

# Prediction and Prognosis of Lung Cancer using Machine Learning Techniques

Dr. Shubhangi Vinayak Tikhe<sup>1</sup>

*Project Guide*

Sonali Bhushan  
Ambadkar<sup>2</sup>

Prachi Shekhar  
Kulkarni<sup>3</sup>

Rashmi Vishwanath  
Papulwar<sup>4</sup>

Swarali Mukund  
Patil<sup>5</sup>

*<sup>1-5</sup>B.Tech, Computer Engineering,*

*MKSSS's Cummins College of Engineering for Women, Pune, India*

-----  
\*\*\*  
-----

**Abstract** - Various types of cancer are on the rise today mainly skin cancer, lung cancer, prostate cancer, breast cancer etc. With limited knowledge on the cure for cancer, curing it becomes even more difficult. There are developments in the field of research to contribute to finding a cure to cancer. Various technologies are now making early detection and prognosis possible. Machine Learning techniques, Deep Learning techniques are on the forefront of this mission. This paper elaborates on a prediction method to achieve early detection of lung cancer.

**Key Words :** Lung Cancer, Artificial Neural Networks (ANN), Deep Learning

## 1. INTRODUCTION

Cancer is the world's second-leading cause of deaths. Abnormal cells develop in the body which divide unmanageably. When they invade and destroy the normal body tissue, it can cause a large number of diseases. They can also spread to other organs. Cancer refers to these diseases. Lung Cancer cases in 2020 stand at 2.21 million. Nearly 10 million deaths have been caused by cancer in 2020, with lung cancer being the highest at 1.80 million deaths [1]. Some of the trigger causes of lung cancer are smoking, intake of tobacco, contact with the radon gas, asbestos and other carcinogens.

In this paper we will be covering a technique of Machine Learning to achieve the outcome of prediction and prognosis of lung cancer. Artificial Neural Networks are useful in solving real body functioning problems using artificial intelligence. These Neural networks are useful for predictive modeling, adaptive control and applications.

Artificial Neural Network mechanism is discussed which works on the principle of neural network of our brain in which information is passed from one neuron to another. A neural network is basically a network of neurons which are designed in layers. It is an artificial network of neuron circuits similar to the ones present in our body. It provides a relatively simpler model as compared to the actual human brain. The input or predictors form the bottom layer, whereas the outputs forecasts form the top layer. The intermediate layers are hidden layers.

The independent variables in the dataset are fed into the input layer. They are then passed on to the hidden layer.

The core of the neural network is the hidden layer. It computes the parameters which are important to the prediction. To achieve this, weights are assigned to the parameters [2]. They are then passed on to the output layer. Assigning weights to the parameters is done by using stochastic gradient descent or a comparatively newer method, adam, can be used. Adam can be said to be a combination of the RMSProp and AdaGrad

algorithms. It is very efficient when working with large data or parameters [3].



Fig 1.2 - Adam Optimizer

Lastly, the output layer receives the prediction result from the hidden layer and produces output.

## 2. METHODOLOGY

In neural networks, the hidden layers work on the data to provide something useful which the output layer can make sense of [4]. Firstly, the ANN needs to be trained to give an accurate output. For this, the ANN goes through two phases, training and testing [5]. A dataset is used to train the model upon. As discussed above, weights are assigned to the parameters.

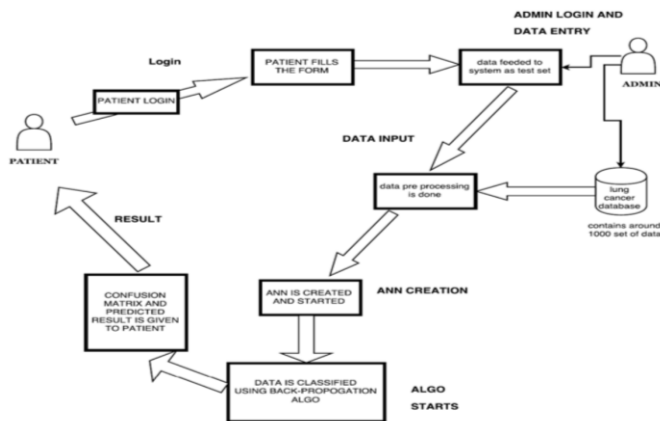


Fig 2.1 - System Architecture

### 2.1 Login & Registration Module

In the login & registration module, the user is validated when his email address and password gets matched. And if the user is not present in the database records then he is supposed to register himself

through the registration module using an email address and set a password as the site is password enabled to protect the privacy of the user. Only then can the user proceed to avail the services. Later on, the user can just login using his/her email id and password.

