# Lung Cancer Prediction Using Machine Learning Techniques

**Sanjay.M[1], N.Ayyanathan[2]**
[1]Dept of Computer Applications
[2]Associate Professor, Dept of Computer Applications
[1, 2] B. S. Abdur Rahman Crescent Institute of Science & Technology, Vandalur, Chennai-600 048, India

***Abstract-*** *Lung cancer is due to uncontrollable growth of cells in the lungs. It causes a serious breathing problem in both inhale and exhale part of chest. Cigarette smoking and passive smoking are the principal contributor for the cause of lung cancer as per world health organization. The mortality rate due to lung cancer is increasing day by day in youths as well as in old persons as compared to other cancers.The aim is to predict machine learning based techniques for lung cancer prediction. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset.To propose a machine learning-based method to accurately predict the lung cancer using supervised classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface of lung cancer prediction by attributes.*

***Keywords***- Dataset, Machine learning-Classification method, python, Prediction of Accuracy result.

## I. INTRODUCTION

Lung cancer considers as the deadlier disease and a primary concern of high mortality in present world. Lung cancer affects human being at a greater extent and as per prediction it now takes 7th position in mortality rate index causing 1.5% of total mortality rate of the world . Lung cancer originates from lung and spreads up to brain and spreads Lung cancer is categorized in to two major group. One is non-small cell lung cancer and another is small cell lung cancer. Some of the symptoms which are associated with the patients like severe chest pain, dry cough, breathlessness, weight loss etc. Looking in to the cultivation of cancer and its causes doctors give stress more on smoking and second-hand smoking as if the primary causes of lung cancer. Treatment of lung cancer involves surgery, chemotherapy, radiation therapy, Immune therapy etc. In-spite of this lung cancer diagnosis process is very weak because doctor will able to know the disease only at the advanced stage . Therefore early prediction before final stage is highly important so that the mortality rate can be easily prevented with effective control. Even after the proper medication and diagnosis survival rate of lung cancer is very promising. Survival rate of lung cancer differs from person to person. It depends on age, sex and race as well as health condition. Machine learning now days plays a crucial role for detection and prediction of medical diseases at early stages of safe human life.

## II. PROBLEM DEFINITION

Machine learning algorithms provide an opportunity to effectively predict detect and prevent negative forms of human behavior, such as Lung cancer is most life-threatening disease, treatment of which must be the primary goal throughout scientific research. The early recognition of cancer can be helpful in curing disease entirely. There are numerous techniques found in literature for detection of lung cancer. Several investigators have contributed their facts for cancer prediction. These papers largely pact about prevailing lung cancer detection techniques. We analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose a machine learning-based method to accurately predict the lung cancer using supervised classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface of lung cancer prediction by attributes.

## III. LITERATURE REVIEW

Texture Analysis Based Feature Extraction and Classification of Lung Cancer
Sanjukta Rani Jena, Dr. Thomas George, Dr. Narain Ponraj

Lung cancer is most life-threatening disease, treatment of which must be the primary goal throughout scientific research. The early recognition of cancer can be helpful in curing disease entirely. There are numerous techniques found in literature for detection of lung cancer. Several investigators have contributed their facts for cancer prediction. These papers largely pact about prevailing lung cancer detection techniques that are obtainable in the literature. A numeral of methodologies has been originated in cancer detection methodologies to progress the efficiency of their detection. Diverse applications like as support vector machines, neural networks, image processing techniques are extensively used in for cancer detection which is elaborated in this work

The early discovery of lung malignancy is a confront, because of the structure of tumour cells, where the greater part of the cells are covered with each other. This paper has surveyed numerous strategies, to distinguish the lung tumour in its beginning periods. The manual examination of the samples is tedious, inaccurate and requires intensive trained person to eliminate diagnostic errors. From the results obtained we could conclude that the Local Binary Pattern performs better than other basic textural patterns as the histogram features obtained were greater than that of the latter. Multiple Resolution Residually Connected Feature Streams For Automatic Lung Tumor Segmentation From CT Images.

