Final Paper

**Submission date:** 23- MAR- 2023 10:56AM (UTC+0530)

**Submission ID:** 1085209946

**File name:** RAAGUL.docx (356.17K)

**Word count:** 5445

**Character count:** 29545

**LUNG CANCER PREDICTION USING**

**DECISION TREE CLASSIFIER ALGORITHM**

## P Vijayalakshmi1, S Raagul2,R Vinothkumar2,R Ram kumar 2

*Assistant Professor1, Under Graduate Students2,*

*Department of CSE, Panimalar Engineering College Chennai, Tamil Nadu, India*

***Abstract* — Lung cancer is a type of cancer that begins in the lungs. Your lungs are two spongy organs in your chest that take in oxygen when you inhale and release carbon dioxide when you exhale. The aim is to predict machine learning-based techniques for Lung cancer prediction results with the best accuracy. Lung cancer is the leading cause of cancer deaths worldwide. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information like, variable identification, univariate analysis, bi-variate, and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preprocess and data visualization will be done on the entire given dataset. To propose a machine learning-based method to accurately predict Lung cancer by prediction results in the form of pulmonary disease classification of best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given dataset with evaluation classification report, identify the confusion matrix, and to categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with the best accuracy with precision, Recall and F1 Score.**

**Index Terms— SMLT, Machine learning,Accuracy,Recall, F1 Score**

### INTRODUCTION

One of the most lethal types of the disease, lung cancer, is responsible for the passing away of about one million people every year. The current state of affairs in the world of medicine makes it absolutely essential to perform lung nodule identification on chest CT scans. This is due to the fact that lung nodules are becoming increasingly common. As a direct result of this, the deployment of CAD systems is required in order to accomplish the objective of early lung cancer identification. When doing a CT scan, sophisticated X-ray equipment is utilized in order to capture images of the human body from a number of

different angles. Following this, the images are fed into a computer, which processes them in such a way as to produce a cross-sectional view of the internal organs and tissues of the body. If lung cancer is detected at an early stage, the American Cancer Society estimates that a patient has a 47 percent chance of surviving the disease. It is quite unlikely that X-ray pictures may accidentally reveal lung cancer in its earlier stages. It is famously difficult to detect lesions that are round and have a diameter of 510 millimeters or less. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.The term “data science” was first coined in 2008 by D.J. Patil, and Jeff Hammer bacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market. Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

This chapter makes a present of an overall introduction of the lung cancer. It discusses the structure of the lung cancer, moreover types of it consisting benign and malignant tumors. In this chapter, aims of the thesis are also described. In addition to this, the contributions of the proposed work are discussed, as well as, the thesis overview and structure. Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body’s cells begin to divide without stopping and spread into surrounding tissues. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. In a wide range of cancer, some portion of the body’s cells, start to separate without halting and spread into encompassing tissues. Tumor can begin anywhere within these cells. In general the human cells develop and separate to form new cells as required by the body. When cells are old, new cells usually substitute them, however when cancer occurs, this process does not operate as it supposed to. The old cells do not die and the new cells are formed without necessity. The cells keep on dividing without any restrictions and forms outgrowths in the body called tumors. (Som et al., 2011) Tumors are usually solid and are strong masses of tissue. Malignancies of the blood, for example, leukemia, by and large do not shape strong tumors. The cancerous tumors can spread into, or attack the tissues close to them, these are called malignant. In addition, as these tumors grow, some cancer cells can break off and travel to distant places in the body through the blood or the lymph system and form new tumors far from the original tumor, but benign tumors are not like malignant tumors, they do not spread or attack the tissues surrounding them, or the tissues close to them. After removal either by surgery or by other treatment procedures, benign tumors do not grow back. This is unlike malignant tumors, which sometimes grow back after removal (Sayar et al., 2004 ). Generally, benign tumors are not life threatening, except for the benign tumors that occur in the brain. The brain benign tumors can be risky and can even be the cause of death of a person. 2 Cancer cells differ from normal cells in many ways that allow them to grow out of control and become invasive. The imperative contrast is that disease cells are not as specialized as normal human cells in the body. Normal human cells can grow into cells performing particular functions in the body but carcinogenic cells cannot. This causes them to spread

