DAYANANDA SAGAR UNIVERSITY

KUDLU GATE, BANGALORE – 560068

Bachelor of Technology
in
COMPUTER SCIENCE AND ENGINEERING

Major Project Phase-II Report

LUNG CANCER ANALYSIS AND PREDICTION

By

Team 19:
Dheeraj D - ENG18CS0090
Donal Jovian Nazareth - ENG18CS0095
Gayathri Devi Nagalapuram - ENG18CS0105
Vansika Singh - ENG18CS0310
Varshashree D - ENG18CS0312

Under the supervision of

Dr. Savitha Hiremath
Associate Professor,
Department of Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
SCHOOL OF ENGINEERING
DAYANANDA SAGAR UNIVERSITY,
BANGALORE

(2021-2022)

DAYANANDA SAGAR UNIVERSITY

School of Engineering
Department of Computer Science & Engineering
Kudlu Gate, Bangalore – 560068
Karnataka, India

CERTIFICATE

This is to certify that the Phase-II project work titled **"LUNG CANCER ANALYSIS AND PREDICTION"** is carried out by **Dheeraj D (ENG18CS0090), Donal Jovian Nazareth (ENG18CS0095), Gayathri Devi Nagalapuram (ENG18CS0105), Vansika Singh (ENG18CS0310), Varshashree D (ENG18CS0312),** bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2021-2022**.

Dr. Savitha Hiremath	**Dr. Girisha G S**	**Dr. A Srinivas**
Associate Professor Dept. of CS&E, School of Engineering Dayananda Sagar University	Chairman CSE School of Engineering Dayananda Sagar University	Dean School of Engineering Dayananda Sagar University
Date:	Date:	Date:

Name of the Examiner **Signature of Examiner**

1.

2

DECLARATION

We, **Dheeraj D (ENG18CS0090), Donal Jovian Nazareth (ENG18CS0095), Gayathri Devi Nagalapuram (ENG18CS0105), Vansika Singh (ENG18CS0310), Varshashree D (ENG18CS0312),** are students of the eighth semester B.Tech in **Computer Science and Engineering**, at School of Engineering, **Dayananda Sagar University**, hereby declare that the phase-II project titled **"LUNG CANCER ANALYSIS AND PREDICTION"** has been carried out by us and submitted in partial fulfillment for the award of degree in **Bachelor of Technology in Computer Science and Engineering** during the academic year **2021-2022**.

Student **Signature**

Name: Dheeraj D
USN: ENG18CS0090

Name: Donal Jovian Nazareth
USN: ENG18CS0095

Name: Gayathri Devi Nagalapuram
USN: ENG18CS0105

Name: Vansika Singh
USN: ENG18CS0310

Name: Varshashree D
USN: ENG18CS0312

Place: Bangalore
Date

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to School of Engineering & Technology, Dayananda Sagar University for providing us with a great opportunity to pursue our Bachelor's degree in this institution.

We would like to thank ***Dr. A Srinivas, Dean***, ***School of Engineering & Technology***, ***Dayananda Sagar University*** *for his constant encouragement and expert advice.*

It is a matter of immense pleasure to express our sincere thanks to ***Dr. Girisha G S, Department Chairman***, ***Computer Science and Engineering***, ***Dayananda Sagar University,*** *for providing right academic guidance that made our task possible.*

We would like to thank our guide ***Dr. Savitha Hiremath***, ***Associate Professor***, ***Dept. of Computer Science and Engineering***, ***Dayananda Sagar University***, *for sparing her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

We would like to thank our ***Project Coordinators Dr. Meenakshi Malhotra and Dr. Bharanidharan N****, and all the staff members of Computer Science and Engineering for their support.*

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

Signature of Students

USN: ENG18CS0090, ENG18CS0095, ENG18CS0105,
ENG18CS0310, ENG18CS0312

Name: Dheeraj D, Donal Jovian Nazareth, Gayathri Devi Nagalapuram,
Vansika Singh, Varshashree D

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CNN	Convolution Neural Network
CT Scan	Compute Tomography Scan
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
3D	3 Dimensional
CMixNet	Customized Mixed Link Network
KNN	K-Nearest Neighbours
RFC	Random Forest Classifier
SVC	Support Vector Classifier
DTC	Decision Tree Classifier
LR	Linear Regression
RFR	Random Forest Regressor

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Lung cancer is one of the most common and deadly cancers worldwide. One of the most effective ways to fight cancer is to discover it early enough to improve the patient's chances of survival. The discovery of lung cancer at an early stage helps in reducing its risk. Various technologies like MRI, isotopes, X-rays, and CT scans are used for diagnosis of lung cancer. The studying of lung nodules helps a doctor to determine if the patient is malignant. These nodules sometimes have a chance of growing undetected by the naked eye.

In this project, Lung cancer is detected with the help of patient details, symptoms and CT scans by using Machine learning and Deep learning algorithms with open-source datasets. The proposed approach uses Machine learning algorithms to study past medical records and determine if the patient has lung cancer. Deep learning models are used to analyze the CT scans to determine the type of lung cancer. The major goal of this project is to find nodules as small as 3 mm to detect cancer stage accurately. Finally, the machine learning model calculates the patient's estimated medical insurance costs. All of these functionalities are combined and provided in the form of a web application. This project is useful for the early detection of lung cancer in individuals and can help them in overcoming these health conditions. The effectiveness of cancer prediction systems helps people to know their cancer risk in an expensive manner and it also helps the people to take the appropriate medical decision based on their cancer risk status.

CHAPTER 1

INTRODUCTION

CHAPTER 1 INTRODUCTION

Lung cancer is repeatedly identified as one of the deadliest diseases in the history of humankind. It is also one of the most frequent malignancies and one of the leading causes of mortality. According to the World Health Organization (WHO), lung cancer causes around 7.6 million deaths worldwide each year. The number of people affected by cancer is expected to continue to rise, reaching around 17 million by 2030. Early detection can aid in treatment.

1.1 BACKGROUND KNOWLEDGE

People who smoke are more likely to develop lung cancer. There are two types of lung cancer that are most common: non-small cell lung cancer and small cell lung cancer.

Smoking, secondhand smoke, exposure to specific chemicals, and a family history of the disease causes lung cancer. Coughing (frequently with blood), chest pain, wheezing, and weight loss are all symptoms. These symptoms aren't usually present until cancer has progressed.

A lung nodule (or mass) is a small abnormal area that is sometimes found during a CT scan of the chest. By studying nodules, a doctor can determine whether this scan is malignant. Humans can evaluate nodules larger than 7 mm in diameter, and doctors can sometimes have a patient wait to see if the nodule will develop or not, as it will be a harmless nodule if it does not. As a result, a nodule has a better chance of growing undetected by the naked eye.

1.2 EXISTING SYSTEM

There are many reasons behind cancer, ranging from behavioral traits such as high body mass index, tobacco and alcohol usage to physical carcinogens, such as exposure to ultraviolet rays and radiation, including certain biological and genetic carcinogens. However, the cause may vary from one patient to another. Common cancer symptoms are

pain, fatigue, nausea, persistent cough, breathing difficulties, weight loss, muscle pain, bleeding, bruising, and many more. Then again, neither of these symptoms are exclusive to cancer, nor are all of them apparent in every patient. As a result, it is hard to determine cancer without a thorough diagnostic procedure, such as Computed Tomography (CT) scan, Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) scan, ultrasound, or biopsy. Most times, the victims show little to no symptoms in the early stages, and when symptoms become apparent, more often than not, it is already too late.

Lung cancer is classified into two types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) (NSCLC). Carcinoid lung cancer is a third, less prevalent kind of lung cancer. There are no two lung cancers that are alike. Lung cancer is classified into several categories and subgroups. However, the focus of this work is solely on the detection of lung cancer based on previous patients records with symptoms, prediction of lung cancer using CT scans and lung nodule detection from CT scan.

1.3 PURPOSE

The purpose of this project is to investigate alternative strategies for identifying and detecting lung cancer in its early stages. As a result, treatment of lung cancer can take place before it reaches the point where treatment is impossible.

1.4 INTENDED AUDIENCE

This project will assist data scientists, machine learning engineers, development teams, and other computer engineering enthusiasts, who want to carry on working on the lung cancer analysis and prediction system.

1.5 INTENDED USE

The benchmarking results can be used by the intended users to redesign or enhance the existing lung cancer analysis and prediction system to meet their requirements.

CHAPTER 2

PROBLEM DEFINITION

CHAPTER 2 PROBLEM DEFINITION

2.1 MOTIVATION

Because of bloodstreams and lymph fluid present in lung tissue that migrates to the center of the chest, cancer cells spread quickly. Lung cancer must be discovered as early as possible because it has a tendency to spread and is incurable if it spreads too far.

CT scan produces accurate pictures of lung cancer diagnosis. One scan can provide up to 500 sections/slices, depending on the slice thickness. A single slice takes an experienced radiologist about 2–3.5 minutes to observe. When only one radiologist views the scan, lung cancer nodules are correctly diagnosed only 68 percent of the time. Radiologists face a challenging, time-consuming, and demanding task in detecting malignant lung nodules at an early stage.

Not just cancer, the lack of proper diagnosis leaves the people of developing and underdeveloped countries extremely vulnerable to other diseases as well. To overcome this challenge, these countries have to invest heavily in the public health sector. If we want to stay in the fight against cancer and give these people a realistic chance of survival, we have to look for alternative ways of diagnosis.

2.2 PROBLEM STATEMENT

Lung cancer is difficult to diagnose since symptoms only appear in the latter stages, and it is really hard to save a person's life at this stage. A single scan can provide up to 500 sections and it takes an experienced radiologist about 2–3.5 minutes to observe each section. Hence, a better mechanism is required.

The proposed project aims to develop a web application which can eye help in the initial stages of lung cancer diagnosis and can be scaled up further for other diseases using Deep

Learning. The proposed model can detect lung cancer at initial stage based on the patient's symptoms and can interpret CT(Computed tomography) images to identify nodules with diameters as tiny as 3mm which are unlikely to be identified by a radiologist with the naked eye.

2.3 OBJECTIVE

The objective of this project is:

- To build an approach to overcome the challenges faced with respect to lung cancer using Machine Learning and Deep Learning techniques to forecast the existence of cancer in the lungs using medical records, as well as interpret CT images to identify nodules with diameters as tiny as 3mm accurately with low cost in less time.
- To be able to make this approach easily accessible to all in the form of a web application.

CHAPTER 3

LITERATURE REVIEW

CHAPTER 3 LITERATURE REVIEW

DETAILS OF LITERATURE SURVEY

Cancer causes about one in six deaths every year [1][2] and lung cancer stands at the top of all of this as it is responsible for 1.76 million deaths up to 2016. Early detection of cancer can provide a suitable treatment to not just prolong life but also save a patient's life and hence increase the survival rate. [1][2][3][4]

The journal paper [1] published by Muthazhagan B, et al. (**2021**) states that with the aid of current lung cancer prediction technologies, predicting and detecting lung cancer at an early stage is a difficult challenge. An early lung tumor prediction might extend a person's life by one to five years. They created a Support Vector Machine based classification model which provided about 98% prediction accuracy in small amount of time. However, the images were merely classified into 'abnormal' or 'normal' and did not take into account the various stages [Stage 0 – stage IV] which is what this project aims to improve on.

The paper [2] proposed by Masud M, Sikder N, et al. (**2021)** uses a CNN based model for classifying the image into one of five kinds: colon adenocarcinomas, benign colonic tissues, lung adenocarcinomas, lung squamous cell carcinomas and benign lung tissues. While a peak accuracy of 96.33% has been achieved in the classification, the authors state that two out of five classes in can have much improved performance with further experimentation. The dataset used is Histopathological and Histopathology is the microscopic examination of a biopsy which is an invasive process. Our approach prefers to work on CT scans which is a non-invasive mechanism to detect cancer.

Sajja T, Devarapalli R, et al. (**2019)** [3] published a paper which worked on detecting lung cancer using the pre-trained CNN model called Google-Net. The deployed 60% of all neurons in the drop out layers to prevent overfitting and achieved a simplified and sparse network for classifying the CT images into benign or malignant. The model still requires testing on various dropout ratios to check for better performance accuracy. Our approach aims to construct a simplified CNN model to classify cancer along with providing medical information costs.

Tripathi P, Tyagi S, et al. (**2019)** [4] published a paper in which they attempt to detect lung cancer using four different segmentation techniques of image processing. They conclude that marker-controlled watershed segmentation provides the most accurate results. Through the comparative analysis, it is found that CT scans tend to provide the best chance at detecting cancer and should be the preferred means to do the same. Hence, we shall use Deep Learning on CT scans to classify the various stages.

Nasrullah Nasrullah et al. (**2019)** [5] study focuses on developing a model that can detect cancerous nodules using CT images. They opt to employ 3D CNN after some researching because of its proven performance in image analysis. To further identify the condition as benign or malignant, they use 3D MixNet to extract nodule features, which are then classified using Gradient Boosting Machine (GBM). The proposed model was validated using the free response receiver operating characteristic (FROC) evaluation matrix to obtain a FROC score of 94.21%. The suggested model outperformed all other models in terms of computational cost and desired output accuracy.

Siddharth Bhatia et al. (**2019)** [6] present a method for detecting lung cancer using deep residual learning. They offer a series of preprocessing strategies for extracting caner-vulnerable lung features using UNet and ResNet models. They examine the likelihood of predicting carcinogenic CT scans by comparing the effectiveness of classifiers such as Random forest and XGBoost. When the authors combine the two classifiers, they get the greatest accuracy of 84%. The constraint in this case is that the best achievable accuracy may have been higher.

Suren Makaju et al. (**2018)** [7] made a comparison of many probable cancer detection approaches and ranked them in order of effectiveness. They decide to upgrade that model to achieve even higher accuracy by selecting the current best approach from their survey of articles. The Median and Gaussian filters were used in the pre-processing stage, and the data was then segmented using the Watershed algorithm. They went on to use support vector machines to identify diagnosed cancerous nodules as benign or malignant. This upgraded model outperformed the previous best model by 5.4%, with an accuracy rate of 92 %. The model's sole flaw is that it does not differentiate between cancer stages (I to IV).

Ali I et al. **(2018)** [8] developed a deep learning algorithm that takes a CT image and perceives it as a collection of states, producing a classification of whether or not a malignant nodule is present. They employ a Reinforcement Learning algorithm that improves with time and with more data. Their research shows that the model's training data has a high accuracy of 99.1%, however the validation data has a low accuracy of 64.4 %. The model appears to be overfitted as a result of this. The authors suggest that because this is the only flaw, the constraint can be solved with more data.

Table 3.1 shows the summary of the extensive literature review described prior.

Table 3.1: Literature Survey

Sl No.	Author Name and Year	Title of Paper	Methodology	Limitations/ Conclusions
1	Muthazhagan B, Ravi T, Rajinigirinath D- **2021**	An enhanced computer-assisted lung cancer detection method using content-based image retrieval and data mining techniques [1]	Support Vector Machine image Classification algorithm	The malignancy is classified as 'Normal' and 'Abnormal', not as Stages 1-4
2	Masud M, Sikder N, et al. - **2021**	A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a	3 Digital Image Processing techniques with CNN	Dataset uses microscopic cells images rather

		Deep Learning-Based Classification Framework [2]		than CT/MRI scans
3	Sajja T, Devarapalli R, Kalluri H- **2019**	Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning [3]	A deep neural network based on Google-Net	Overfitted data causing the need for max dropout ratio
4	Tripathi P, Tyagi S, Nath M- **2019**	Comparative Analysis of Segmentation Techniques for Lung Cancer Detection [4]	comparative analysis- image segmentation techniques	marker-controlled watershed segmentation provides more accurate results.
5	Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H - **2019**	Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies [5]	Two 3D CNN with CMixNet architectures	3D CMixNet had better accuracy feature exploitation than other models compared with.

