

# **Lung Cancer Prediction**

## **Mini Project Report**

Submitted by

**Neema Poulse**

*Submitted in partial fulfillment of the requirements for the award of  
the degree of*

***Master of Computer Applications  
Of***

***A P J Abdul Kalam Technological University***



**FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®**

**ANGAMALY-683577, ERNAKULAM(DIST)**

**DECEMBER 2021**

## **DECLARATION**

I, **Neema Poulse**, hereby declare that the report of this project work, submitted to the Department of Computer Applications, Federal Institute of Science and Technology (**FISAT**), Angamaly in partial fulfillment of the award of the degree of Master of Computer Application is an authentic record of our original work.

The report has not been submitted for the award of any degree of this university or any other university.

**Date : 04-03-2022**

**Place: Angamaly**

**FEDERAL INSTITUTE OF SCIENCE AND  
TECHNOLOGY (FISAT)®  
ANGAMALY, ERNAKULAM-683577**

**DEPARTMENT OF COMPUTER APPLICATIONS**



**CERTIFICATE**

This is to certify that the project report titled "**Lung Cancer Prediction**" submitted by **Neema Poulse** towards partial fulfillment of the requirements for the award of the degree of Master of Computer Applications is a record of bonafide work carried out by them during the year 2022.

**Project Guide**

**Head of the Department**

Submitted for the viva-voice held on ..... at .....

**Examiner1 :**

**Examiner2 :**

## **ACKNOWLEDGEMENT**

We are extremely glad to present our main project which we did as a part of our curriculum. We take this opportunity to express our sincere thanks to those who helped us in bringing out the report of my project.

We are deeply grateful to Dr. George Manoj George, Principal, FISAT, Angamaly and Dr. C. Sheela, Vice Principal, FISAT, Angamaly. Our sincere thanks to Dr. Deepa Mary Mathews, Head of the department of MCA, FISAT, who had been a source of inspiration. During the period of our project work, we have received generous help from Miss Anju Lekha, my project guide, which I like to put on record here with deep gratitude and great pleasure.

I express our heartfelt thanks to all the faculty members in our department for their constant encouragement and never ending support throughout the project.

Finally I am grateful to all our friends who gave us a lot of suggestions for the successful completion of this project.

## **ABSTRACT**

The prominent cause of cancer-related mortality throughout the globe is "Lung Cancer". Hence beforehand detection, prediction and diagnosis of lung cancer has become essential as it expedites and simplifies the consequent clinical board. To erect the progress and medication of cancerous conditions machine learning techniques have been utilized because of its accurate outcomes. Various types of machine learning algorithms(ML) like Support Vector Machine (S V M) , Logistic regression, Artificial Neural Network(ANN), have been applied in the healthcare sector for analysis and prediction of lung cancer.This research uses data mining technology such as classification, clustering and prediction to identify potential cancer patients. The gathered data is preprocessed, fed into the database and classified to yield significant patterns using decision tree algorithm. Then the data is clustered using Kmeans clustering algorithm to separate cancer and non cancer patient data. Further the cancer cluster is subdivided into six clusters. Support Vector Machine (SVM) as classifier, has accuracy of 99.2

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>4</b>
<b>2</b>	<b>PROOF OF CONCEPT</b>	<b>6</b>
<b>3</b>	<b>IMPLEMENTATION</b>	<b>8</b>
<b>4</b>	<b>RESULT ANALYSIS</b>	<b>15</b>
<b>5</b>	<b>CONCLUSION AND FUTURE SCOPE</b>	<b>17</b>
<b>6</b>	<b>APPENDIX</b>	<b>19</b>
<b>7</b>	<b>REFERENCES</b>	<b>26</b>

# Chapter 1

## INTRODUCTION

Cancer is very dangerous and common disease that causes death worldwide. Early diagnosis of cancer provide more possibility of getting cured. Cancer disease generates abnormal growth of cells which spreads to all parts of body. In this project I discuss, the early prediction of lung cancer with help of data mining techniques. Lung are spongy organs that affected by cancer cells that leads to loss of life. The common reasons of lung cancer are smoking habits, working in smoke environment or breathing of industrial pollutions, air pollutions and genetic. In this project I have proposed a genetic algorithm based dataset classification for prediction of multiple models. The usage of genetic algorithm (GA) have shown better performance when compared with Particle swarm optimization and differential evolutions.