### 2.2 Data Design Module

After the user is authenticated, they are eligible to provide data to the system. The details entered will go to the dataset. The details are saved using POST variables in PHP. They can also be used for prediction by storing them in a dataset using PHP. From the given details, the relevant attributes are shortlisted and inserted in the new dataset. This refined data is then used for further ANN training. A few basic details like age, sex, occupation are required along with the individual's smoking status.

Cigar Smoking	Cigarette Smoking	Pipe Smoking
Packs smoked per day	Regular Smoker? Filtered or Unfiltered?	Current BMI
Age when started smoking	Age when stopped smoking	Number of years since stopped smoking

Table 2.1 - Smoking Attributes

### Global Data Structure

- Features: A stack used to store the read records
- Next element: Used to hold the current batch values. It isn't activated until run in a tensorflow session.
- Current batch: Used to store the current number of batches.
- Dataset: A dataframe data structure used to store the entire dataset in a format that can be used by tensorflow to create batches.

- X test, y test, X train, y train: Training and testing variable arrays for X and y
- Accuracy array: An array which stores the accuracies at every epoch which is then later used to visualize the accuracy increase with each epoch.
- Output probabilities: An array storing the probabilities of test data for occurrence of lung cancer.

### 2.3 Data Preprocessing and Visualization Module

The dataset will first be preprocessed to manage any data inconsistency. Preprocessing technique involves managing missing values, categorical data encoding and noisy data. For noisy data and outliers, FeatureScaling is used and missing values are filled by calculating mean or global constant. The LabelEncoder library is used for encoding categorical data. This helps in cleaning the data and making it suitable for the machine learning model. This data is divided into equal size batches to be then provided to the ANN model. The Analysis can be viewed in the pictorial form with the help of Pie Chart or Histogram.

### 2.4 Deployment of Neural Network Module

This module is the foundation for the machine learning model which consists of neural network structure with layers. The neural network comprises an input layer which is a matrix of the number of records as the rows and number of attributes as the columns. Since we have 23 attributes, we get [Number of records] X [23]. Then there are two hidden layers which associate weights with the parameters. Optimizer used is adam and the activation used is sigmoid and rectifier linear function i.e. relu. In the output layer, we have 2 classes, which would be either a patient has lung cancer or not.

### 2.5 ANN Backpropagation Module

In machine learning, Backpropagation is one of the most common algorithms used. It works by computing the gradient of the loss function. In this model, the dataset is divided into training and testing data in the ratio of 4:1 respectively. The training data is provided in batches to the ANN. It is worked upon in epochs and the accuracy increases with each epoch. The trained model will then be used to test the accuracy of the system using the test data and the new input data. Tensorflow can be used to implement this algorithm.

### 2.6 Result Extraction & Display Module

Output to the user: The result derived from the ANN module will be evaluated and Lung Cancer risk percentage of the user will be evaluated. The user will also be able to see the confusion matrix and if the user has a higher chance of getting lung cancer the proper precautionary measures will be suggested along with changes in diet and lifestyle.

Server side mailing mechanism: This information and report will be sent to the user through mail using server push technology. First user requests will be checked in the database. If the request is pending and the report is ready then the file is pushed to the email address of the user by checking the correct email address of the user in the database. Once the email is sent, the request is completed and these steps are repeated until all requests are met.

Flask provides tools, libraries and technologies that allow you to build a web application. In this project Flask is used to integrate the front end html code with the machine learning model. It enables the system to be deployed on the web, thus the trained model is converted into a web application. On a user request, we run the interface using the saved model in the server and return results in the UI. The database is integrated with the front end where the user would input the login credentials. The report is made available to the user using the email id.

### 3. SYSTEM TESTING AND RESULT

#### 3.1 Unit Testing

Login and registration module: A user is validated before he can provide his details and username and password are cross-checked through the MySQL database for different cases. Registration is tested against incorrect data and has server side validation.

Questionnaire: Server side validation ensures that incorrect data is not entered in the dataset. Radio buttons are used at most places which ensure that only specified data is entered in the system.