Jue Jiang , Yu-chi Hu , Chia-Ju Liu ,Darragh Halpenny

Volumetric lung tumor segmentation and accurate longitudinal tracking of tumor volume changes from computed tomography (CT) images are essential for monitoring tumor response to therapy. Hence, we developed two multiple resolution residually connected network (MRRN) formulations called incremental-MRRN and dense-MRRN. Our networks simultaneously combine features across multiple image resolution and feature levels through residual connections to detect and segment lung tumors. We evaluated our method on a total of 1210 non-small cell (NSCLC) lung tumors and nodules from three datasets consisting of 377 tumors from the open-source Cancer Imaging Archive (TCIA), 304 advanced stage NSCLC treated with anti- PD-1 checkpoint immunotherapy from internal institution MSKCC dataset, and 529 lung nodules from the Lung Image Database Consortium (LIDC). The algorithm was trained using the 377 tumors from the TCIA dataset and validated on the MSKCC and tested on LIDC datasets. The segmentation accuracy compared to expert delineations was evaluated by computing the Dice Similarity Coefficient (DSC), Hausdorff distances, sensitivity and precision metrics. Our best performing incremental-MRRN method produced the highest DSC of 0.74±0.13 for TCIA, 0.75±0.12 for MSKCC and 0.68±0.23 for the LIDC datasets. There was no significant difference in the estimations of volumetric tumor changes computed using the incremental-MRRN method compared with expert segmentation proposed two neural networks to segment lung tumors from CT images by adding multiple residual streams of varying resolutions. Our results clearly demonstrate the improvement in segmentation accuracy across multiple datasets.

Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis
Naji Khosravan and Ulas Bagci

Early detection of lung nodules is of great importance in lung cancer screening. Existing research recognizes the critical role played by CAD systems in early detection and diagnosis of lung nodules. However, many CAD systems, which are used as cancer detection tools, produce a lot of false positives (FP) and require a further FP reduction step. Furthermore, guidelines for early diagnosis and treatment of lung cancer are consist of different shape and volume measurements of abnormalities. To support this hypothesis we proposed a 3D deep multi-task CNN to tackle these two problems jointly. We tested our system on LUNA16 dataset and achieved an average dice similarity coefficient (DSC) of 91% as segmentation accuracy and a score of nearly 92% for FP reduction. As a proof of our hypothesis, we showed improvements of segmentation and FP reduction tasks over two baselines. e proposed a 3D deep multi-task CNN for simultaneously performing segmentation and FP reduction. We showed that sharing some underlying features for these tasks and training a single model using shared features can improve the results for both tasks, which are critical for lung cancer screening. Furthermore, we showed that a semisupervised approach can improve the results without the need for large number of labeled data in the training.

## IV. EXISTING SYSTEM

Detecting lung nodules with low-dose computed tomography (CT) can predict the future risk suffering from lung cancers. There are a few studies on lung nodules with low-dose CT and detecting rate is very low at present. In order to accurately detect lung nodules with low-dose CT, this paper proposes a solution based on an integrated deep learning algorithm. The CT images are preprocessed via image clipping, normalization and segmentation, and the positive samples are expanded to balance the number of positive and negative samples. The features of candidate lung nodule samples are learned by using convolutional neural network and residual network, and then import into long short-term

memory network, respectively. We then fuse these features, continuously optimize the network parameters during the training process, and finally obtain the model with an optimal performance. The experimental results prove that compared to other algorithms, all metrics in the proposed algorithm are improved. This model has an obvious anti-interference ability. It is stable and can identify lung nodules effectively, which is expected to provide auxiliary diagnostic for early screening of lung cancers.

**DISADVANTAGES OF EXISTING SYSTEM**

They are using CT images and large amount of data required to classify accurately and it take a lot of time to train the model.Any modification made takes lots of effort to change the model.

## V. PROPOSED SYSTEM

Machine learning supervised classification algorithms will be used to give dataset and extract patterns, which would help in predicting the likely patient affected or not, thereby helping them for making better decisions in the future.The data set collected for predicting the network attacks is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using by ensemble learning model are applied on the Training set and based on the test result accuracy, Test set prediction is done.