6	Bhatia S, Sinha Y, Goel L – **2019**	Lung Cancer Detection: A Deep Learning Approach [6]	deep residual networks with XGBoost and Random Forest classifiers and ensemble	The highest accuracy was 84% using an ensemble of both models tried which still a comparatively low accuracy
7	Makaju S, Prasad PW, et al. - **2018**	Lung Cancer Detection using CT Scan Images [7]	Watershed algorithm with SVM	Classification of different stages of cancer is not done
8	Ali I, Hart GR, et al. - **2018**	Lung Nodule Detection via Deep Reinforcement Learning [8]	Reinforcement learning algorithm	The model is overfit as training accuracy obtained was 99.1% whereas the testing accuracy was 64.4%

CHAPTER 4

PROJECT DESCRIPTION

CHAPTER 4 PROJECT DESCRIPTION

4.1 PROPOSED DESIGN

The primary goal of this project is to examine prior medical information in order to detect lung cancer, predict if the patient has lung cancer using the CT scans and then predict the stage of lung cancer by nodule detection. Additional aid is provided by anticipating medical insurance expenditures. The proposed design includes the following stages:

1. **Lung Cancer Detection based on symptoms:** The Random Forest Classifier is trained with a dataset containing previous patients' symptoms records and is used to classify if the patient has lung cancer or not.
2. **Lung Cancer Classification using CT Scans:** The Convolutional Neural Network is trained using the IQ-OTHNCCD lung cancer dataset. This CNN model classifies the CT scan into normal, benign or malignant case.
3. **Lung Cancer Nodule detection using Deep Learning:** UNET model trained using the LUNA 16 dataset for nodule detection. The mhd files in the dataset are converted to png files for training the UNet. Also, mask extraction is performed to train the CNN to identify nodules and variety of preprocessing steps are performed to segment out the ROI (the lungs) from the surrounding regions of bones and fatty tissues.
4. **Medical Insurance Cost Prediction:** The Random Forest Regressor is trained with a dataset to estimate a patient's insurance expenses if the patient has lung cancer.
5. **Data Analysis using Plotly:** The analysis web pages contain interactive graphs made by Plotly. The visualizations provide a better understanding of the datasets.
6. **Develop a Web Application: A** web application is created for better user interaction, which includes all the above stages. The application helps in real-time lung cancer diagnosis and medical insurance cost estimation.

The proposed project is a web application with the main web page comprising four buttons. The first button directs the user to the lung cancer detection based on symptoms, the second

button to the lung cancer classification based on CT scans page, and the third button takes the user to the lung cancer stage detection using deep learning. Finally, the fourth button directs the user to the medical insurance cost prediction. A flowchart of the proposed design is shown in Fig 4.1.

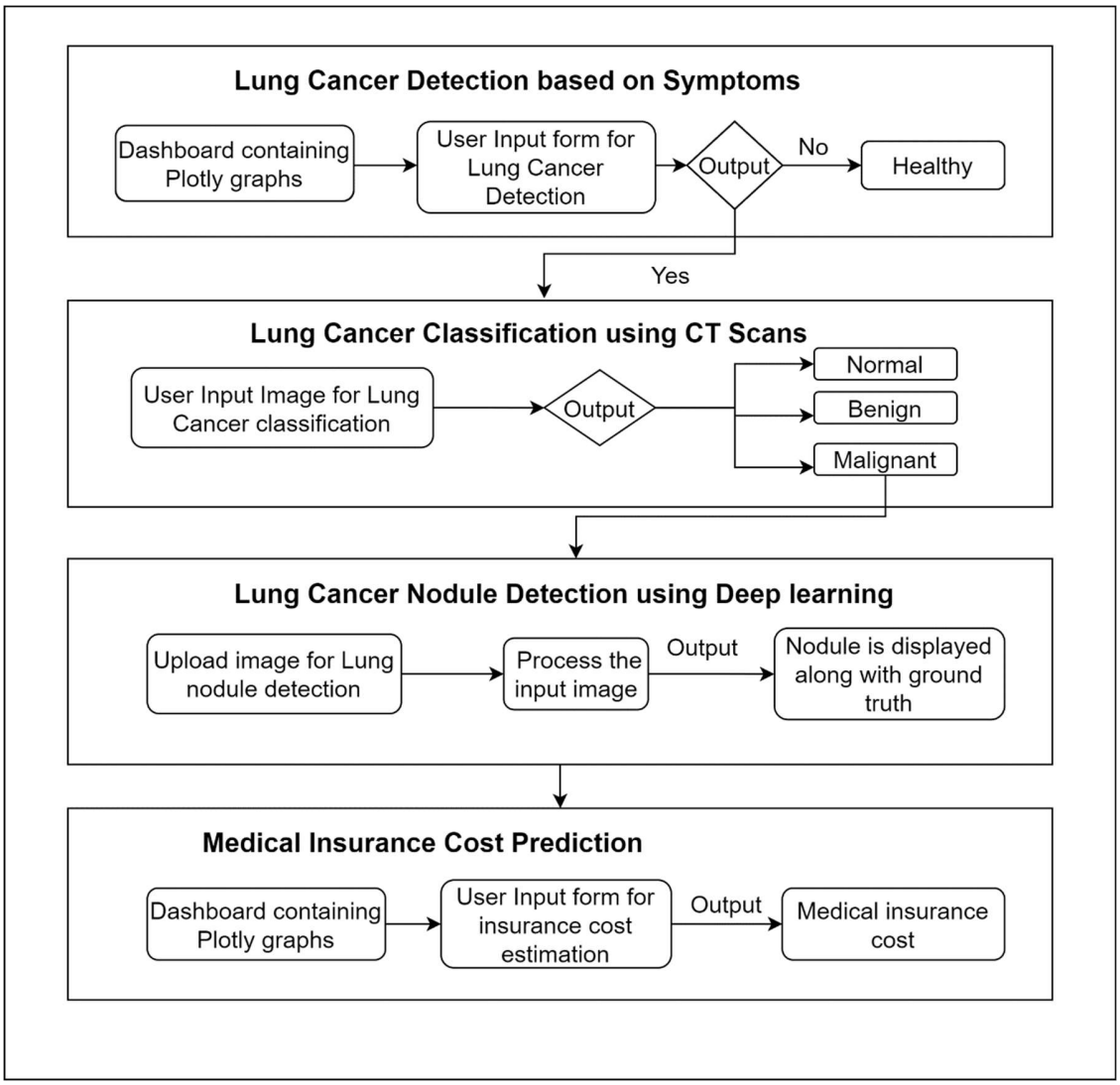


Figure 4.1 Flowchart of Proposed Design

The lung cancer detection based on symptoms page contains interactive graphs made by Plotly. The visualizations provide a better understanding of the datasets. The lung cancer prediction page comprises a form that takes user symptoms as input of patient symptoms. Here, The Random Forest Classifier is trained with a dataset containing previous patients' symptoms records and is used to classify whether or not the patient has lung cancer. If the

user does not have lung cancer, then it directs the user to a page displaying a message that the user is healthy. But, if the patient has lung cancer, then it directs the patient to lung cancer classification page.

The lung cancer classification using CT scans contains an image upload button where the patient can upload the image and then click submit button. The Convolutional Neural Network is trained with class weights using the IQ-OTHNCCD lung cancer dataset to perform this classification. If the output is malignant, it directs the user to the lung cancer stage detection page.

UNET model trained using the LUNA 16 dataset for nodule detection. The mhd files in the dataset are converted to png files for training the UNet. Also, mask extraction is performed to train the CNN to identify nodules and variety of preprocessing steps are performed to segment out the ROI (the lungs) from the surrounding regions of bones and fatty tissues. The output of UNet model is then passed to other defined ConvNet layers having activation as ReLU. The final output is then reshaped to 512X512. At last, we have defined the model that takes input (inp) and gives us the output (x_out) using the base_model.

The Medical insurance page contains plotly dashboard to provide insights about the dataset and the estimation form takes user details like age, gender, region as input and provides the estimated lung cancer treatment costs. Here, a dataset containing previous patients' details is used to train the Random Forest Regressor to predict the costs.

This proposed flask web application provides accurate results to all the users in less time effortlessly.

4.2 ASSUMPTIONS AND DEPENDENCIES

The datasets are used for lung cancer detection and insurance estimation by assuming:

- The inputs given by patients on the website are accurate
- The CT scans are obtained by examining the patients and are approved by doctors
- The radiologists have marked the nodules accurately in the CT scans
- No false inputs from users
- For insurance that the medical insurance costs are correct without any fraud entries by the hospitals or patients

CHAPTER 5

REQUIREMENTS

CHAPTER 5 REQUIREMENTS

5.1 FUNCTIONAL REQUIREMENTS

The Functional requirements for the proposed system are:

- The user can upload his age, height, weight, and other details for lung cancer prediction based on symptoms.
- A user's CT scan can be uploaded for lung cancer classification to know if the tumor is benign, malignant or no tumor.
- A CT scan of the user can be uploaded for lung nodule detection.
- The user can acquire a diagnosis and learn how advanced his/her lung cancer is.
- A cost estimate for the user's medical insurance is available.

5.2 NON - FUNCTIONAL REQUIREMENTS

5.2.1 Usability:

It should be simple to use and navigate through the end-to-end user application (website). Patients and medical professionals should find it useful.

5.2.2 Reliability:

Lung cancer detection and classification should be precise. Any error could result in the patient's death.

5.2.3 Performance:

Any new dataset should have a relatively short training time. The classification must be correct; hence the margin of error must be small.

5.2.4 Supportability:

The website should function properly in the most recent browsers (Google Chrome/Mozilla Firefox).

5.2.5 Data set requirements:

The better the training and testing phases may be, the more CT scans and medical data there are available.

The Non-Functional requirements for the proposed system are:

- The system uses the patient's information to anticipate lung cancer.
- The system classifies the CT scan images into normal, benign and malignant.
- The system reads the CT scan and performs image processing.
- To detect lung cancer stage, the system runs the preprocessed images through a nodule classifier and then detects the largest malignant nodule.
- The system calculates the expenses based on the insurance information.

CHAPTER 6
METHODOLOGY

CHAPTER 6 METHODOLOGY

The project has different modules implemented with different methodologies as shown in Fig 6.1.

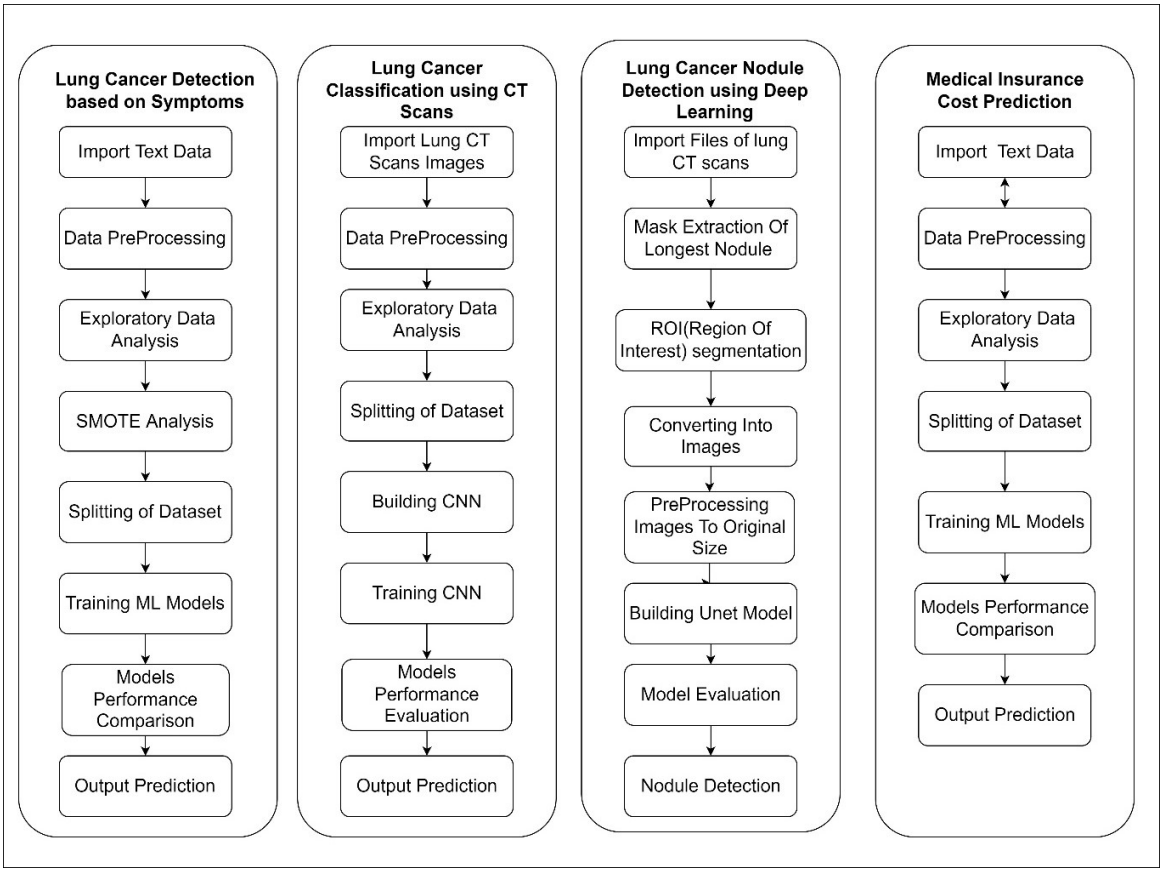


Figure 6.1 Flowchart of implementation

6.1 LUNG CANCER DETECTION BASED ON SYMPTOMS

Various symptoms and habitual practices can lead to lung cancer. Using such data from users, we build models to predict if or not a patient has lung cancer.

6.1.1 Data Analysis

A thorough data analysis to be able to see the kind of data we are dealing with. The dataset used was the Survey Lung Cancer dataset collected from data.world website. The data is collected from the website online lung cancer prediction system and gets feedback from

the user. This site was implemented during the period of August 2013 by the people who visited this site. The data has 309 records with 16 variable columns as shown in Fig 6.2.

	304	305	306	307	308
GENDER	F	M	M	M	M
AGE	56	70	58	67	62
SMOKING	1	2	2	2	1
YELLOW_FINGERS	1	1	1	1	1
ANXIETY	1	1	1	2	1
PEER_PRESSURE	2	1	1	1	2
CHRONIC DISEASE	2	1	1	1	1
FATIGUE	2	2	1	2	2
ALLERGY	1	2	2	2	2
WHEEZING	1	2	2	1	2
ALCOHOL CONSUMING	2	2	2	2	2
COUGHING	2	2	2	2	1
SHORTNESS OF BREATH	2	2	1	2	1
SWALLOWING DIFFICULTY	2	1	1	1	2
CHEST PAIN	1	2	2	2	1
LUNG_CANCER	YES	YES	YES	YES	YES

Figure 6.2 Preview of the Dataset

6.1.2 Data processing

Certain modifications which are to be done on the data are done in this stage. Renaming columns and encoding categorical data were some changes. Since the dataset was imbalanced, as seen in Fig 6.3, resampling was used to balance it. Synthetic Minority Oversampling Technique (SMOTE) was implemented as an oversampling technique to increase the number of cases in the dataset in a balanced way.