Data entry module: Different data array sizes were tested whether all of them are being appended in the test dataset. Appropriate data types are used in the php code.

Preprocessing module: Empty and missing values are fixed using global constants. All records are checked to make sure float is not present as a data type.

Batch creation module: Batch should be tested that it is a numpy array. Training dataset was going out of index at a point after some batches, it was resolved by saving the current batch and only iterating the data through the number of batches remaining.

Neural network model: Neural network is tested against different numbers of data, batches and epochs. It is concluded while testing that a batch of 100 with 30 epochs gave the most desirable (accuracy/performance) ratio. Test data was initially very obvious and made it difficult to predict unobvious test data hence the entire dataset was shuffled before prediction. At first accuracy was very low while testing then the data was normalized and accuracy was desirable then.

Visualization module: Visualization was tested against different values and checked if the visualization was accurate.

#### 3.2 Result

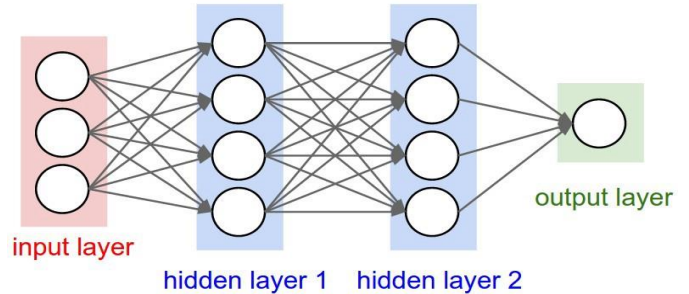


Fig 3.1 - Artificial Neural Network creation

This neural network module consists of an input layer, two hidden layers and an output layer.

The user gets the probability of contradicting lung cancer. The probability ranges from 0 to 1, with 1 being the highest chance of infection. The report will be sent to them through mail. Along with the suggestions on how to change their diet or lifestyle, if the user has a higher chance of getting cancer. Visual report is also provided.