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected

to learn a way of classifying unseen examples. In association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity. In this study, to classify the data and to mine frequent patterns in data set Decision Tree algorithm is used. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The top most node is the root node. The attribute value of the data is tested against a decision tree. A path is traced from root to leaf node, which holds the class prediction for that data. Decision trees can be easily converted into classification rules. This decision tree is used to generate frequent patterns in the dataset. The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific cancer types and are helpful in predicting the cancer and its type is known as Significant frequent pattern. Using this significant patterns generated by decision tree the data set is clustered accordingly and risk scores are given.



## **Chapter 2**

# **PROOF OF CONCEPT**

### **Objectives**

Lung cancer is the principal cause for cancer-related death. Overusage of tobacco, cigarettes and beedis, are the major risk factor that leads to lung cancer in Indian men; however, among Indian women, smoking is not so common, which indicate that there are other factors which lead to lung cancer. Other risk factors include exposure to radon gas, air-pollutions and chemicals in the workplace. A cancer that starts in lung is primary lung cancer whereas those which starts in lung and spread to other parts of body is secondary lung cancer. Size of tumour and how far it has spread determines the stage of cancer. An early stage cancer is a small cancer that is diagnosed in lung and advanced cancer is the one that has spread into surrounding tissue or other part of body . A better understanding of risk factors can help to prevent lung cancer disease.

The Lung Cancer datasets used for this study are taken from kaggle. First, the given datasets are divided into training and test data by using k-fold cross validation technique. Then using the classification algorithms such as SVM ,Logistic Regression, Naïve Bayes and Decision Tree, respective classification models are

implemented using the given training data. The classification models are created using training data and the corresponding models are evaluated using test data to get the accuracy of the models. Finally, we compared the accuracy rates of each and every classification models that we implemented and arrived at a conclusion.

## Chapter 3

# IMPLEMENTATION

The dataset we took is a news dataset from Kaggle.com. I have isolated the label column from the data frame. Preprocessing steps are applied and then we split the dataset into train and test data set. Feature are extracted using tfidf vectorizer.

From sklearn, I imported svm model. Linear kernel is used in creating SVM classifier and response was predicted for test data set. Accuracy of SVM classifier is 99.2

### TOOLS OR PROGRAMMING LANGUAGE

- FRONT END :

- Html
- CSS
- flask

- BACK END :

- Python

## **MODULES**

### **1. Data Preprocessing :**

Data preprocessing is an iterative process for the transformation of the raw data into understandable and useable forms. Raw datasets are usually characterized by incompleteness, inconsistencies, lacking in behavior, and trends while containing errors. The preprocessing is essential to handle the missing values and address inconsistencies

### **2. Feature Extraction :**

Feature extraction involves reducing the number of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