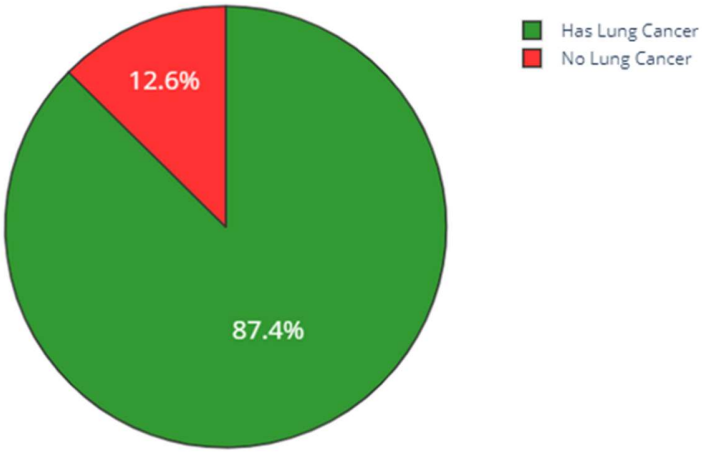


Figure 6.3 Distribution of target variable

6.1.3 Data exploration

To understand the correlation of various variables. Fig 6.4 shows the Feature Correlation to the target variable Lung_Cancer, and Fig 6.5 shows the Correlation Matrix.

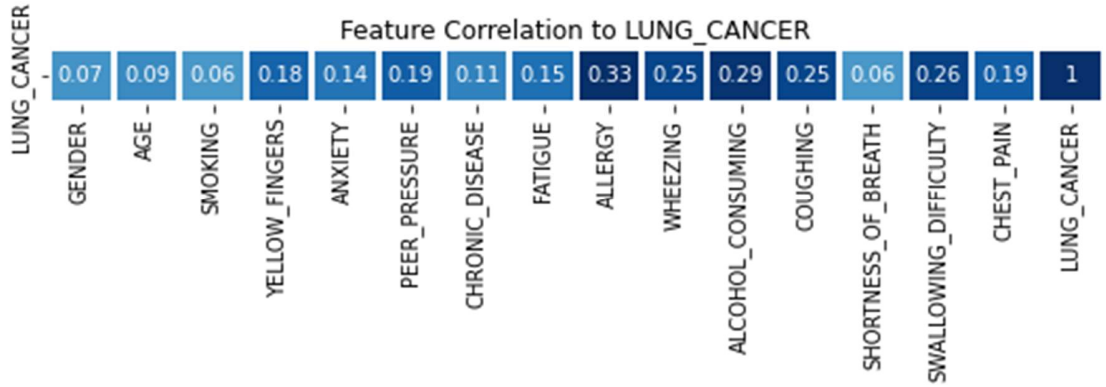


Figure 6.4 Features Correlation to target variable

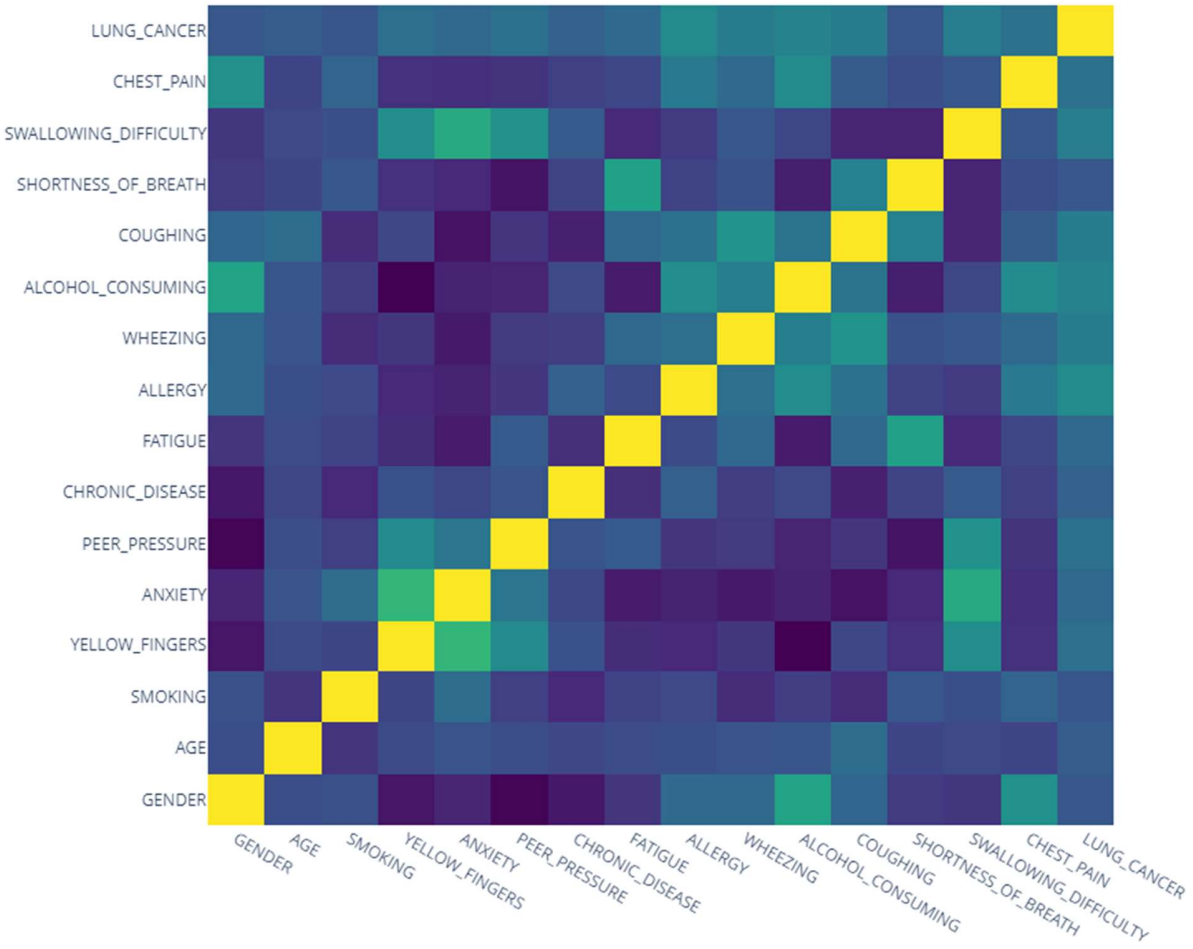


Figure 6.5 Features Correlation Matrix

Certain conclusions are drawn to help us in further comprehension, such as:

- Count of males are more than females
- Mean age of males is 0.4 more than females
- Males smoke more than females
- Yellow fingers are more common among females
- Anxiety is also commonly found symptoms in females
- Peer pressure is also more for females
- Chronic disease is also more for females

Further exploration was done with respect to all variables. The obtained results were used to create a dashboard page, shown in the following chapters.

6.1.4 Model Building

The creation of different samples for training and testing helps us evaluate model performance. Hence the split of our modelling dataset into training and testing samples is performed using the train_test_split() function of the scikit-learn library.

Following the data split, the train data is fed to various models in order to train them. The models used are:

- KNN: K-Nearest Neighbors
- RFC: Random Forest Classifier
- SVC: Support Vector Classifier
- DTC: Decision Tree Classifier

After training the models, they are made to predict. The predicted values are matched against the validation data to obtain accuracies of each of the models. The model with the highest accuracy was the RFC, which was chosen as the best model to predict the presence of cancer based on user's symptoms.

6.2 MEDICAL INSURANCE COST PREDICTION

The insurance dataset is used to build a model to be able to predict an approximate cost of insurance that a cancer patient will be in need for. The prediction will be based on the given user's details.

6.2.1 Data Analysis

A thorough data analysis to be able to see the kind of data we are dealing with. The dataset used was the Medical Insurance Dataset collected from Kaggle. The data has 1338 records with 7 variable columns as shown in Fig 6.6.

	1333	1334	1335	1336	1337
age	50	18	18	21	61
sex	male	female	female	female	female
bmi	30.97	31.92	36.85	25.8	29.07
children	3	0	0	0	0
smoker	no	no	no	no	yes
region	northwest	northeast	southeast	southwest	northwest
charges	10600.5483	2205.9808	1629.8335	2007.945	29141.3603

Figure 6.6 Preview of Insurance dataset

6.2.2 Data processing

Certain modifications which are to be done on the data are done in this stage. Encoding categorical data was done.

6.2.3 Data exploration

Gaining insights by observing the spread of data, Fig 6.8 shows a part of the scatterplot. To understand the correlation of various variables, Fig 6.7 shows the Correlation Matrix.

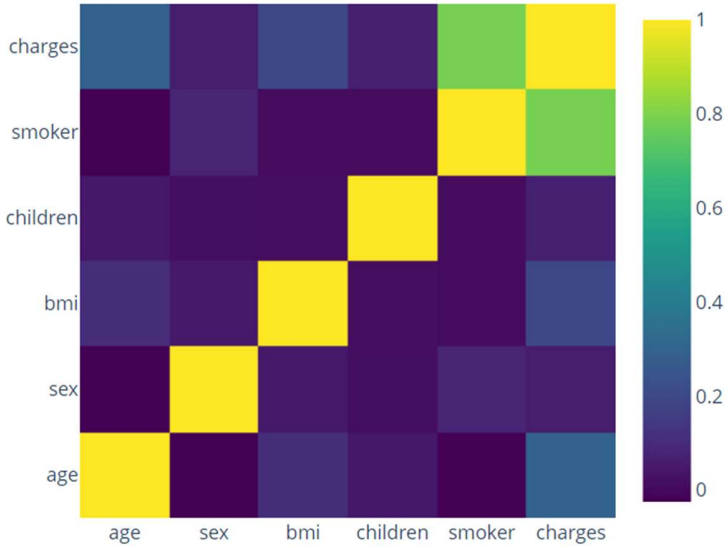


Figure 6.7 Correlation Matrix of Insurance dataset

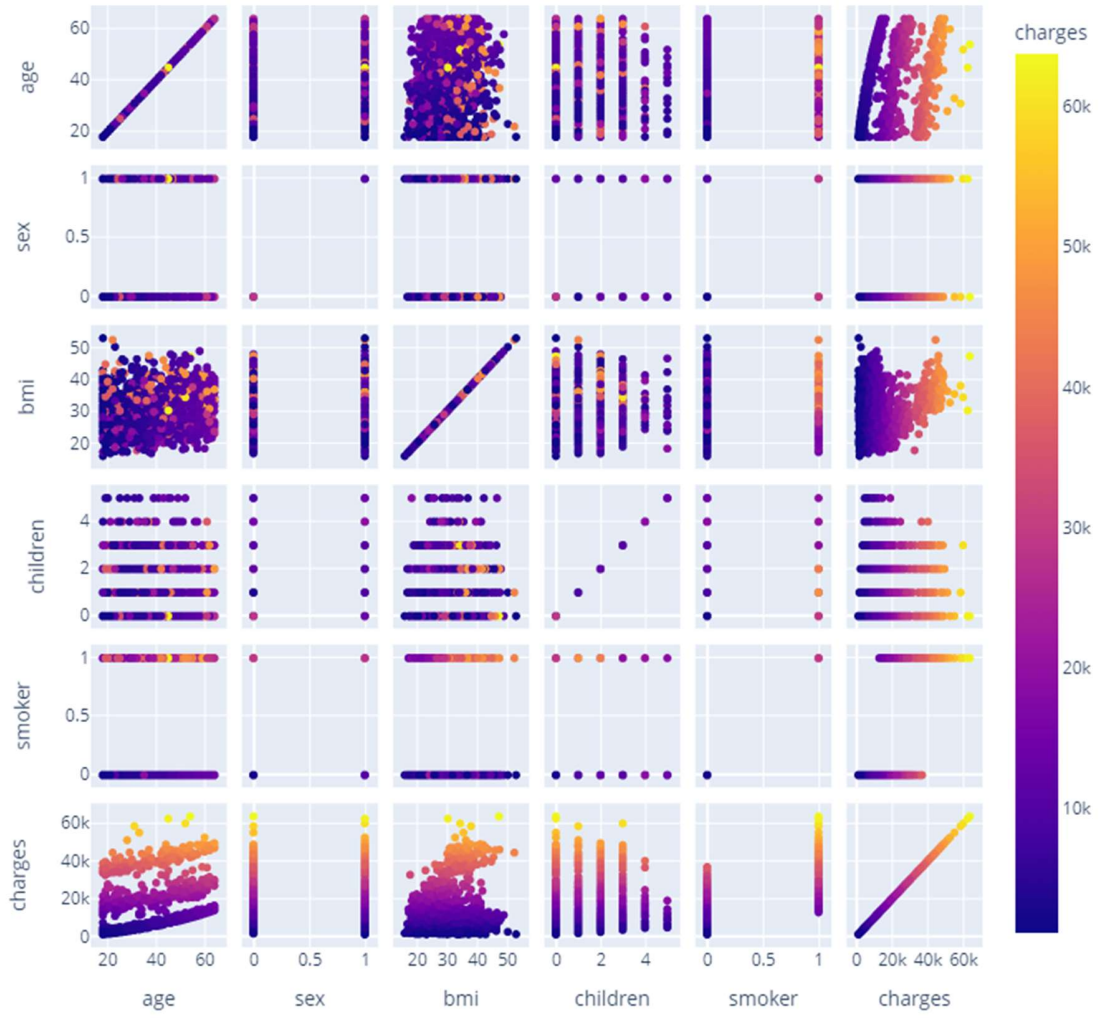


Figure 6.8 Scatter Plot of Insurance dataset

6.2.4 Model Building

The creation of different samples for training and testing helps us evaluate model performance. Hence the split of our modelling dataset into training and testing samples is performed using the train_test_split() function of the scikit-learn library. With about 1070 samples for training and 268 for testing, models were implemented.

Following the data split, the train data is fed to various models in order to train them. The models used are:

- Linear Regression(LR)
- Random Forest Regression (RFR)

- Decision Tree Regression
- Lasso Regression

The models are built to predict after they have been trained. The predicted values are compared to the validation data to determine the accuracy of each model. The RFR model had the highest accuracy and was chosen as the best model to forecast insurance rates based on user information.

6.3 LUNG CANCER CLASSIFICATION USING CT SCANS

6.3.1 Data Analysis

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) lung cancer dataset was collected in the above-mentioned specialist hospitals over a period of three months in fall 2019. It includes CT scans of patients diagnosed with lung cancer in different stages, as well as healthy subjects. IQ-OTH/NCCD slides were marked by oncologists and radiologists in these two centers. The dataset contains a total of 1190 images representing CT scan slices of 110 cases. These cases are grouped into three classes: normal, benign, and malignant. of these, 40 cases are diagnosed as malignant; 15 cases diagnosed with benign; and 55 cases classified as normal cases. The 110 cases vary in gender, age, educational attainment, area of residence and living status. Fig 6.9 shows the images of the three cases below.

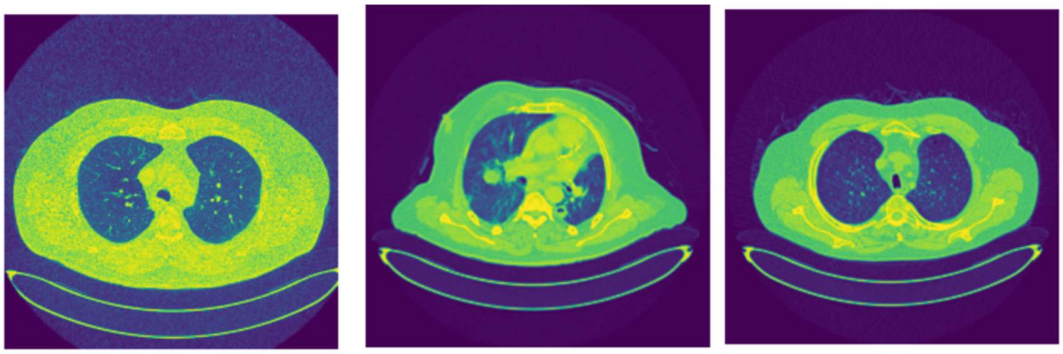


Figure 6.9 Benign, Malignant and normal cases respectively

6.3.2 Data processing and exploration

The image data is shuffled then viewed using colormaps like Fig 6.10 that uses hot cmap. The data is normalized, reshaped and encoded using the one hot encoding.