widely without a halt. The body uses a mechanism called programmed cell death also called as apoptosis, where it does away with unwanted cells (Travis et al., 2013). The cancer cells do not listen to signals sent by the body to stop dividing. The area around tumor cells in some cases, like the non-carcinogenic human cells, blood vessels, gets affected. This area is called microenvironment. The cancer cells can affect the normal cells in such a way that they are forced to create blood vessels to feed the tumors and can get rid of the excreta from the tumors. The immune system, which is a system of organs and concentrated cells that shields the body from diseases and different conditions are frequently dodged by the cancer cells. Tumors can also use the immune system to stay alive and grow. For example, with the help of certain immune system cells that normally prevent a runaway immune response, cancer cells can actually keep the immune system from killing cancer cells (Lin and Yan, 2002). The lungs are the parts of our body that we use to breathe. They supply oxygen to the organs and tissues of the body. The lungs are divided into areas called lobes. The right lung has three lobes and the left lung has two. Lung cancer is the type of cancer which unchecks the growth of unusual cells either in one or in both the lungs. These anomalous cells do not perform the functions of healthy human cells and do not mature into normal cells.

Lung cancer is a disease of abnormal cells multiplying and growing into a tumour. Cancer cells can be carried away from the lungs in blood, or lymph fluid that surrounds lung tissue. Lymph flows through lymphatic vessels, which drain into lymph nodes located in the lungs and in the centre of the chest. Lung cancer often spreads toward the centre of the chest because the natural flow of lymph out of the lungs is toward the centre of the chest.

Metastasis occurs when a cancer cell leaves the site where it began and moves into a lymph node or to another part of the body through the blood stream . Cancer that starts in the lung is called primary lung cancer. There are several different types of lung cancer, and these are divided into two main groups: Small cell lung cancer and non-small cell lung cancer which has three subtypes: Carcinoma, Adenocarcinoma and Squamous cell carcinomas. The rank order of cancers for both males and females among Jordanians in 2008 indicated that there were 356 cases of lung cancer accounting for (7.7 %) of all newly diagnosed

males and 59 (2.5%) females with a male to female ratio of 5:1 which Lung cancer ranked second among males and 10th among females

### LITERATURE REVIEW

1. “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques“by William et.al (2015) The methodology used in this edition involve use of immunohistochemistry throughout the classification, a new emphasis on genetic studies, in particular, integration of molecular testing to help personalize treatment strategies for advanced lung cancer patients, a new classification for small biopsies and cytology.
2. ” Deep Convolution Network Based Prediction Model For Medical Diagnosis Of Lung Cancer - A Deep Pharmacogenomic Approach” by Akash iyer et.al(2018)

.In this paper, the well-established pre-trained model, VGG19 has been used to extract information from the pathological images specific to PTEN, EGFR, ERBB2, BRAF and CDKN2A for Lung cancer. The model finds to be suitable for the prediction of these mutations, specific to Indian populations. The pathological image processing using deep convolutional network can been used as an ingenious and effective technique for further investigation of the suspected cases.

1. ” Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier” by Janee Alam (2018). This paper proposed an efficient lung cancer detection and prediction algorithm using multi-class SVM (Support Vector Machine) classifier. Multi-stage classification was used for the detection of cancer. This system can also predict the probability of lung cancer. The main problem of these techniques is that the detection was confined to solitary pulmonary nodules . The precision of the system can be extemporized via preparing it on a huge image set and arrangement in light of hereditary calculation of genetic algorithm and deep neural network.
2. ” Random Forest based Classification Model for Lung Cancer Prediction on Computer Tomography

automated computer aided model has been developed to identify lung cancer on CT images. The presented model operates on four main stages namely preprocessing, segmentation, feature extraction and classification

1. ”Texture Analysis Based Feature Extraction and Classification of Lung Cancer” by Sanjuktha rani et.al(2019). The primary objective of pre-processing is to enhance image quality to build it for diminishing or evacuating the irrelevant parts of the images. Pre- processing stage is important to enhance quality of image. The noise and other high recurrence segments are evacuated by filters and prepare the datasets for additional processing.
2. ” Detect lung cancer nodule using fuzzy interference system and active contour model” by Roy et al(2019)This system uses gray transformation for image contrast enhancement. Image binarization is performed before segmentation and resulted image is segmented using active contour model. Cancer classification is performed using fuzzy inference method. Features like area, mean, entropy, correlation, major axis length, minor axis length are extracted to train the classifier. Overall, accuracy of the system is 94.12%. Counting its limitation it does not classify the cancer as benign or malignant which is future scope of this proposed model
3. ” K mean unsupervised learning algorithm for clustering or segmentation” by Sangamithra et.al(2019) . It groups the pixel dataset according to certain characteristics. For classification this model implements back propagation network. Features like entropy, correlation, homogeneity, PSNR, SSIM are extracted using gray-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. Image pre processing median filter is used for noise removal which can be useful for our new model to remove the noise and improve the accuracy.
4. ”Lung Disease Detection Using Deep Learning” by Siddarth (2021). In this project , they will conduct a study and analysis of this data set, then apply Machine Learning and Deep Learning to predict that the patient has a lung disease. This project is a binary classification with input is

patient's data (age, gender, X-ray images, View Position) and output is found for diseases or not. The selected model has not really had small loss and rapid convergence, so still need more improvement can be resolved to identify lung disease in the hospital. Increasing the size of training shots, this will increase the chance of getting important features, but it also means that the model will be more complex and the training time and prediction will be longer.

1. ” Lung Cancer Prediction Using Robust Machine Learning and Image Enhancement Methods on Extracted Gray-Level Co-Occurrence Matrix Features by Lal Hussain(2022). This study aimed to improve lung cancer image quality by utilizing and employing various image enhancement methods, such as image adjustment, gamma correction, contrast stretching, thresholding, and histogram equalization methods. The present study was carried out on a small lung cancer dataset provided by the Lung Cancer Alliance on lung cancer types, i.e., NSCLC and SCLC. In the future, we shall apply the proposed methods based on image enhancement, feature extraction, and ranking and machine learning methods on larger datasets with more clinical details, disease severity levels, and more types, and larger datasets acquired on different imaging modalities.
2. ” Deep Learning-Based Classification of Reduced Lung Ultrasound Data From COVID-19 Patients by Umair khan et.al(2022). This study evaluates the performance of DL algorithm overLUS data with varying pixel and gray- level resolution. The algorithm is evaluated over a dataset of 448 LUS videos captured from 34 examinations of 20 patients. All videos are resampled by a factor of 2, 3, and 4 of original resolution, and quantized to 128, 64, and 32 levels, followed by score prediction. However, its high cost,limited availability, and exposure of patients to radiations make it less preferable. As a future work, it focus on evaluating our study on different DL algorithms and use this study as a benchmark for comparison of prediction performance on reduced LUS data.

### EXISTING SYSTEM

In existing system, prediction of lung cancer is done using lung ultrasound images.Deep learning uses convolutional neural network [CNN] to recognize and classify the images.output obtained in grey scale reduce the accuracy.As only one algorithm is used,82.3% is the

maximum accuracy that can be achieved.In this system,only medical officials, can process& predict the output.