### **3. Training the Model :**

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model

### **4. Evaluation :**

Model evaluation techniques in machine learning are helping us to find a better model among all other models in machine learning. It is simply the selection of machine learning models or measuring the performance of machine learning models.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	GENDER	AGE	SMOKING	YELLOW_F	ANXIETY	PEER_PRE	CHRONIC	FATIGUE	ALLERGY	WHEEZING	ALCOHOL	COUGHING	SHORTNE	SWALLOW	CHEST PAI	LUNG_CANCER	
2	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES	
3	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES	
4	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO	
5	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO	
6	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO	
7	F	75	1	2	1	1	2	2	2	2	1	2	2	1	1	YES	
8	M	52	2	1	1	1	1	2	1	2	2	2	2	1	2	YES	
9	F	51	2	2	2	2	1	2	2	1	1	1	2	2	1	YES	
10	F	68	2	1	2	1	1	2	1	1	1	1	1	1	1	NO	
11	M	53	2	2	2	2	2	1	2	1	2	1	1	2	2	YES	
12	F	61	2	2	2	2	2	2	1	2	1	2	2	2	1	YES	
13	M	72	1	1	1	1	2	2	2	2	2	2	2	1	2	YES	
14	F	60	2	1	1	1	1	2	1	1	1	1	2	1	1	NO	
15	M	58	2	1	1	1	1	2	2	2	2	2	2	1	2	YES	
16	M	69	2	1	1	1	1	1	2	2	2	2	1	1	2	NO	
17	F	48	1	2	2	2	2	2	2	2	1	2	2	2	1	YES	
18	M	75	2	1	1	1	2	1	2	2	2	2	2	1	2	YES	
19	M	57	2	2	2	2	2	1	1	1	2	1	1	2	2	YES	
20	F	68	2	2	2	2	2	2	1	1	1	2	2	1	1	YES	
21	F	61	1	1	1	1	2	2	1	1	1	1	2	1	1	NO	
22	F	44	2	2	2	2	2	2	1	1	1	1	2	2	1	YES	
23	F	64	1	2	2	2	1	1	2	2	1	2	1	2	1	YES	
24	F	71	2	1	1	1	2	2	2	1	1	1	2	1	1	NO	

**DATASET**

The dataset used to test the efficiency of the model is produced by kaggle, containing size of 310.

- Attributes of dataset :
  - Gender
  - Age
  - Smoking
  - yellow fingers
  - Anxiety
  - Allergy
  - Wheezing
  - Alcohol
  - Coughing
  - Chest Pain
  - Peer pressure

## ALGORITHMS TO BE USED

### 1 Naïve Bayes :

Naïve Bayes is a type of classifier and it's a supervised learning algorithm. The prediction occurs on the basis of probability. It makes quick predictions for the machine learning models. It works best with text classification. Naïve Bayes is used for calculating conditional probability. It is derived from Bayes theorem. It states that “probability that something will happen, given that something else has already occurred” (Saxena, 2017). In Naïve Bayes, occurrence of one feature is independent of other feature. Naïve Bayes is a type of classifier and it's a supervised learning algorithm. The prediction occurs on the basis of probability. It makes quick predictions for the machine learning models. It works best with text classification.

Naïve Bayes Classifier is used for multiclass and binary classifications. But the disadvantage of using it is, classifier fails to learn the relationship between features as it treats all features independent of each other. There are three types of naïve bayes model i.e. Gaussian, Multinomial, Binomial. Gaussian model follows normal distribution. Multinomial model is used mostly for text classification problems and prediction is on the basis of frequency of words. Binomial classifier is similar to multinomial, it is used for classification task. In our model, we have used Multinomial naïve bayes classifier.

$$P(A|B) = P(B|A)P(A) / P(B) \quad (1)$$

$P(A|B)$ : Probability of event A such that event B has already occurred.

## **2. Support Vector Machine (SVM) :**

A support vector machine (SVM), is a managed learning calculation. Hence, the model is constructed after it has already been trained. The main motive of SVM is to categorize new data that comes under. There is a decision boundary or hyperplane that splits dataset into two class. For the considered class, a point is chosen such that it is close to the opponent class. A line is drawn touching the point parallel to hyperplane. Hyperplane is drawn considering maximum margin. SVM are more accurate on smaller dataset. The disadvantage of using SVM on large dataset is training time is high.

There are two types of SVM model. Linear SVM is used for linearly separable data. When Single straight line classify two classes of a dataset, such data is called as linearly separable data and the classifier used for this type of data is Linear SVM.