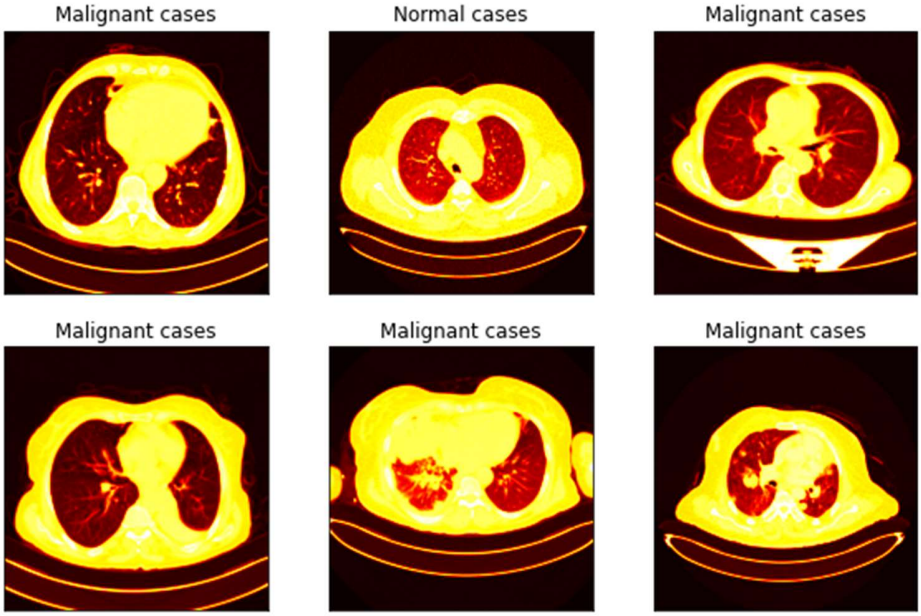


Figure 6.10 Image data with hot colormap

6.3.3 Model Building

The data is then split into two: Training and Testing set using the train_test_split() function. The split of data has images from all three classes in both the sets.

A Convolutional Neural Network to predict the correct classes of cells from the images was created. We have used 3 Conv2D layers with MaxPool2D layers after each for the feature extraction from the images. The activation function used is ReLU. The output layer has only three neurons corresponding to the three classes of tumors (Benign, Malignant, Normal), with SoftMax activation function. Fig 6.11 shows the model summary.

Then, the model is compiled with Adam as the optimizer and Categorical Crossentropy as the loss function. We will train the model for 7 epochs with the class weights. The training was stopped at the 7^{th} epoch as a good accuracy was obtained of about 92%.

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 254, 254, 32)      320

 max_pooling2d (MaxPooling2D  (None, 127, 127, 32)     0
 )

 conv2d_1 (Conv2D)           (None, 125, 125, 64)      18496

 max_pooling2d_1 (MaxPooling  (None, 62, 62, 64)       0
 2D)

 conv2d_2 (Conv2D)           (None, 60, 60, 128)       73856

 max_pooling2d_2 (MaxPooling  (None, 30, 30, 128)      0
 2D)

 flatten (Flatten)           (None, 115200)            0

 dense (Dense)               (None, 256)               29491456

 dense_1 (Dense)             (None, 3)                 771

=================================================================
Total params: 29,584,899
Trainable params: 29,584,899
Non-trainable params: 0
_________________________________________________________________
```

Figure 6.11 CNN Model Summary

6.4 LUNG CANCER NODULE DETECTION USING DEEP LEARNING

6.4.1 Data Analysis

The dataset excluded scans with a slice thickness greater than 2.5 mm. In total, 888 CT scans are included. The LIDC/IDRI database also contains annotations which were collected during a two-phase annotation process using 4 experienced radiologists. Each radiologist marked lesions they identified as non-nodule, nodule < 3 mm, and nodules >=

3 mm. See this publication for the details of the annotation process. The reference standard of challenge consists of all nodules >= 3 mm accepted by at least 3 out of 4 radiologists. Annotations that are not included in the reference standard (non-nodules, nodules < 3 mm, and nodules annotated by only 1 or 2 radiologists) are referred as irrelevant findings. The subset0 is a zip file which contains all CT images and annotations.csv file contains the annotations used as reference standard for the 'nodule detection' track.

In subset0, CT images are stored in MetaImage (mhd/raw) format. Each mhd file is stored with a separate raw binary file for the pixel data. The annotation file is a csv file that contains one finding per line. Each line holds the SeriesInstanceUID of the scan, the x, y, and z position of each finding in world coordinates and the corresponding diameter in mm. The annotation file contains 1186 nodules.

6.4.2 Data processing and exploration

Before inputting the CT images into the U-net architecture, it is important to reduce the domain size for more accurate results. A variety of preprocessing steps are performed to segment out the ROI (the lungs) from the surrounding regions of bones and fatty tissues. These include

- Binary Thresholding
- Selecting the two largest connected regions
- Erosion to separate nodules attached to blood vessels
- Dilation to keep nodules attached to the lung walls
- Filling holes by dilation
- Converting the mhd files to png

An index of the image file is captured and the directory of that index image is stored. After that, the image is opened, resized and converted into a grayscale image and its index is stored. The mask of the image corresponding to the index is also stored along with the grayscale image. After that, masks can be read by giving the directory of the mask. Finally, the mask image is preprocessed by resizing it and normalizing the pixel value then stored

at the pre-processed mask image at the output array at the same index position. X[n] stores the image and y[n] stores the corresponding mask.

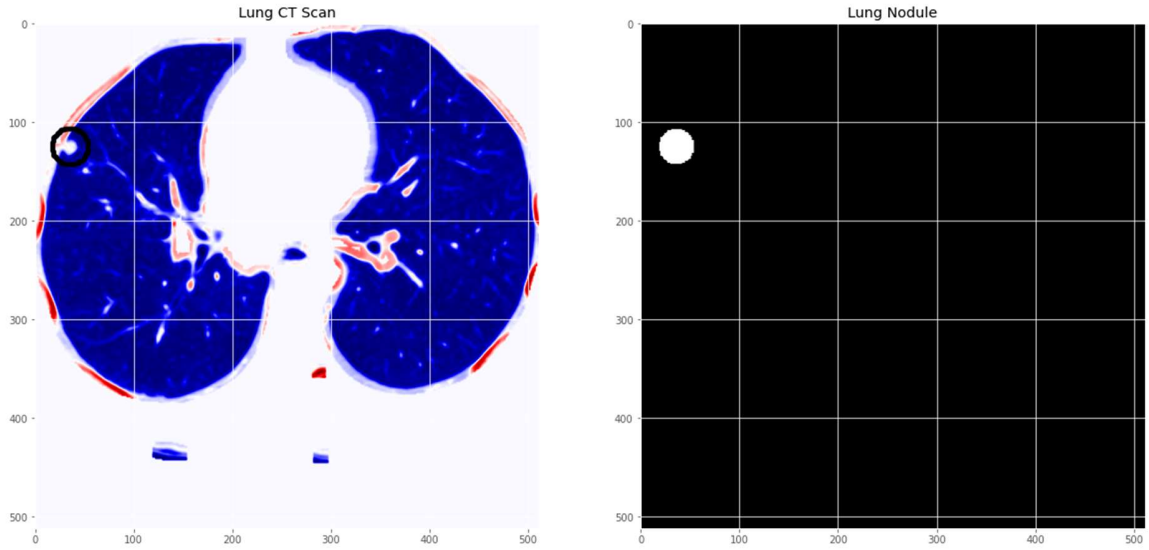


Figure 6.12 CT Scan and its corresponding lung nodule mask visualization

The plot of the pre-processed sample image and mask of the image displaying the nodule is shown above in Fig 6.12. The above image shows the pre-processed image as well as its mask.

6.4.3 Model Building

An input shape is defined as 512x512x1 that is expected by the base model and the custom layer that takes that base mode input whose output is then passed to the UNet model. The output of UNet model is then passed to other defined ConvNet layers having activation as ReLU. The final output is then reshaped to 512X512. At last, we have defined the model that takes input (inp) and gives us the output (x_out) using the base_model.

Then, the model is compiled with Adam as the optimizer and Binary Crossentropy as the loss function. We will train the model for 3 epochs with the class weights. The training was stopped at the 10^{th} epoch as a good accuracy was obtained of about 98%.

CHAPTER 7

EXPERIMENTATION

CHAPTER 7 EXPERIMENTATION

7.1 LUNG CANCER DETECTION BASED ON SYMPTOMS

Various models as previously discussed were used and the best of them was chosen to be used. The various models used their train and test accuracies are provided below in Table 7.1.

Table 7.1 Accuracy of models

Model	Train Accuracy	Test Accuracy
RFC	98.020312	0.969136
SVC	96.296296	0.969136
DTC	98.876574	0.919753
KNN	94.444444	0.907407

Random Forest Classifier has yielded the best performance.

7.2 MEDICAL INSURANCE COST PREDICTION

Various models as previously discussed were used and the best of them was chosen to be used. The various models used their train and test accuracies are provided below in Table 7.2.

Table 7.2 Accuracy of models

Model	Train Accuracy	Test Accuracy
Random Forest Regression	88.9149	86.2906
Decision Tree Regression	88.0294	86.1846
Lasso Regression	74.6111	76.1826
Linear Regression	74.6111	76.1825

Random Forest Regressor has yielded the best performance.

7.3 LUNG CANCER CLASSIFICATION USING CT SCANS

On using built CNN model to predict a random image, the following output was given. The first array is the prediction values of possibility of each of the classes. The next value is the class of the argument in the array with the maximum value. The third value is the category label of the class. The image was predicted rightly as shown in Fig 7.1.

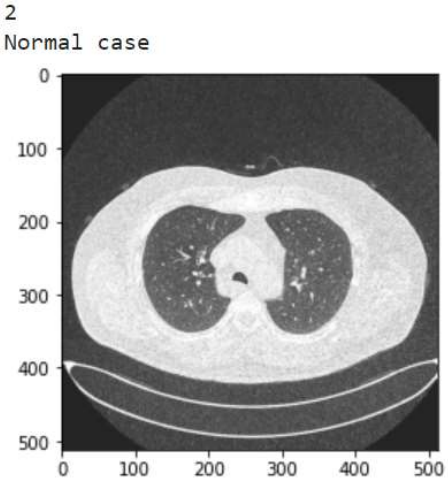


Figure 7.1 Prediction of randomly chosen image

On further experimentation, a data frame with predicted value and its equality to its actual value is created. Fig 7.2 shows a preview of the data frame.

	File	Actual	Path	Predicted	Equal
0	/content/drive/MyDrive/LungCancer/splitted_dat...	0	/content/drive/MyDrive/LungCancer/splitted_dat...	0	True
1	/content/drive/MyDrive/LungCancer/splitted_dat...	0	/content/drive/MyDrive/LungCancer/splitted_dat...	0	True
2	/content/drive/MyDrive/LungCancer/splitted_dat...	0	/content/drive/MyDrive/LungCancer/splitted_dat...	0	True
3	/content/drive/MyDrive/LungCancer/splitted_dat...	0	/content/drive/MyDrive/LungCancer/splitted_dat...	0	True
4	/content/drive/MyDrive/LungCancer/splitted_dat...	0	/content/drive/MyDrive/LungCancer/splitted_dat...	0	True

Figure 7.2 Predicted vs Actual value data frame

7.4 LUNG CANCER NODULE DETECTION USING DEEP LEARNING

A learning curve is a plot of model learning performance over experience or time. Learning curves are a widely used diagnostic tool in machine learning for algorithms that learn from a training dataset incrementally. The model can be evaluated on the training dataset and on a holdout validation dataset after each update during training and plots of the measured performance can created to show learning curves.

Reviewing learning curves of models during training can be used to diagnose problems with learning, such as an underfit or overfit model, as well as whether the training and validation datasets are suitably representative. Fig 7.3 shows the learning curve for the lung cancer nodule detection. The curve demonstrates a case of a good fit.

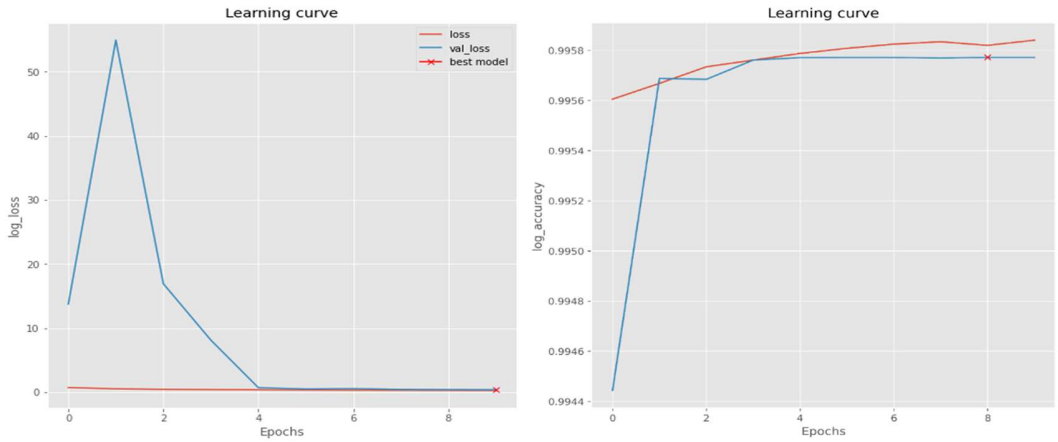


Figure 7.3 The learning curve for lung cancer nodule detection

CHAPTER 8

TESTING AND RESULTS

CHAPTER 8 TESTING AND RESULTS

8.1 LUNG CANCER PREDICTION ON THE BASIS OF SYMPTOMS

A prediction with values from dataset to test if prediction is right or wrong provided 100% correct results. Fig 8.1 shows the result of the predictions.

```
[58] prediction = rfc.predict([[0,63,1,2,1,1,1,1,1,2,1,2,2,1,1]])
     print(prediction)

     [0]

     prediction = rfc.predict([[0,59,1,1,1,2,1,2,1,2,1,2,2,1,2]])
     print(prediction)

     [1]
```

Figure 8.1 Results for Symptom based predictions

The models used provided variable scores but RFC and SVC proved to be the best as shown in Fig 8.2.

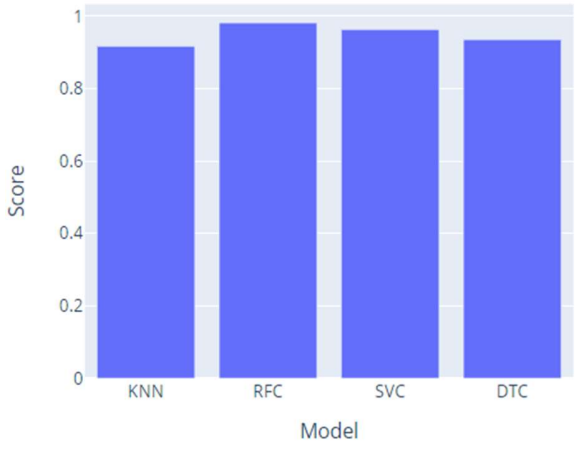


Figure 8.2 Scores of all models for Lung Cancer Prediction

8.2 INSURANCE PREDICTION

A prediction with values from dataset to test if prediction is right or wrong provided 100% correct results. Fig 8.3 shows the result of the predictions.