### PROPOSED SYSTEM

In proposed system,data is preprocessed and visualized initially machine learning algorithm is used to classify the output.Here four types of algorithm are used.Hence highest accuracy can be achieved upto 98.2%.By comparing the four outputs from different algorithms ,the most accurate model is chosen and trained for deployment to predict.In this system, anyone can feed the data and analyse the presence of lung cancer.the advantages: Machine Learning Technology is implemented,Performance metrics will be calculated and Deployment is done

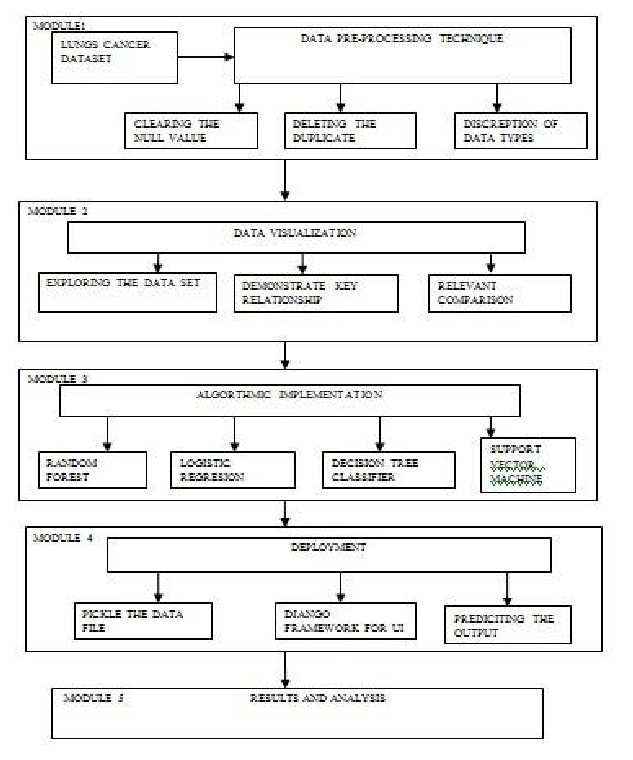
### SYSTEM ARCHITECTURE

* 1. **DATA PRE-PROCESSING**

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.A number of different **data cleaning** tasks using Python’s [Pandas library](https://pandas.pydata.org/) and specifically, it focus on probably the biggest data

cleaning task, **missing values** and it able to **more** [**quickly**](https://www.dataoptimal.com/data-cleaning-with-python-2018/)[**clean data**.](https://www.dataoptimal.com/data-cleaning-with-python-2018/) It wants to **spend less time cleaning data**, and more time exploring and modeling. Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It’s important to understand these [different types of missing data](https://en.wikipedia.org/wiki/Missing_data) from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for [dealing with missing data.](https://github.com/matthewbrems/ODSC-missing-data-may-18/blob/master/Analysis%20with%20Missing%20Data.pdf) Before, joint into code, it’s important to understand the sources of missing data. Here are some typical reasons why data is missing: User forgot to fill in a field.Data was lost while transferring manually from a legacy database. There was a programming error.Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted



### DATA VISUALIZATION

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.How to chart time series data with line plots and categorical quantities with bar charts.

### Fig 4.1 SYSTEM ARCHITECTURE

* 1. **ALGORITHMS**

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies. The below 4 different algorithms are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine

### 4.3.1. LOGISTIC REGRESSION:

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable

= response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).The assumptions are 1. Binary logistic regression requires the dependent variable to be binary.2.For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome, 3.Only the meaningful variables should be included.4.The independent variables should be independent of each other. That is, the model should have little.