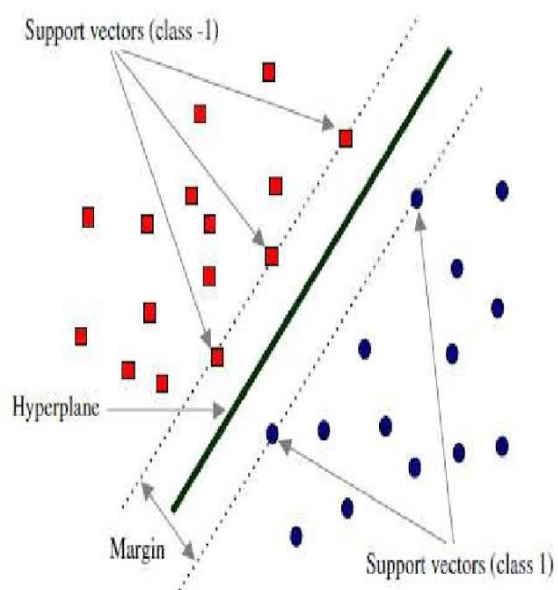


Figure 3.1: Classification of two different categories using hyperplane.

## Chapter 4

# RESULT ANALYSIS

Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used: In most cases, high accuracy value represents a good model, but considering the fact that we are training a classification model in our case, an article that was predicted as true while it was actually false (false positive) can have negative consequences; similarly, if an article was predicted as false while it contained factual data, this can create trust issues.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

**Confusion Matrix**

A Confusion matrix is an  $N \times N$  matrix used for evaluating the performance of a classification model, where  $N$  is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

For a binary classification problem, we would have a  $2 \times 2$  matrix as shown below with 4 values:

1. TP = True Positives
2. FP = False Positives
3. TN = True Negatives
4. FN = False Negatives

## **Chapter 5**

# **CONCLUSION AND FUTURE SCOPE**

### **Conclusion**

In earlier times, the doctor has to do multiple tests in order to detect whether a given patient has lung cancer or not . But this was a very time consuming process. In a diagnosis sometimes a patient has to undergo unnecessary check-ups or different tests to identify the disease of lung cancer. To minimize the process time and unnecessary check-ups there needs to be a preliminary test in which both the patient and the doctor will be notified with the possibilities of lung cancer.

Nowadays the machine learning algorithms plays an important role in the prediction and classification of medical data. Logistic Regression, SVM, decision tree and Naïve Bayes are the machine learning algorithms used for this comparative study. A comparative analysis of accuracy rates of each classifier are presented. The predictive performance of classifiers are compared quantitatively. In the performance chart, different results are produced for each classifier on the lung cancer dataset. Looking at the correct classification (CA) and other metrics; the best result is given by the support vector machine algorithm. SVM algorithm used

high dimension to classify the observation so it's performance is the best. More accurate lung cancer detection can be done using this technique. Therefore, there is less mistakes. Finally, by adding extra pre-processing the accuracy rate can be enhanced.

### **Future Scope**

In the emerging nations, there is an intense growth in the occurrence of smoking. This is estimated to stimulate the frequency of lung distortions in the upcoming period. The growing occurrence of cancer is directly proportional to the development of the market because it drives the demand for timely screening and identification of cancer. The lung cancer diagnostics market on the source of Type of Test could span Molecular Test, Biopsy, Sputum Cytology, Imaging Test, and Others. The subdivision of the molecular tests is estimated to witness an important development above the prediction period by the way of a CAGR of more than 10% The Lung Cancer Diagnostic market on the source of Type could span Non-small Cell Lung Cancer, Small Cell Lung Cancer. The source of distinction of the tumor categories is the dimensions of the tumor cells. Small Cell Lung Cancer (SCLC) extends rapidly in the body. It marks discovery of this cancer, difficult, at an initial phase.