```
prediction = rfr.predict([[28,1,33,3,0]])
print(prediction)
```

```
[6332.10105162]
```

```
result = prediction * (75.62)
res = result[0]
round(res,2)
```

```
478833.48
```

Figure 8.3 Results for Insurance cost prediction

The models used provided variable scores but RFR and DTR proved to be the best as shown in Fig 8.4.

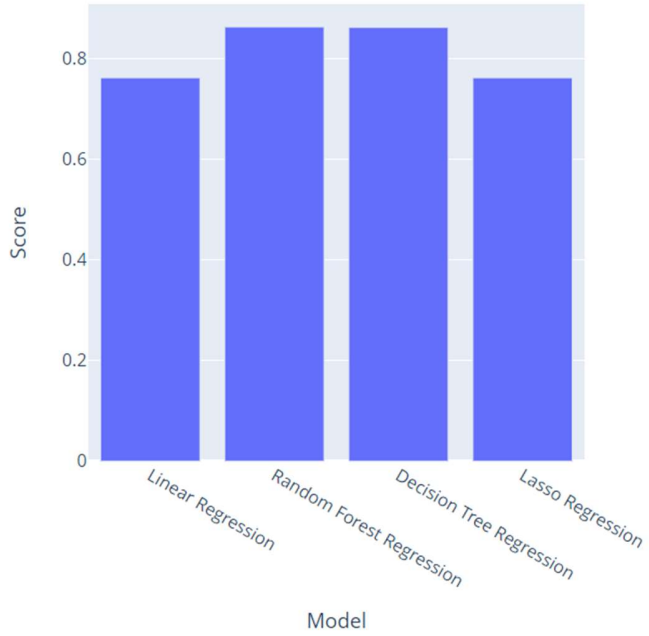


Figure 8.4 Scores of all models for Insurance cost prediction

8.3 LUNG CANCER CLASSIFICATION

If the accuracy is high and the loss is low, then the model makes small errors on just some of the data, which would be the ideal case. The following plots in Fig 8.5 show the loss at and accuracy of the CNN model on both training and testing sets.

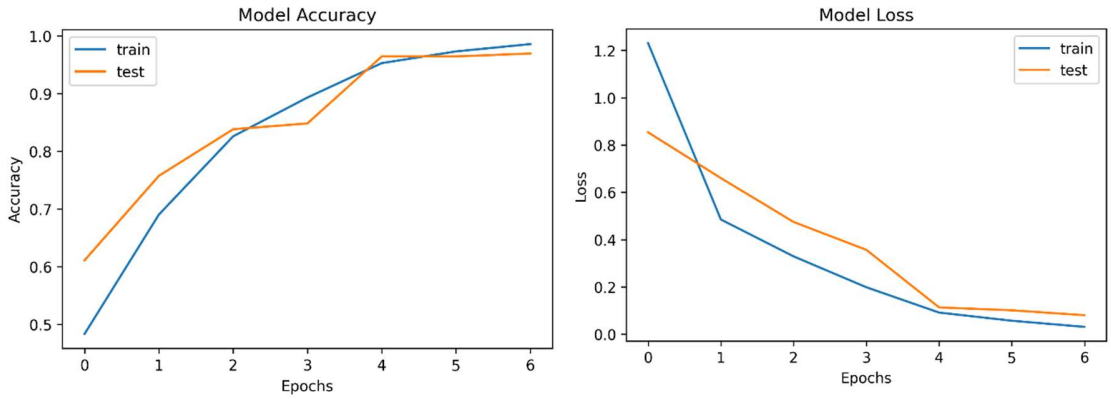


Figure 8.5 Learning curve of CNN Model

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model. 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. Fig 8.6 shows the confusion matrix of the CNN Model.

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, and F1 Score of your trained classification model. Fig 8.7 shows the trained CNN model's classification report.

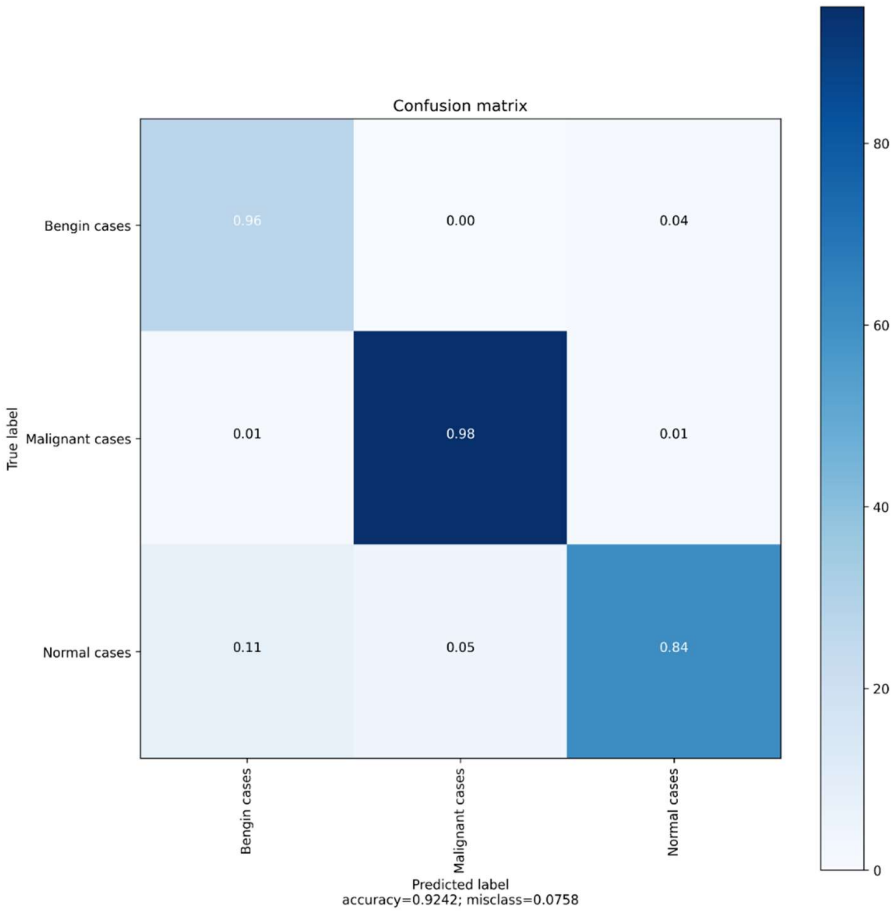


Figure 8.6 CNN Model Confusion Matrix

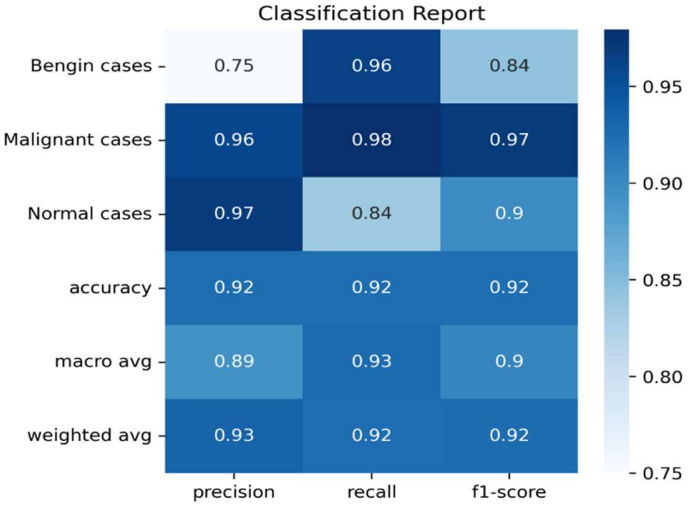


Figure 8.7 CNN Model Classification Report

8.4 LUNG CANCER NODULE DETECTION USING DEEP LEARNING

Fig 8.8 shows the output image that includes the Lung CT Scan, Ground Truth, Predicted Nodule and Predicted Nodule Binary.

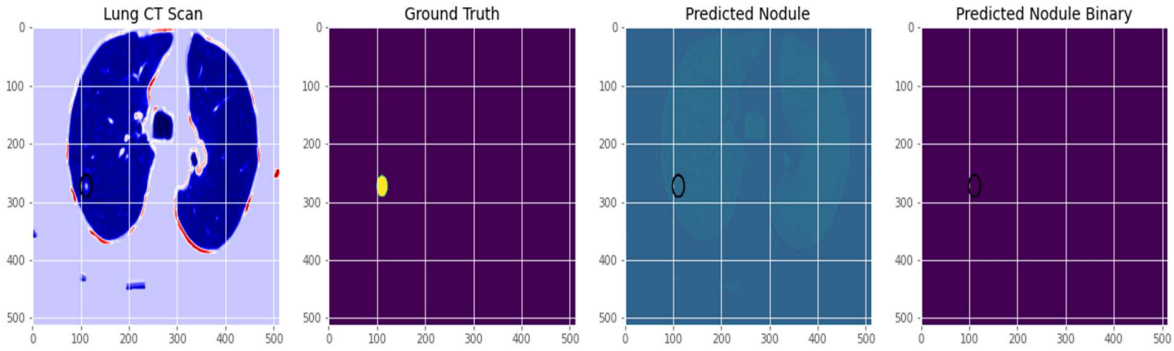


Figure 8.8 Output image of detected nodule

8.5 COMPARISON OF PROPOSED MODELS

The accuracy of the proposed model is compared with different models from previously done experiments as shown in Table 8.1. The proposed models outperformed the others in terms of performance. The proposed UNet with an accuracy of 98% outperformed CNN and ResNet models for nodule detection. The classification of CT Scans had accuracy upto about 90% in previous done research which was also outperformed with our proposed CNN model by achieveing an accuracy of 92.42%. Similarly the proposed RFC and RFR gave comparatively better results.

Table 8.1 Comparison of Proposed Models

Modules	Ref	Model Used	Accuracy Achieved
Lung Cancer Nodule Detection	[19]	RestNet50 + SVM RBF	93.19
	[18]	CNN	92.63
	[20]	DFCNet	89.52
	[8]	CNN	64.40
	[10]	SVM-LASSO	84.60
		Proposed Unet	98.00
Lung cancer classification using CT scans	[21]	3D MixNet	88.83
	[21]	3D MixNet + GBM	90.57
	[6]	Ensemble of UNet+ Random Forest and ResNet+ XGBoost	84.00
	[7]	SVM	86.67
	[22]	fine-tuning Conv4 of AlexNet	85.21
		Proposed CNN	92.42
Lung cancer prediction on the basis of symptoms	[9]	ANN	96.67
		Proposed RFC	96.91
	[23]	Stochastic Gradient Boosting	85.82
Medical Insurance Cost Prediction	[23]	XGBoost	85.36
	[24]	RFR	85.00
	[25]	Stochastic Gradient Boosting	86
		Proposed RFR	86.29

CHAPTER 9

CONCLUSION

CHAPTER 9 CONCLUSION

The goal of this project was to create a method for overcoming the challenges of lung cancer by utilizing Machine Learning and Deep Learning techniques to predict the presence of cancer in lungs using medical records, as well as interpret CT images to accurately identify nodules with diameters as small as 3mm at a low cost and in less time. Along with that, to be able to make this technology available to everyone in the form of a web application.