### 4.3.2 DECISION TREE CLASSIFIER:

Decision tree learning is a supervised learning approach used in [statistics,](https://en.wikipedia.org/wiki/Statistics) [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine](https://en.wikipedia.org/wiki/Machine_learning) [learning.](https://en.wikipedia.org/wiki/Machine_learning) In this formalism, a classification or regression [decision tree](https://en.wikipedia.org/wiki/Decision_tree) is used as a [predictive model](https://en.wikipedia.org/wiki/Predictive_model) to draw conclusions about a set of observations. Tree models where the target variable can take a discrete set of values are called [classification](https://en.wikipedia.org/wiki/Classification) [trees](https://en.wikipedia.org/wiki/Decision_tree); in these tree structures, [leaves](https://en.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically [real numbers](https://en.wikipedia.org/wiki/Real_numbers)) are called [regression](https://en.wikipedia.org/wiki/Regression_analysis) [trees.](https://en.wikipedia.org/wiki/Decision_tree) Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity

### 4.3.3. RANDOM FOREST CLASSIFIER

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of over fitting to their training set. Random forest is a type of

supervised machine learning algorithm based on [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) [learning.](https://en.wikipedia.org/wiki/Ensemble_learning) Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The [random forest](https://en.wikipedia.org/wiki/Random_forest) algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

.

### 4.3.4 SUPPORT VECTOR MACHINE:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

### DEPLOYMENT:

In this module the trained machine learning model is converted into pickle data format file (.pkl file) which is then deployed in our django framework for providing better user interface and predicting the output of how much the given data is emitting Co2.

### FLASK (WEB FRAMEWORK) :

Flask is a micro web framework written in Python.It is classified as a micro-framework because it does not require particular tools or libraries.It has no

database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.However, Flask supports extensions that can add application features as if they were implemented in Flask itself.Extensions exist for object- relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

### FEATURES:

Flask was designed to be **easy to use and extend**. The idea behind Flask is to build a solid foundation for web applications of different complexity. From then on you are free to **plug in any extensions** you think you need. Also you are free to build your own modules. Flask is great for all kinds of projects. It's especially good for prototyping. Built-In Development server and Fast debugger, integrated support for unit testing, dispatching, Unicode based Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Flask framework and is also easier to learn because it has less base code to implement a simple web-Application Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.Flask is a web framework. This means Flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, and a wiki or go as big as a web-based calendar application or a commercial website. The advantages of Flask are Higher compatibility with latest technologies and high scalability.

### SYSTEM TESTING

**5.1 PERFORMANCE METRICS :**

**False Positives (FP):** A person who will pay predicted as

defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive

but predicted class tells you that this passenger will survive.

**False Negatives (FN):** A person who default predicted as payer. When actual class is yes but predicted class in no.

E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

**True Positives (TP):** A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

**True Negatives (TN):** A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

**Accuracy**: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

**Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

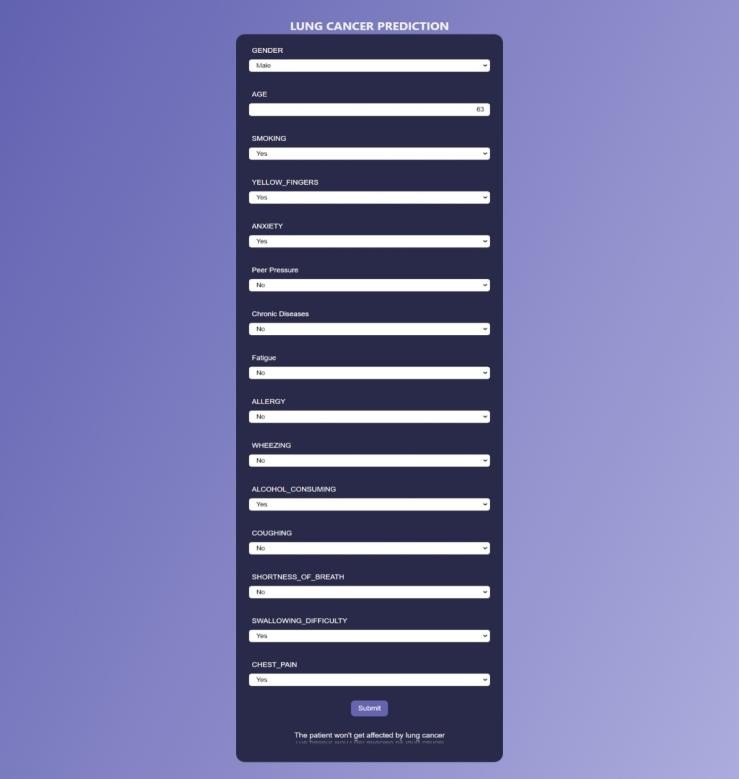
**Recall:** The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict).Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