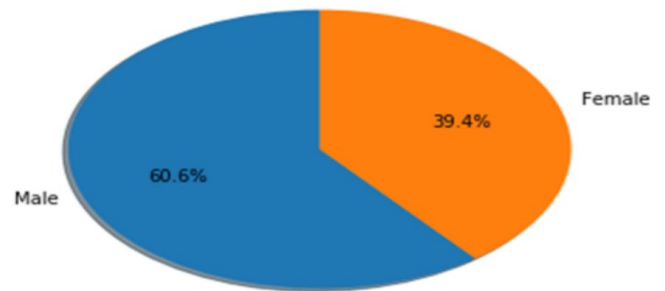


Fig 3.2 - Pie Chart

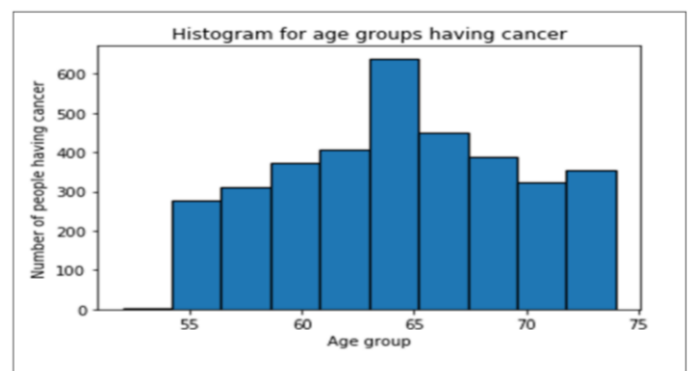
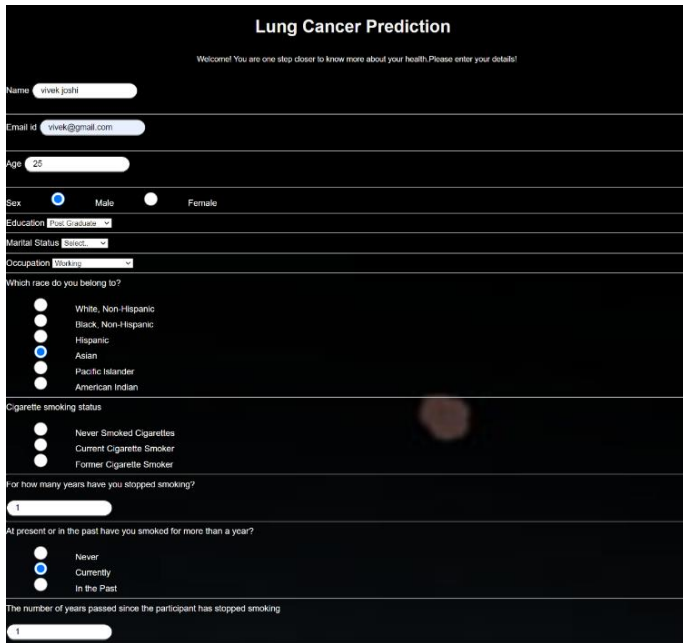
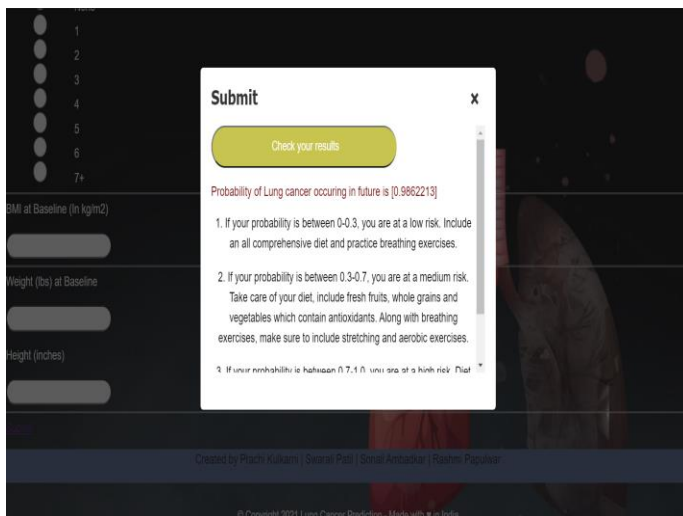


Fig 3.3 - Histogram for age groups having cancer



The screenshot shows a web form titled "Lung Cancer Prediction". It includes fields for Name (vivek joshi), Email id (vivek@gmail.com), Age (25), Sex (Male selected), Education (Post Graduate), Marital Status (Single), and Occupation (Writer). It also has radio buttons for race (Asian selected), cigarette smoking status (Former Cigarette Smoker selected), and years of smoking (1 selected). A "Submit" button is visible at the bottom.

### 3.3 - Questionnaire



The screenshot shows the prediction and prognosis results. A "Submit" dialog box is open, displaying the "Probability of Lung cancer occurring in future is [0.9862213]". Below this, there are three numbered risk levels: 1. Low risk (0-0.3), 2. Medium risk (0.3-0.7), and 3. High risk (0.7-1.0). The background shows a partially visible questionnaire form with fields for BMI, Weight, and Height.

Fig 3.5 - Prediction and Prognosis

### 4. CONCLUSION AND FUTURE SCOPE

Deep learning with the help of ANN and historical data can help in providing an approximate prediction to know whether a person lies within a risk of being diagnosed with lung cancer in the future which could help to increase awareness and make efficient use of deep learning for cancer prediction mechanisms.

The future perspective can be of expanding this system to use different forms of data attributes such as images and better deep learning algorithms for prediction. This system can even help in giving a better and refined view of different cancer causing factors to the people.

It also includes increasing the scope of the project to diagnose types of cancer other than lung cancer. It can increase awareness towards the factors of the people and help them in preventing a future chance of being diagnosed with lung cancer.

### 5. REFERENCES

- [1] Article on Cancer and Death Statistics. <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [2] Articles on Neural network models <https://otexts.com/fpp2/nnetar.html>
- [3] Article on Intuition of Adam Optimizer <https://www.geeksforgeeks.org/intuition-of-adam-optimizer/>
- [4] L. Dormehl, "Digital Trends", 5 1 2019. [Online]. Available: <https://www.digitaltrends.com/cool-tech/what-is-an-artificialneural-network/>. [Accessed 15 3 2019].
- [5] I. E. Liveris, K. Drakopoulou and P. Pintelas, "Predicting students' performance using artificial neural networks," in 8th PanHellenic Conference with International Participation Information and Communication Technologies in Education, Volos, Greece, 2012.