All of the objectives were met, as evidenced by the following outcomes. Random Forest Classifier with 96.9% accuracy for symptom-based recognition and Random Forest Regressor with 86.3 percent accuracy for predicting medical insurance costs. The accuracy of the CNN model used to analyze CT images was 92.42 percent. Finally, the UNet model designed to detect nodules on CT scans performed excellently, with a 98% accuracy rate. The developed strategy is, in general, highly dependable for users.

CHAPTER 10

FUTURE WORK

CHAPTER 10 FUTURE WORK

Further research and studies are to be conducted and validation of the proposed models of convolutional neural networks has to be performed. Verification of the presented models is necessary before they may be used in lung cancer screening procedures, enhancing the detection rate at an earlier stage. Models can also be enhanced by training with additional data from a wider range of scenarios. Also, consider employing multi-segmentation models with a high number of processors to train models to detect additional nodules. This can help train and predict output in less time. More research and trials exploiting technology developments are needed, and clinicians must rise to the task of improvising and implementing them.

REFERENCES

[1] Muthazhagan B, Ravi T, Rajinigirinath D, "An enhanced computer-assisted lung cancer detection method using content-based image retrieval and data mining techniques", *Journal of Ambient Intelligence and Humanized Computing,* 2:1-9, 2020.

[2] Masud M, Sikder N, Nahid AA, Bairagi AK, AlZain MA, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework", *Sensors*, 21(3):748, 2021.

[3] Sajja T, Devarapalli R, Kalluri H, "Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning", *Traitement du Signal*, 36(4):339-44, 2019.

[4] Tripathi P, Tyagi S, Nath M, "A comparative analysis of segmentation techniques for lung cancer detection", *Pattern Recognition and Image Analysis*, 167-73, 2019.

[5] Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H, "Automated lung nodule detection and classification using deep learning combined with multiple strategies", *Sensors*, 19(17):3722, 2019.

[6] Bhatia S, Sinha Y, Goel L, "Lung cancer detection: a deep learning approach. InSoft Computing for Problem Solving", *Springer*, (pp. 699-705), 2019.

[7] Makaju S, Prasad PW, Alsadoon A, Singh AK, Elchouemi A, "Lung cancer detection using CT scan images", *Procedia Computer Science*, 1;125:107-14, 2018.

[8] Ali I, Hart GR, Gunabushanam G, Liang Y, Muhammad W, Nartowt B, Kane M, Ma X, Deng J, "Lung nodule detection via deep reinforcement learning", *Frontiers in oncology*, 16;8:108, 2018.

[9] Nasser IM, Abu-Naser SS, "Lung cancer detection using artificial neural network.", *International Journal of Engineering and Information Systems (IJEAIS),* Mar;3(3):17-23, 2019.

[10] Choi W, Oh JH, Riyahi S, Liu CJ, Jiang F, Chen W, White C, Rimner A, Mechalakos JG, Deasy JO, Lu W, "Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer", *Medical physics*, Apr;45(4):1537-49, 2018.

[11] Kadir, Timor, and Fergus Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques.", *Translational lung cancer research*, 7.3: 304, 2018.

[12] Raoof, Syed Saba, M. A. Jabbar, and Syed Aley Fathima, "Lung Cancer prediction using machine learning: A comprehensive approach.", *2nd International conference on innovative mechanisms for industry applications (ICIMIA). IEEE*, 2020.

[13] Xie, Ying, et al, "Early lung cancer diagnostic biomarker discovery by machine learning methods.", *Translational oncology*, 14.1: 100907, 2021.