# **Chapter 6**

## **APPENDIX**

**Sourcecode ( page no : 20 - 21)**

- Svm.py ( page no : 21)
- frond end : index.php (page no : 20)

**Dataset ( page no : 25)**

**Screenshots ( page no : 22 - 24)**

```
main.py 1 X
C:\Users\HP>HP>AppData>Local>Temp>Rar$DI153544> main.py > ...

70 @app.route('/lungresult', methods= ["POST"])
71 def lungResult():
72     title = 'Prediction'
73     data = []
74     try:
75         q = ''
76         q = (request.form.get('gender'))
77         if q=='F':
78             data.append(0)
79         elif q=='M':
80             data.append(1)
81         else:
82             return render_template('invalid.html',title=title,output='Invalid Gender')
83     data.append(int(request.form.get('age')))
84     data.append(int(request.form.get('smoking')))
85     data.append(int(request.form.get('yellow_fingers')))
86     data.append(int(request.form.get('anxiety')))
87     data.append(int(request.form.get('peer_pressure')))
88     data.append(int(request.form.get('chronic_disease')))
89     data.append(int(request.form.get('fatigue')))
90     data.append(int(request.form.get('allergy')))
91     data.append(int(request.form.get('wheezing')))
92     data.append(int(request.form.get('alcohol_consuming')))
93     data.append(int(request.form.get('coughing')))
94     data.append(int(request.form.get('shortness_of_breath')))
95     data.append(int(request.form.get('swallowing_difficulty')))
96     data.append(int(request.form.get('chest_pain')))
97     print(data)
98     output = PredictLC(data)
99     return render_template("lungPredict.html", title=title,output = output ) #render template se hum puri html file browser ko bhej dete hai,
100 except:
```

The screenshot shows a Jupyter Notebook environment. On the left is a file explorer with a search bar and icons for file operations. It displays a directory structure with a folder named 'sample\_data' containing a file 'survey\_lung\_cancer.csv'. Below the file explorer is a disk usage indicator showing '65.88 GB available'.

The main area contains three code cells, each with a green checkmark and a '0s' execution time:

- Cell 1: `import pandas as pd`
- Cell 2: `data = pd.read_csv('survey_lung_cancer.csv')`
- Cell 3: `data.head()`

Below the code cells, a preview of the first five rows of the 'survey\_lung\_cancer.csv' file is displayed as a table:

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALC CONSU
0	M	69	1	2	2	1	1	2	1	2	
1	M	74	2	1	1	1	2	2	2	1	
2	F	59	1	1	1	2	1	2	1	2	
3	M	63	2	2	2	1	1	1	1	1	
4	F	63	1	2	1	1	1	1	1	2	