**F1 Score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

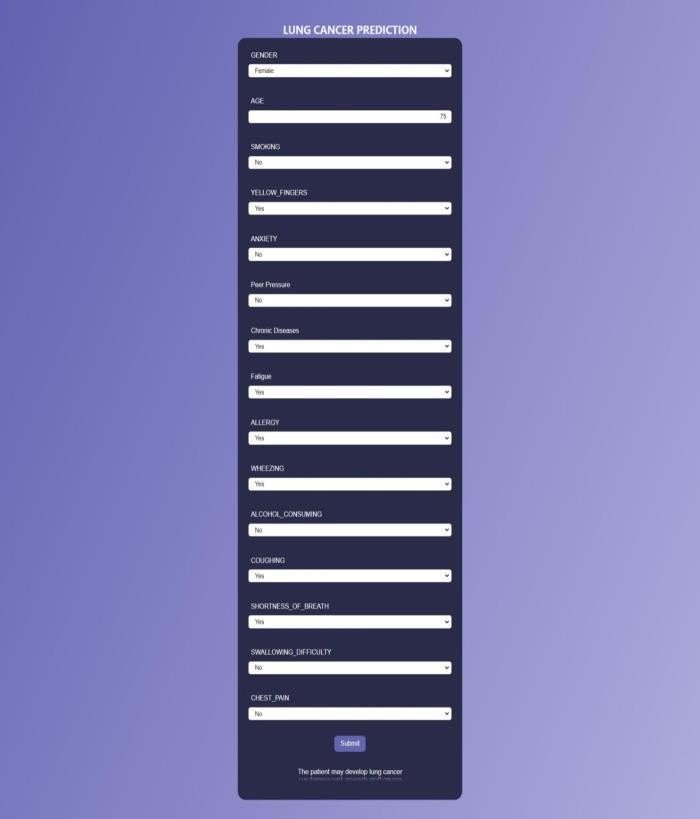
### RESULTS & DISCUSSION

The best combination of supervision algorithm has been used and its been decided that Decision Tree Classifier has showed the best accuracy rate and this project has used this algorithm for the further process. This model has showed the accuracy rate of 98.148%. The output for this project can be divided has into two types :

1.Person not - affected with Lung Cancer 2. Person may develop Lung Cancer



### FIG 6.1 PERSON NOT AFFECTED WITH LUNG CANCER



**FIG 6.2 PERSON MAY DEVELOP LUNG CANCER**

### CONCLUSION AND FUTURE ENHANCEMENTS:

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the lung cancer prediction. The future enhancements are deploying the project in the cloud and to optimize the work to implement in the IOT system.

### REFERENCES

* 1. P. Joshi, R. K. Tyagi, and K. M. Agarwal, “Technological resources for fighting COVID-19 pandemic health issues,” J. Ind. Integr. Manage., vol. 6, no. 2, pp. 271–285, Jun. 2021.
  2. Y. Yang et al., “Laboratory diagnosis and monitoring the viral shedding of SARS-CoV-2 infection,” Innov., vol. 1, no. 3, Nov. 2020, Art. no. 100061.
  3. S. Woloshin, N. Patel, and A. S. Kesselheim, “False negative tests for SARS-CoV-2 infection—Challenges and implications,” New EnglandJ. Med., vol. 383, no. 6, p. e38, Aug. 2020.
  4. S. Salehi, A. Abedi, S. Balakrishnan,, “Coronavirus disease 2019 (COVID-19): A systematic review of imaging find ings in 919 patients,” Amer. J. Roentgenol., vol. 215, no. 1, pp. 87–93, Jul. 2020.
  5. G. Soldati et al., “Is there a role for lung ultrasound during the COVID-19 pandemic?” J. Ultrasound Med., vol. 7 no. 39,pp. 1459–1462, 2020.
  6. L. Demi, “Lung ultrasound: The future ahead and the lessons learned from COVID-19,” J. Acoust. Soc. Amer., vol. 148, no. 4, pp. 2146–2150,