[14] Singh, Gur Amrit Pal, and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans.", *Neural Computing and Applications,* 31.10: 6863-6877, 2019.

[15] Shin, Hyunku, et al, "Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes.", *ACS nano*, 14.5: 5435-5444, 2020.

[16] Hosny, Ahmed, et al, "Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study.", *PLoS medicine,* 15.11: e1002711, 2018.

[17] Lakshmanaprabu, S. K., et al. "Optimal deep learning model for classification of lung cancer on CT images.", *Future Generation Computer Systems,* 92: 374-382, 2019.

[18] de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., Gattass, M, "Classification of patterns of benignity and malignancy based on CT using topology-based phylogenetic diversity index and convolutional neural network", *Pattern Recognition*, 81, 200-212, 2018.

19. da N´obrega, R. V. M., Rebou¸cas Filho, P. P., Rodrigues, M. B., da Silva, S. P., Dourado J´unior, C. M., de Albuquerque, V. H. C, "Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks", *Neural Computing and Applications*, 32(15), 11065-11082, 2020.

20. Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., Feng, D, "Computer assisted decision support system in pulmonary cancer detection and stage classification on CT images", *Journal of biomedical informatics*, 79, 117-128, 2018.

21. Sang, J., Alam, M. S., Xiang, H, "Automated detection and classification for early stage lung cancer on CT images using deep learning", In Pattern Recognition and Tracking XXX (Vol. 10995, p. 109950S), *International Society for Optics and Photonics*, 2019.

22. Shan, H., Wang, G., Kalra, M. K., de Souza, R., Zhang, J, "Enhancing transferability of features from pretrained deep neural networks for lung nodule classification", *In Proceedings of the 2017 International Conference on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, 2017.

23. Hanafy, Mohamed, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models", *International Journal of Innovative Technology and Exploring Engineering*, Volume-10. 137, 2021.

24. Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A., Sajid Ullah, S, "A Computational Intelligence Approach for Predicting Medical Insurance Cost" *Mathematical Problems in Engineering*, 2021.

APPENDIX A

APPENDIX A: OUTPUT SCREENSHOTS

The first page of the web application creating using flask web framework has 4 main buttons as shown in Fig A.1. These buttons lead to the other pages of the app.

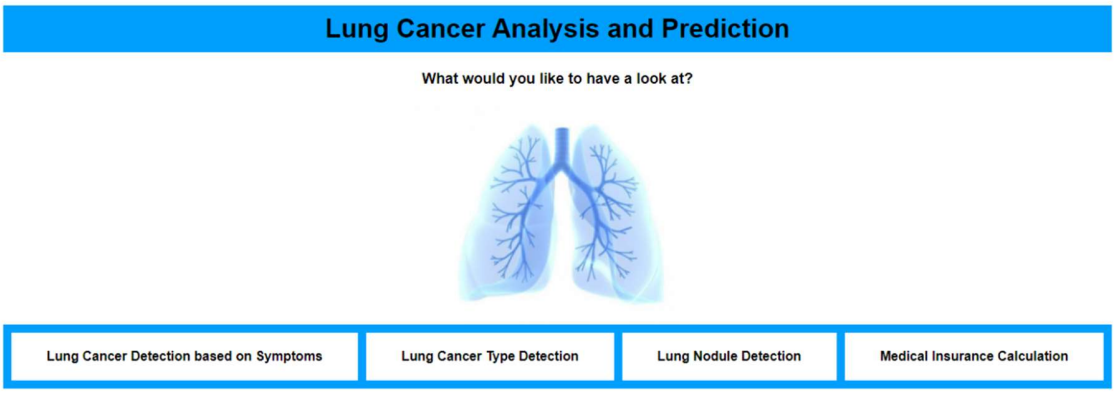


Fig A.1 Main Screen

On selecting the first button, the user is redirected to the screen as shown in Fig A.2.

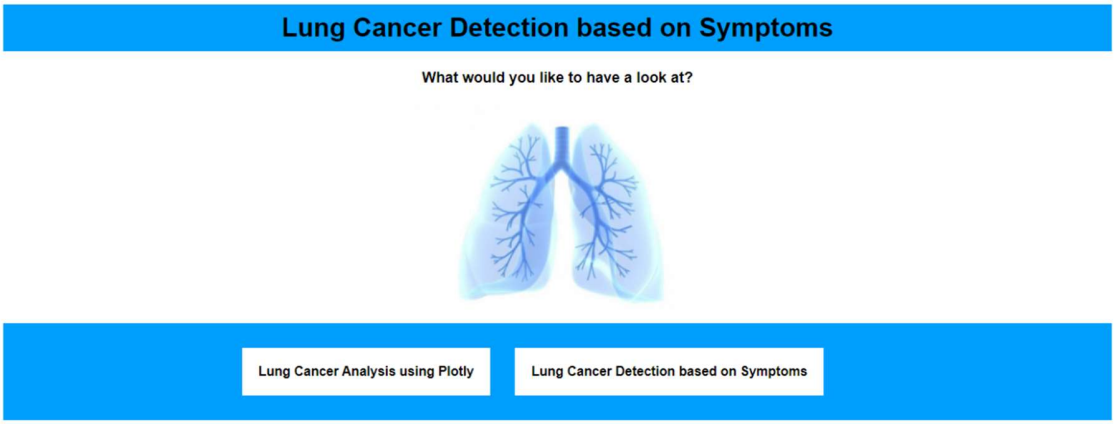


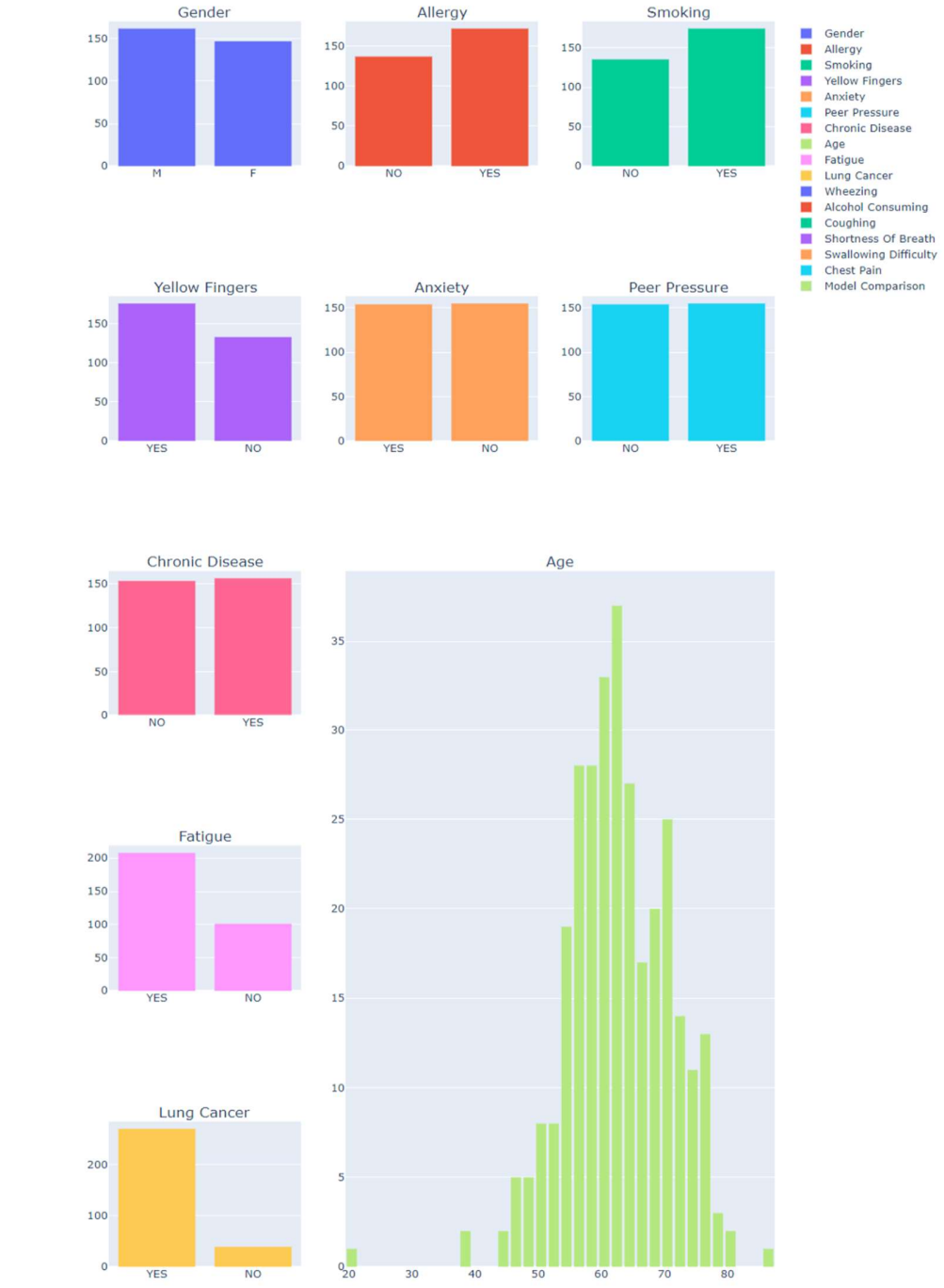
Fig A.2 Lung Cancer Detection based on Symptoms

To view the analysis done for the lung cancer detection based on symptoms, the respective button must be chosen to lead to the webpage as shown in Fig A.3.

Home

Lung Cancer Analysis

Data Distribution, X-axis indicates 'Features' and Y-axis indicates 'Count'



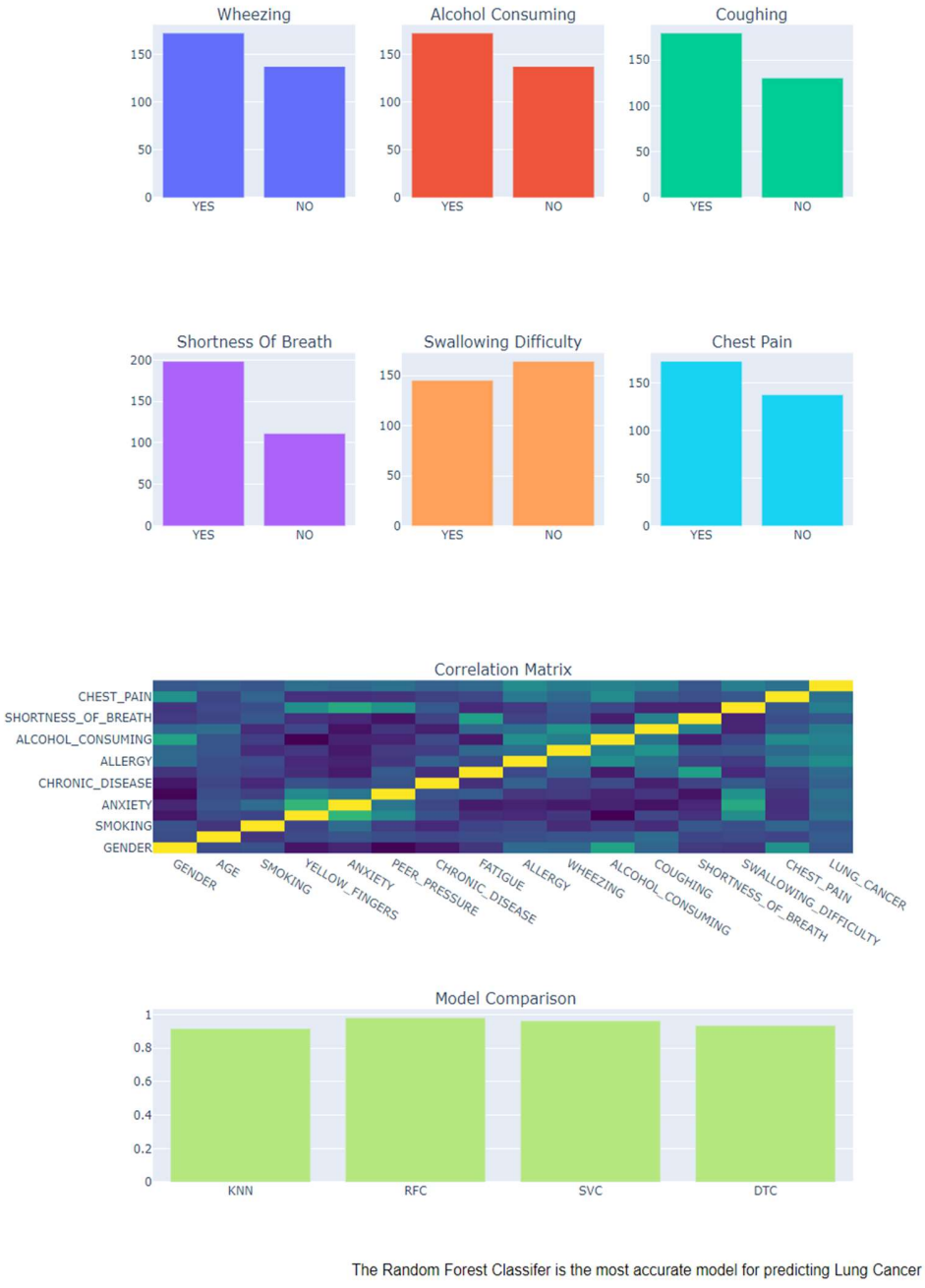


Fig A.3 Lung Cancer Analysis using Plotly

Similarly, to detect the presence of cancer based on symptoms knowledge the other button must be clicked to lead to page shown in Fig A.4. Here, a form with all the needed data will be provided. Based on user's input, the detection is taken place.

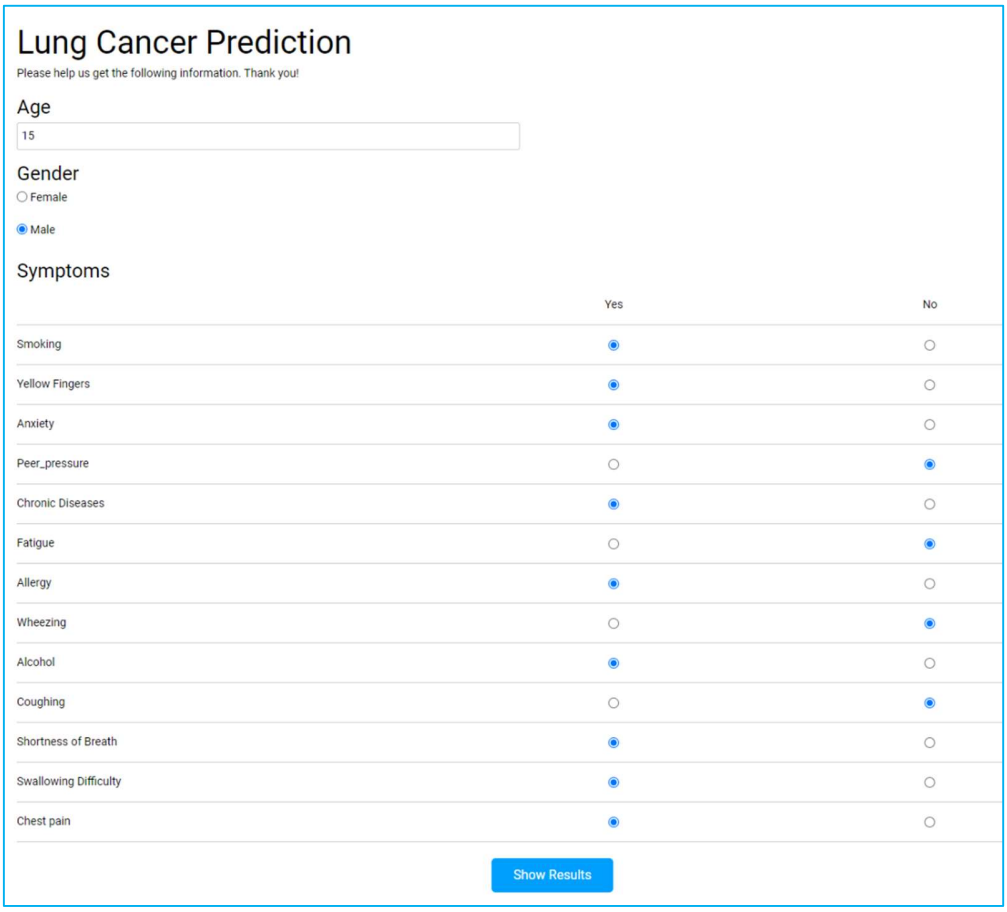


Fig A.4 Lung Cancer Symptoms Form

If the input provided produces a negative output, user is directed to the page shown in Fig A.5. Else, is directed to page shown in Fig A.6.