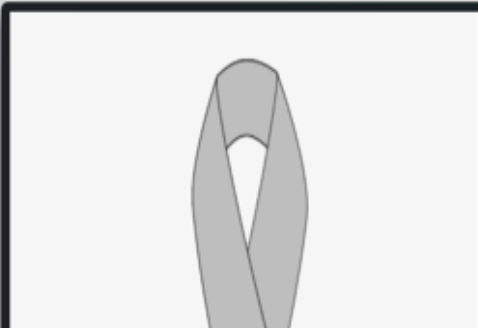


← → 127.0.0.1:5000/lungcancer

## Lung Cancer Detection

Home

Predict whether or not you are susceptible to Lung Cancer



Enter Cell Attributes

Gender


Age

Smoking

Yellow Fingers

Anxiety

Peer Pressure



Chronic Disease

Fatigue

Allergy

Wheezing

Alcohol Consuming

Coughing

Shortness of Breath

Swallowing Difficulty

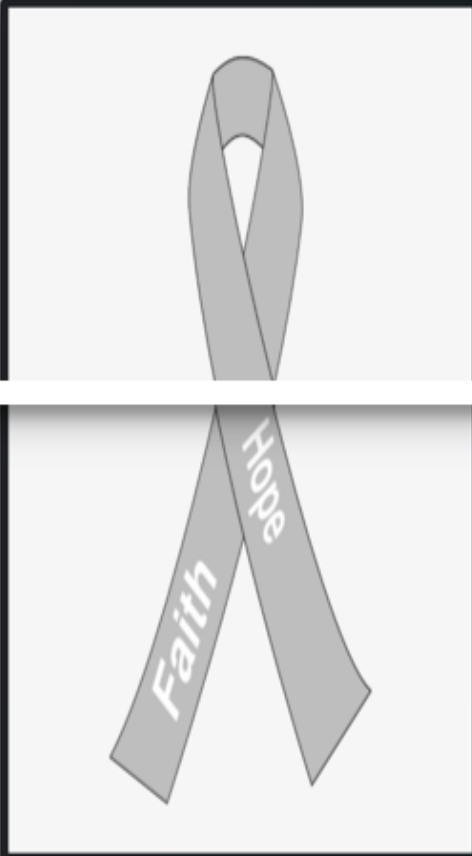
Chest Pain

Predict Cancer

Lung Cancer Detection

Home

Predict whether or not you are susceptible to Lung Cancer



Enter Cell Attributes

M

69

1

2

1

1

1

2

1

2

1

2

2

2

2

2

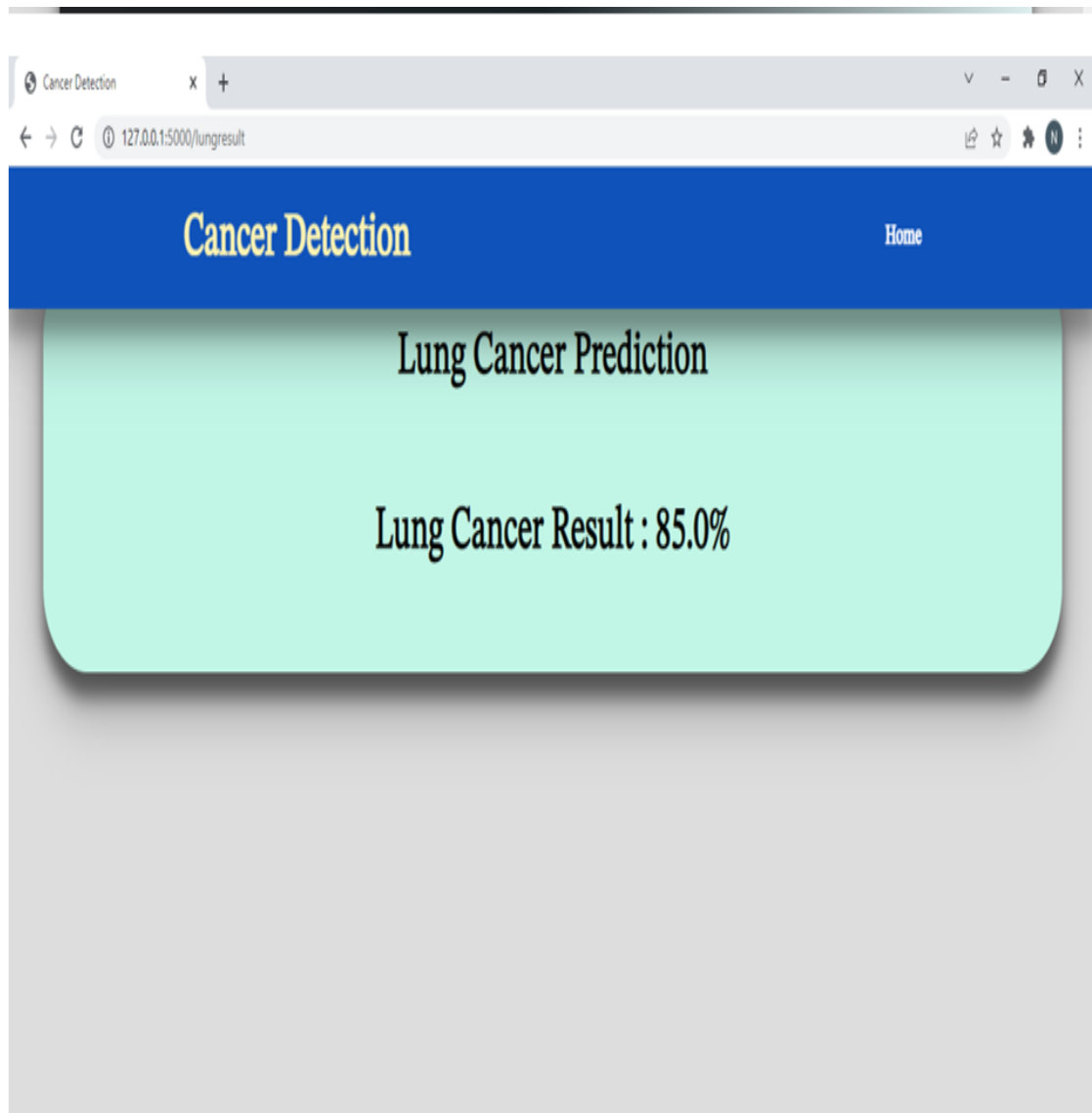
2

2

2

2

Predict Cancer



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	GENDER	AGE	SMOKING	YELLOW_F	ANXIETY	PEER_PRE	CHRONIC	FATIGUE	ALLERGY	WHEEZING	ALCOHOL	COUGHING	SHORTNE	SWALLOW	CHEST PAI	LUNG_CANCER	
2	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES	
3	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES	
4	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO	
5	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO	
6	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO	
7	F	75	1	2	1	1	2	2	2	2	1	2	2	1	1	YES	
8	M	52	2	1	1	1	1	2	1	2	2	2	2	1	2	YES	
9	F	51	2	2	2	2	1	2	2	1	1	1	2	2	1	YES	
10	F	68	2	1	2	1	1	2	1	1	1	1	1	1	1	NO	