**ADVANTAGES OF PROPOSED SYSTEM**

It improves accuracy score by comparing popular machine learning algorithms.These reports are to the investigation of applicability of machine learning techniques for detecting cancer in operational conditions by attribute prediction.

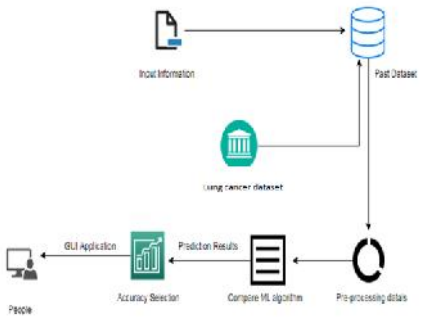



Fig1

## VI. REQUIREMENTS SPECIFICATION

### 6.1 HARDWARE REQUIREMENTS

Processor : Pentium IV/III
Hard disk : minimum 80 GB
RAM : minimum 2 GB

### 6.2 SOFTWARE REQUIREMENTS

Operating System : Windows
Tool : Anaconda with Jupyter Notebook

## . MODULES

### 7.1 Data validation process EDA(Exploratory Data Analysis)

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 uses 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. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset)

### 7.2 Exploration data analysis of 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.

**7.3Logistic 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.).

In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

**7.4 Support Vector Machines (SVM)**

A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

- How to disentangle the many names used to refer to support vector machines.
- The representation used by SVM when the model is actually stored on disk.
- How a learned SVM model representation can be used to make predictions for new data.
- How to learn an SVM model from training data.
- How to best prepare your data for the SVM algorithm.
- Where you might look to get more information on SVM.