Oct. 2020.

* 1. G. Soldati, M. Demi, A. Smargiassi, R. Inchingolo, and

L. Demi, “The role of ultrasound lung artifacts in the diagnosis of respiratory diseases,” Expert Rev. Respirat. Med., vol. 13, no. 2, pp. 163–172, 2019.

* 1. E. Picano and P. A. Pellikka, “Ultrasound of extravascular lung water: A new standard for pulmonary congestion,” Eur. Heart J., vol. 37, no. 27 pp. 2097–2104, 2016.
  2. G. Soldati, M. Demi, R. Inchingolo, A. Smargiassi, and

L. Demi, “On the physical basis of pulmonary sonographic interstitial syndrome,”J. Ultrasound Med., vol. 35, no. 10, pp. 2075–2086, Oct. 2016.

* 1. R. Copetti, G. Soldati, and P. Copetti, “Chest sonography: A useful tool to differentiate acute cardiogenic pulmonary edema from acute respiratory distress syndrome,” Cardiovascular Ultrasound, vol. 6, no. 1, pp. 1–10, Apr. 2008.
  2. L. Demi, M. Demi, R. Prediletto, and G. Soldati, “Real- time multi- frequency ultrasound imaging for quantitative lung ultrasound–first clinical results,” J. Acoust. Soc. Amer., vol. 148, no. 2, pp. 998–1006, 2020.
  3. F. Mento, G. Soldati, R. Prediletto, M. Demi, and L. Demi, “Quantitative lung ultrasound spectroscopy applied to the diagnosis of pulmonary fibrosis: The first clinical study,” IEEE Trans. Ultrason., Ferroelectr.Freq. Control, vol. 67, no. 11, pp. 2265–2273, Nov. 2020.
  4. F. Mento and L. Demi, “On the influence of imaging parameters on lung ultrasound B-line artifacts, in vitro study,” J. Acoust. Soc. Amer., vol. 148, no. 2, pp. 975–983, Aug. 2020.
  5. L. Demi, W. van Hoeve, R. J. G. van Sloun, G. Soldati, and M. Demi, “Determination of a potential quantitative measure of the state of the lung using lung ultrasound spectroscopy,” Sci. Rep., vol. 7, no. 1,pp. 1–7, Dec. 2017.
  6. K. Mohanty, J. Blackwell, T. Egan, and M. Müller, “Characterization of the lung parenchyma using ultrasound multiple scattering,” Ultrasound Med. Biol., vol. 43, no. 5, pp. 993–1003, 2017.

|  |  |  |  |
| --- | --- | --- | --- |
| Final Paper |  | | |
| ORIGINALITY REPORT |  |  |  |
| 12% | 8% | 1% | 6% |
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |
| PRIMARY SOURCES |  |  |  |

1 pdfs.semanticscholar.org

7%

Internet Source

# 2 Submitted to TechKnowledge

4%

Student Paper

# 3 Yang Yang, Yu Zhao. "Quantification and evaluation of atmospheric pollutant emissions

from open biomass burning with multiple methods: A case study for Yangtze River Delta region, China", Atmospheric Chemistry and Physics Discussions, 2018

Publication

<1%

Exclude quotes On Exclude bibliography On

Exclude matches < 3 words