11	M	53	2	2	2	2	2	1	2	1	2	1	1	2	2	YES	
12	F	61	2	2	2	2	2	2	1	2	1	2	2	2	1	YES	
13	M	72	1	1	1	1	2	2	2	2	2	2	2	1	2	YES	
14	F	60	2	1	1	1	1	2	1	1	1	1	2	1	1	NO	
15	M	58	2	1	1	1	1	2	2	2	2	2	2	1	2	YES	
16	M	69	2	1	1	1	1	1	2	2	2	2	1	1	2	NO	
17	F	48	1	2	2	2	2	2	2	2	1	2	2	2	1	YES	
18	M	75	2	1	1	1	2	1	2	2	2	2	2	1	2	YES	
19	M	57	2	2	2	2	2	1	1	1	2	1	1	2	2	YES	
20	F	68	2	2	2	2	2	2	1	1	1	2	2	1	1	YES	
21	F	61	1	1	1	1	2	2	1	1	1	1	2	1	1	NO	
22	F	44	2	2	2	2	2	2	1	1	1	1	2	2	1	YES	
23	F	64	1	2	2	2	1	1	2	2	1	2	1	2	1	YES	
24	F	71	2	1	1	1	2	2	2	1	1	1	2	1	1	NO	

# Chapter 7

## REFERENCES

[1] KwetisheJoroDanjuma, ” Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients” Department of Computer Science, ModibboAdama University of Technology, Yola, Adamawa State, Nigeria

[2]International Journal of Engineering Research Technology (IJERT) ISSN: 2278-0181 Published by, [www.ijert.org](http://www.ijert.org) RTICCT - 2019 Conference Proceedings

[3]Discovering interesting prediction rules with a genetic algorithm. In Proceedings of 1999 Congress on Evolutionary Computation (CEC' 99), pp. 13221329.

[4]Zehra Karhan<sup>1</sup>, Taner Tunç<sup>2</sup>, ”Lung Cancer Detection and Classification with Classification Algorithms” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 22788727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 71-77.