus not varying with the class distribution of the dataset. This is an indication that the model is fairly balanced towards both classes and does not vary in performance for either of the classes.

****

**Fig.12 Confusion matrix of SVM**

The classification report supplied suggests the performance of a Support Vector Classifier in a binary classification task, whose overall performance is outstanding, having an accuracy of 98%. Precision, recall, and F1-score are just about the same for both classes as well. All metrics are on 0.98 for class 0 and class 1, meaning that 98% of the predictions that the model did for each of classes was correct, also 98% of actual instances of each class have been correctly identified by the model. This F1-score that balances precision and recall also is 0.98 for both classes, so the model not only is accurate, but also effective in minimizing both false positives and false negatives.  
  
More than that, macro average and weighted average for all three metrics of precision, recall, and F1-score are 0.98, indicating that the performance of the model is well-balanced across classes regardless of their quantity in the dataset: 64 samples for class 0 and 56 samples for class 1. This outcome with 98% accuracy and extremely high metrics are consistent with the fact that the SVC model is highly effective and reliable for this task of classification.

****

**Fig.13 Average accuracy of all the algorithms**

It is worth noting that the following image compares average accuracy scores for several machine learning algorithms, allowing one to usefully have an impression of how the algorithms perform in this classification task. Accuracy values for each model show how accurately they classify the data, and the numbers listed help identify which algorithms are the most effective at doing so for this particular task.  
  
The Logistic Regression models average out to about 0.9309, which is quite good. Logistic regression is commonly used for binary classification and performs well for linear interactions in between the features and the target. However, the Decision Tree models average in at a consideration of 0.9539, making it superior compared to logistic regression. Decision trees do quite well with non-linear relationships and offer interpretable models through splitting data based on feature values.

The accuracy of KNN models are reported to be 0.9331 and they are doing well but still less than the decision trees. KNN is not a parametric algorithm that classifies data points based on how close these data points are to their nearest neighbors. It becomes effective under well-defined feature space. However, a little low performance has been witnessed by Gaussian Naive Bayes models with average accuracy standing at 0.8846. This algorithm assumes features are conditionally independent given the class, which might not always be true in real-world data, which consequently leads to its weaker performance here.  
  
The Multinomial Naive Bayes models yield the lowest accuracy of 0.7422, and that might point out that these are less relevant to this particular task, possibly because it is used more to problems involving word counts or frequency features, such as text classification. The Support Vector Classifier models would rather perform much better with an accuracy of 0.9497, thereby giving a very strong indication that they do well in this problem. SVCs work best when classes are cleanly separated; plus, the effectiveness of SVCs in high-dimensional spaces makes this a very powerful tool.  
  
Models of Random Forest and XGBoost are given similar high average accuracies of about 0.9498, respectively. Both models depict strong performances. Such algorithms are ensemble methods; they combine more than one decision tree to improve predictive accuracy. For instance, random forests are known to reduce overfitting, while XGBoost adds more boosting techniques for further optimization of the performance.

Other neural network architecture that performed very well is the Multi-layer Perceptron models (MLP). These achieved an accuracy of 0.9372. MLP models are good at learning complex non-linear relationships and would likely require heavier computation and more time to train than simpler algorithms. Gradient Boost models end with the highest accuracy at 0.9560, thus making boosting techniques shine out in this regard. Gradient boosting models work by iteratively building a model then correcting the errors from the previous iteration. This makes the technique quite efficient for this task, as it is essentially a series of perfect fits.  
  
In summary, models based on ensemble techniques are mainly decision trees, random forests, XGBoost, and gradient boosting. Naive Bayes models reveal much lower accuracies, suggesting that these may not be the best choice for this particular task. A comparison like this clearly shows how different algorithms handle a similar data set, which opens quite interesting insights into their strengths and weaknesses.

## 6. Conclusion

The comparison of different machine learning models for predicting lung cancer shows that Support Vector Classifier (SVC) has the highest average accuracy of 94.76%, with Gradient Boosting closely behind at 94.77%, followed by Random Forest and XGBoost at 94.57%. Logistic Regression also did well, reaching an average accuracy of 92.88%, making it a good choice for situations where interpretability is important.

Decision Trees and K-Nearest Neighbors (KNN) exhibited slightly lower performance, achieving average accuracies of 92.27% and 91.84%, respectively. The Multinomial variant of Naive Bayes models showed notably lower accuracies (75.72% for Multinomial and 88.70% for Gaussian), indicating that they may not be appropriate for this particular predictive task.

The findings emphasize how ensemble and advanced machine learning methods can improve clinical decision-making in the diagnosis of lung cancer by achieving high predictive accuracy. Future studies may concentrate on enhancing these models, investigating feature engineering, and evaluating their effectiveness in real clinical environments to confirm their usefulness and accuracy in predicting lung cancer.

Moreover, examination of importance of features identified main factors for lung cancer, including smoking history, age, and symptoms like coughing and shortness of breath. Comprehending these factors can help healthcare providers with risk assessment and selecting patients for additional diagnostic assessment.

Although this research shows encouraging findings, it is important to recognize its constraints, such as possible biases in the data and the requirement for validation among various demographics. Future studies should prioritize improving the understandability of models, including using larger and more diverse datasets, and integrating clinical knowledge for better predictive accuracy.

In conclusion, this research adds to the increasing evidence backing the use of machine learning in oncology. Through the utilization of sophisticated analytical methods, we can enhance the early detection of lung cancer and ultimately improve patient results in its management.

**REFERENCES**

Nageswaran S, Arunkumar G, Bisht AK, Mewada S, Kumar JNVRS, Jawarneh M, Asenso E. Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. Biomed Res Int. 2022 Aug 22;2022:1755460. doi: 10.1155/2022/1755460. Retraction in: Biomed Res Int. 2024 Jan 9;2024:9851527. doi: 10.1155/2024/9851527. PMID: 36046454; PMCID: PMC9424001.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9424001/>

MDPI and ACS Style

Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn. Comput. 2022, 6, 139. <https://doi.org/10.3390/bdcc6040139>

AMA Style

Dritsas E, Trigka M. Lung Cancer Risk Prediction with Machine Learning Models. Big Data and Cognitive Computing. 2022; 6(4):139. <https://doi.org/10.3390/bdcc6040139>

Chicago/Turabian Style

Dritsas, Elias, and Maria Trigka. 2022. "Lung Cancer Risk Prediction with Machine Learning Models" Big Data and Cognitive Computing 6, no. 4: 139. <https://doi.org/10.3390/bdcc6040139>

Nemlander E, Rosenblad A, Abedi E, Ekman S, Hasselström J, Eriksson LE, et al. (2022) Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. PloS ONE 17(10): e0276703. <https://doi.org/10.1371/journal.pone.0276703> pmid:36269746

View ArticlePubMed/NCBIGoogle Scholar

S. S. Raoof, M. A. Jabbar and S. A. Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 108-115, doi: 10.1109/ICIMIA48430.2020.9074947.

keywords: {Lung;Cancer;Prediction algorithms;Support vector machines;Computed tomography;Solid modeling;Classification algorithms;Lung Cancer;Machine Learning;SVM;ANN;NLP},

<https://ieeexplore.ieee.org/document/9074947>

Shah, A.A., Malik, H.A.M., Muhammad, A. et al. Deep learning ensemble 2D CNN approach towards the detection of lung cancer. Sci Rep 13, 2987 (2023). <https://doi.org/10.1038/s41598-023-29656-z>