**. EXPERIMENTALSCREENSHOT OUTPUT**

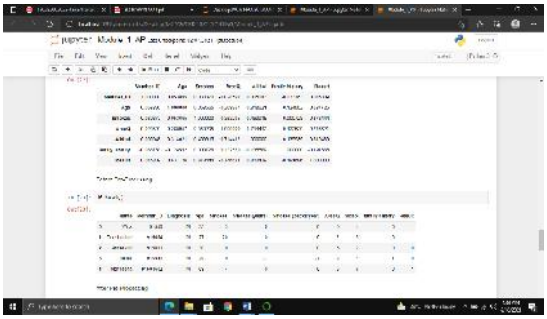


Fig 7.1

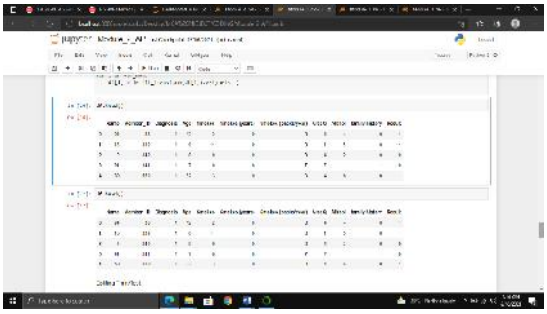


Fig7.2

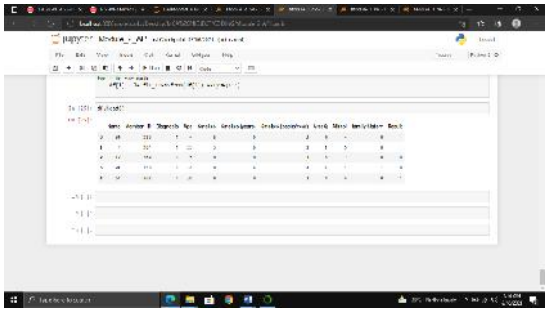


Fig 7.2

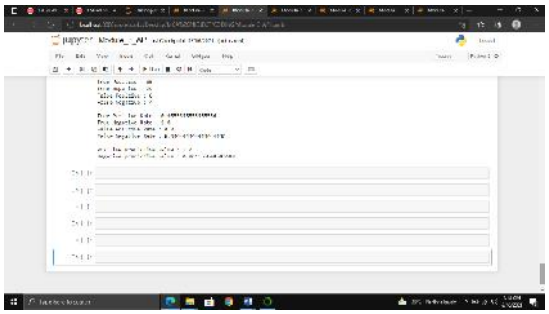
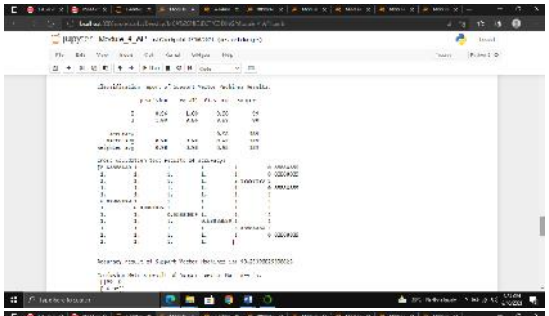


Fig 7.3



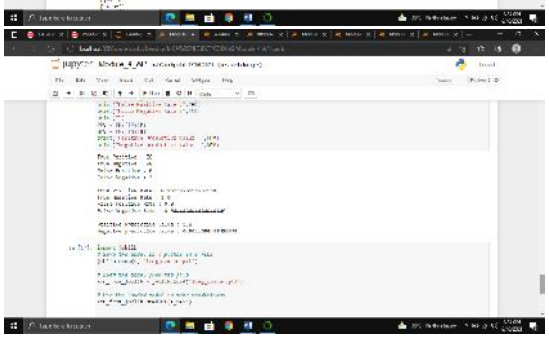
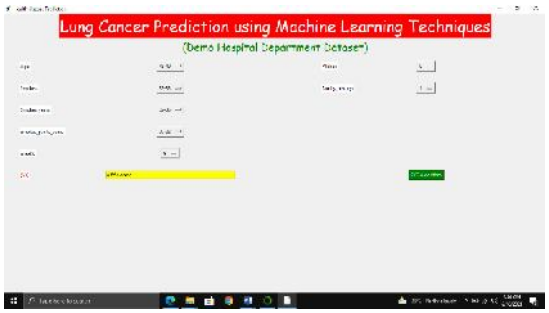


Fig 7.4

**OUTPUT**




## VII. CONCLUSION

The process of the project started from data cleaning and pre-processing, missing value identification and exploratory analysis of the lung cancer dataset and finally model building and evaluation. The algorithms comparison is done between the machine learning algorithms and the logistic regression performed better so the model (.pkl) is taken from there using joblib and it is deployed in Graphical User Interface( GUI ) which is built using tkinter packages and the output is predicted whether the patient is affected or not based on the given input.

## VIII. FUTURE ENHANCEMENTS

To automate this process by show the prediction result in web application or desktop application.To optimize the work to implement in Artificial Intelligence environment.

## REFERENCES

[1] Texture Analysis Based Feature Extraction and Classification of Lung Cancer
Author: Sanjukta Rani Jena, Dr. Thomas George, Dr. Narain Ponraj

[2] Multiple Resolution Residually Connected Feature Streams For Automatic Lung Tumor Segmentation From CT Images
Author: Jue Jiang , Yu-chi Hu , Chia-Ju Liu , Darragh Halpenny

[3] Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis
Author: Naji Khosravan and Ulas Bagci

[4] Detection of Lung Cancer in CT Images using Image Processing
Author: Nidhi S. Nadkarni, Prof. Sangam Borkar

[5] High incidence of radiation pneumonitis in lung cancer patients with chronic silicosis treated with radiotherapy
Author: Tianle Shen, Liming Sheng, Ying Chen , Lei Cheng and Xianghui Du

[6] Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening
Author: SAEED S. ALAHMARI , DMITRY CHEREZOV , DMITRY B. GOLDGOF

[7] A graphical model of smoking-induced global instability in lung cancer
Author: Yanbo Wang , Weikang Qian , Bo Yuan