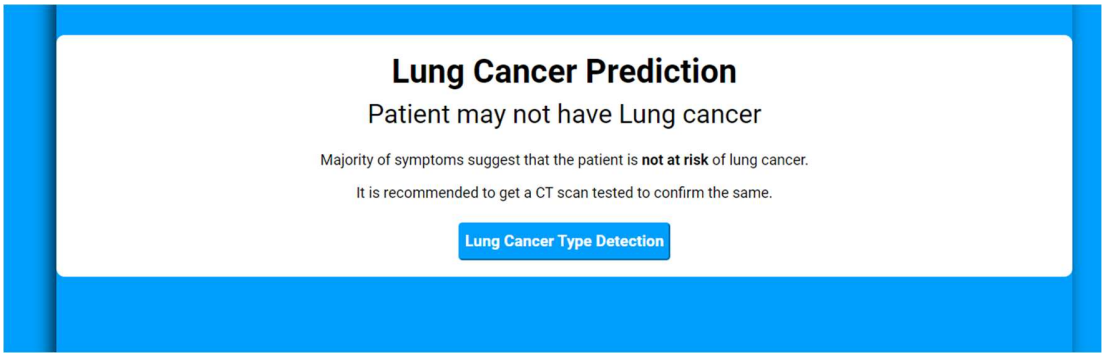


Fig A.5 Lung Cancer Detection based on Symptoms Output: No

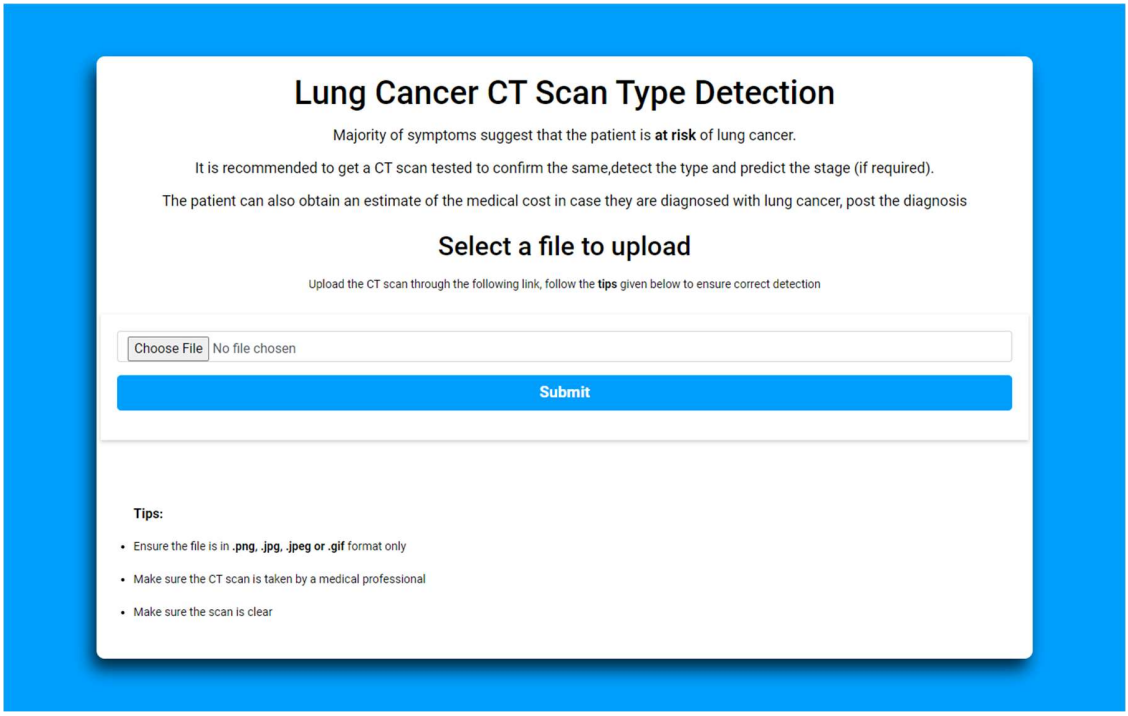


Fig A.6 Lung Cancer Detection based on Symptoms Output: Yes

On the main page, selecting the second button leads to the page shown in Fig A.7.

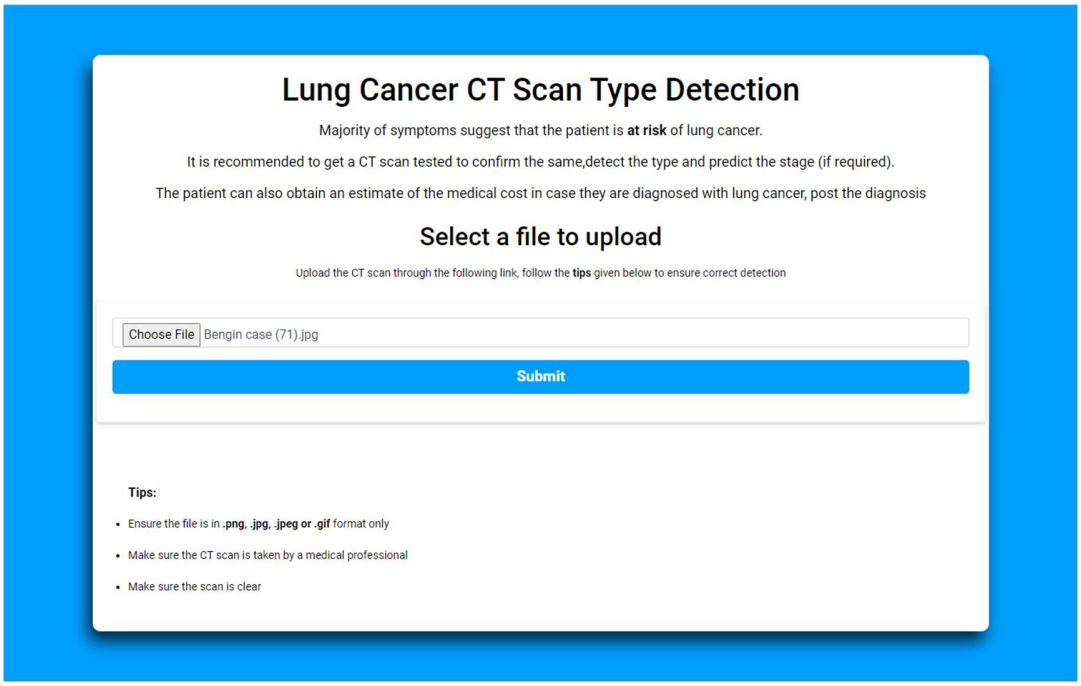


Fig A.7 Type Detection: Benign Case Upload

On uploading a CT Scan to detect the type of case, the results are displayed as in Fig A.8 which shows the example of benign case being predicted.

Similarly, for malignancy case Fig A.9 and A.10 shows what is displayed and Fig A.11 and A.12 for normal case example.

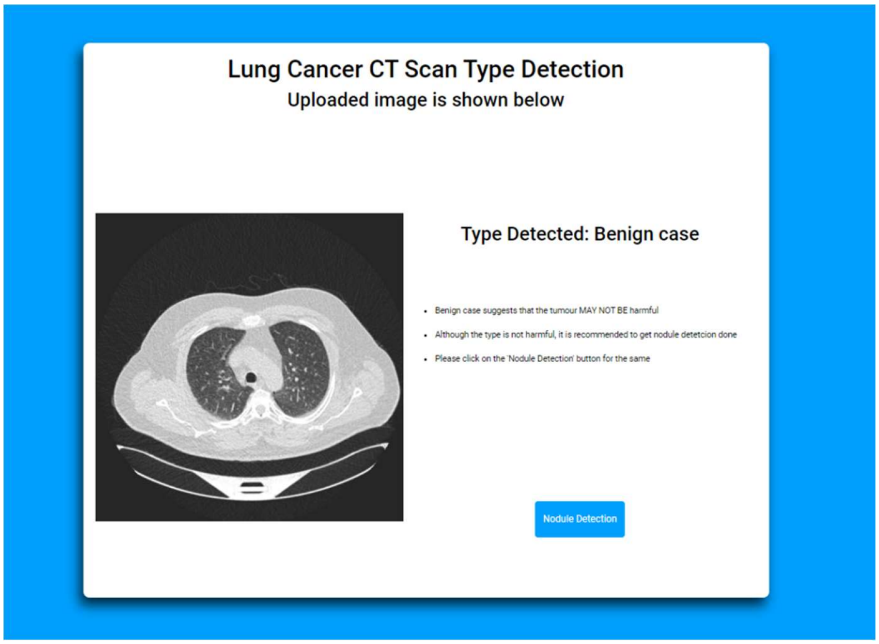


Fig A.8 Type Detection Output: Benign Case

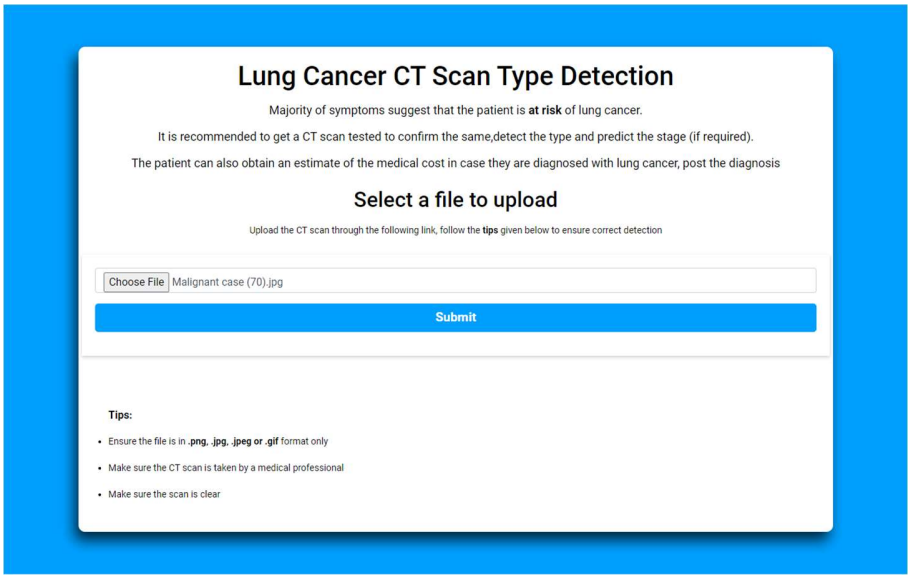


Fig A.9 Type Detection: Malignant Case Upload

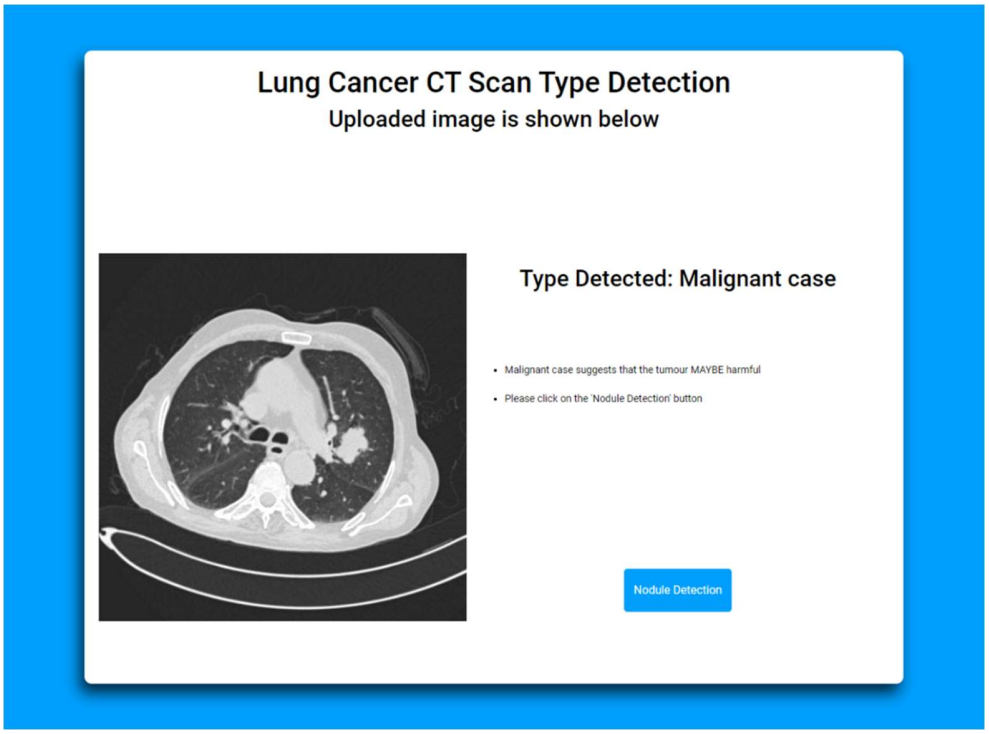


Fig A.10 Type Detection Output: Malignant Case

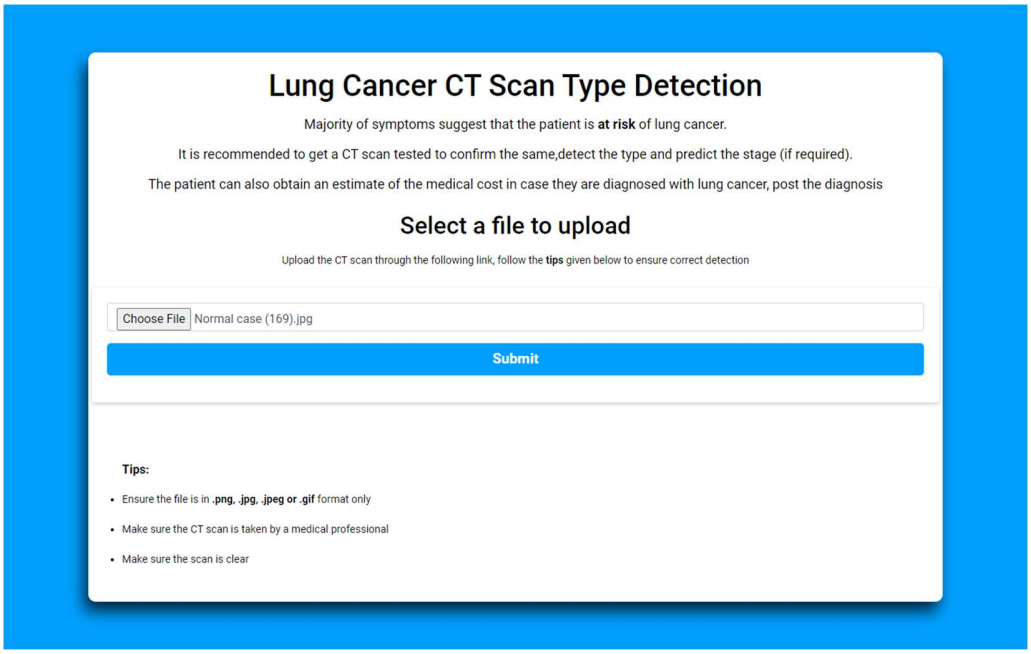


Fig A.11 Type Detection: Normal Case Upload

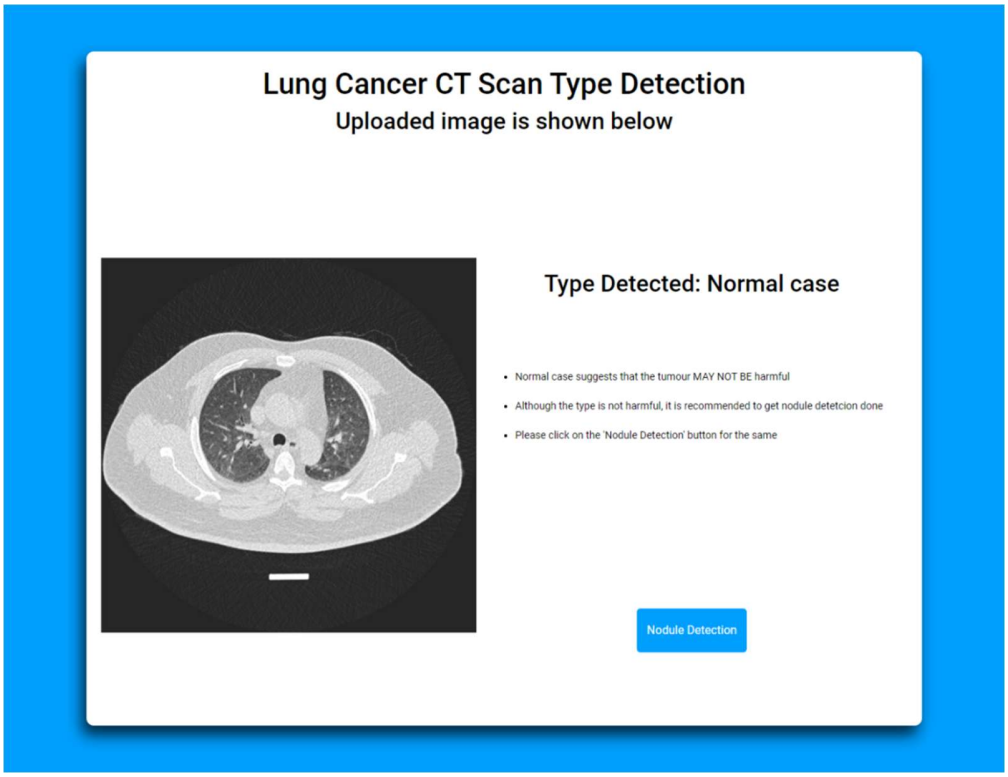


Fig A.12 Type Detection Output: Normal Case

Fig A.13 shows the webpage displayed on clicking the third button on the main page which allows you to upload a CT Scan to detect nodules. Following this, the output is displayed where the CT scan with the ground truth is displayed as shown in Fig A.14.

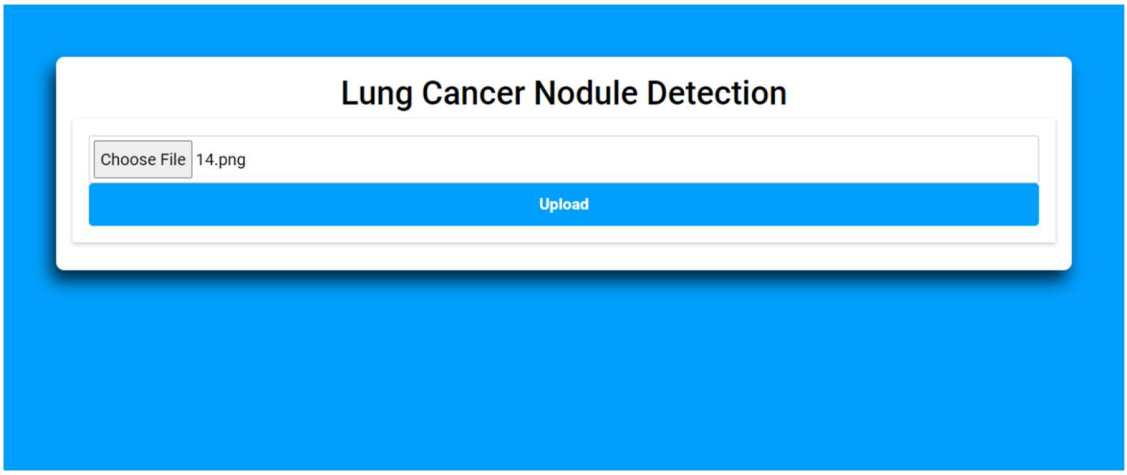


Fig A.13 Lung Cancer Nodule Detection Upload Page

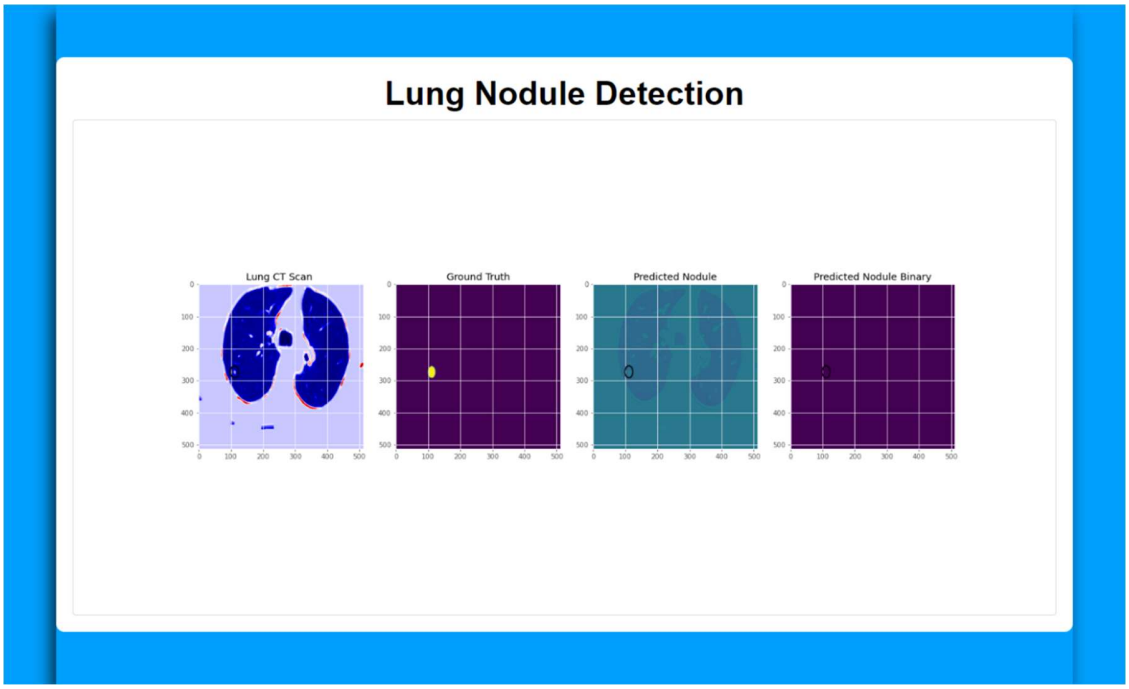


Fig A.14 Lung Cancer Nodule Detection Output Image

The last button on the main page leads to the Medical Insurance Calculation as shown in Fig A.15 with two buttons.

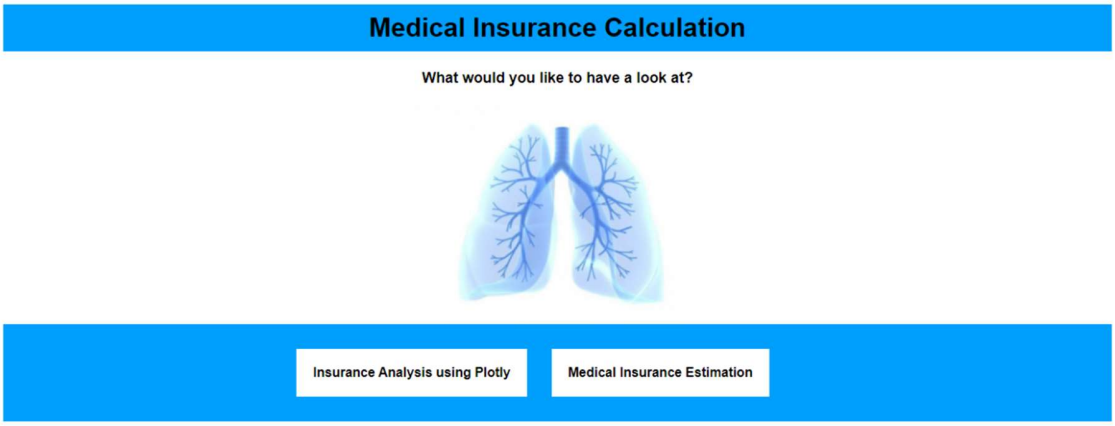


Fig A.15 Medical Insurance Calculation

To view the medical insurance analysis, the first button must be clicked which will lead the user to the page shown in Fig A.16.

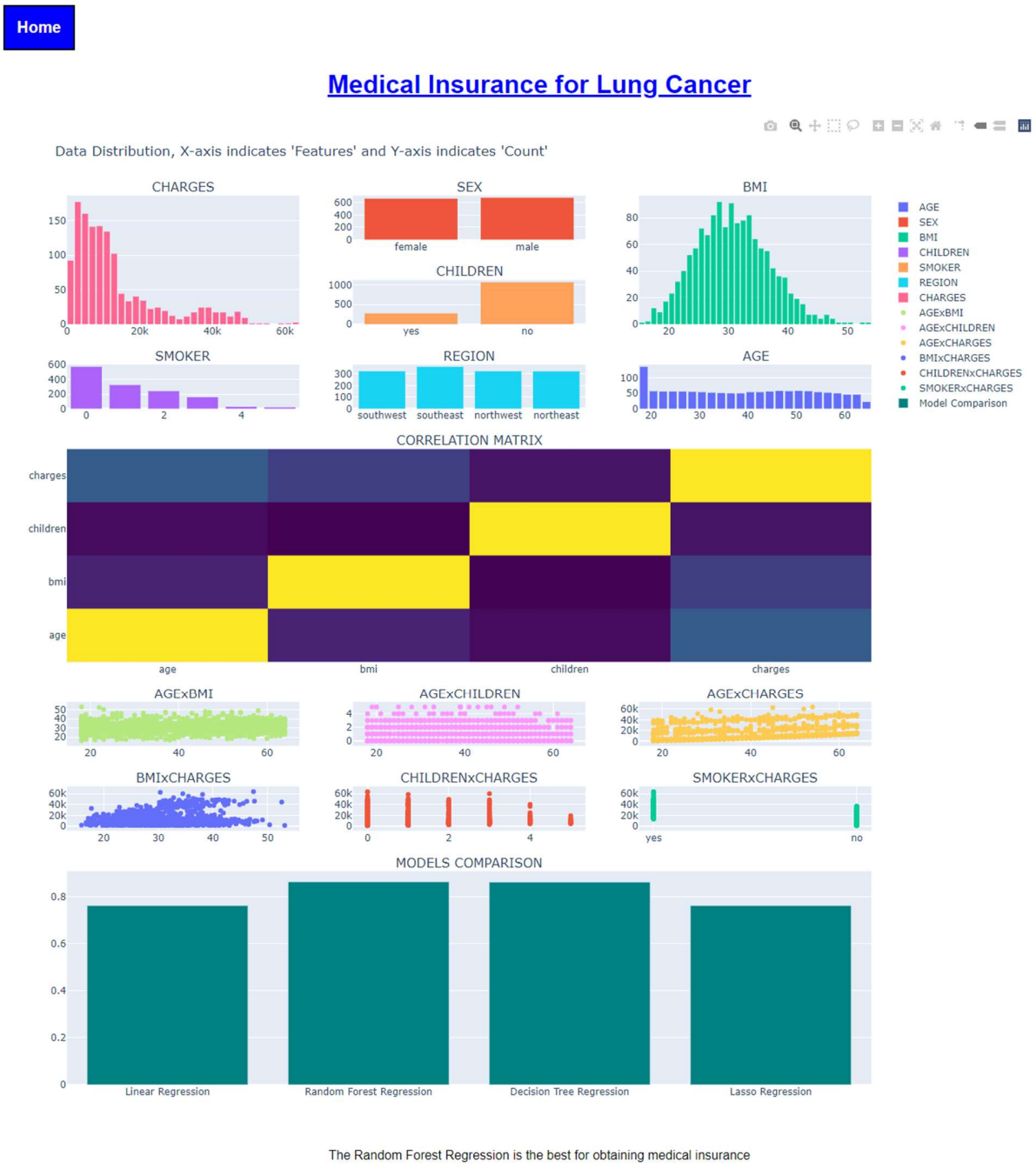


Fig A.16 Medical Insurance Prediction using Plotly

To obtain an estimate of medical insurance costs, the respective button is clicked on to lead to a form-like page shown in Fig A.17. On filling the form, the predicted output is displayed on a new page as in Fig A.18

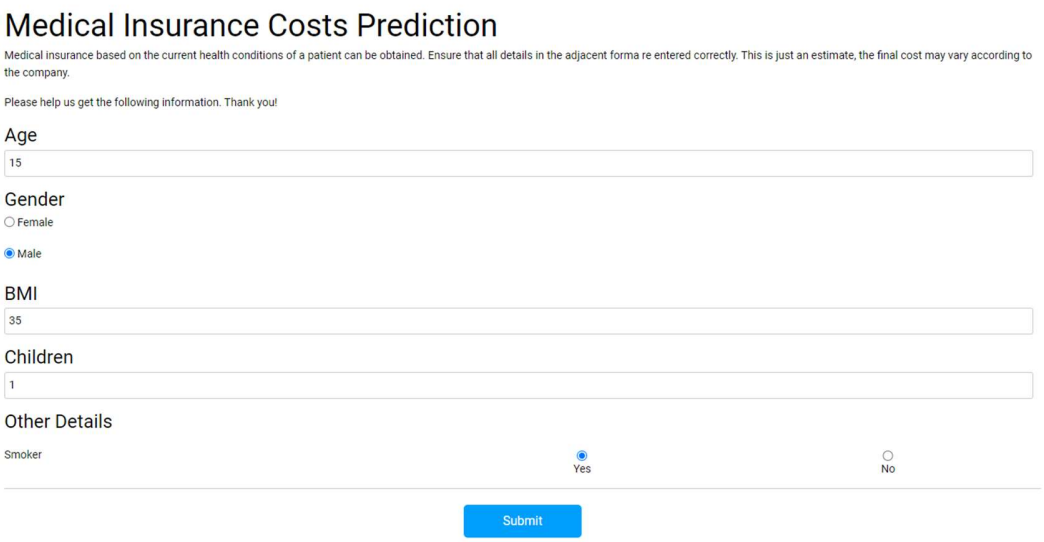


Fig A.17 Medical Insurance Form

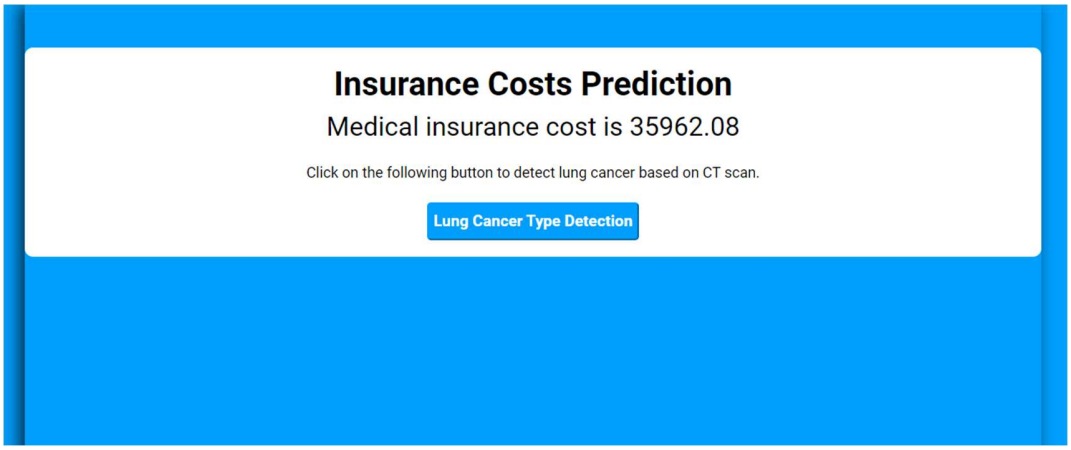


Fig A.18 Medical Insurance Output

PUBLISHED PAPER DETAILS

International Journal for Research in Applied Science and Engineering Technology (IJRASET) is a **UGC** recognized, international peer-reviewed, open-access, multidisciplinary online journal with high impact factor published for the enhancement of research in various disciplines of Applied Science & Engineering Technologies.

Survey Paper Submission Details:

- Paper Title: Literature Survey for Lung Cancer Analysis and Prediction
- Paper ID: IJRASET40245
- Authors: Gayathri Devi Nagalapuram, Varshashree D, Vansika Singh, Dheeraj D, Donal Jovian Nazareth, Dr. Savitha Hiremath
- Publish Date: 05-02-2022
- ISSN: 2321-9653
- Publisher Name: IJRASET, published in Volume 10 Issue II February 2022
- Website Link: https://www.ijraset.com/research-paper/lung-cancer-analysis-and-prediction
- DOI Link: https://doi.org/10.22214/ijraset.2022.40245

Research Paper Acceptance Details:

- Paper Title: A Web Application for Lung Cancer Analysis and Detection using Deep Learning Approach
- Paper ID: IJRASET44134
- Authors: Gayathri Devi Nagalapuram, Varshashree D, Vansika Singh, Dheeraj D, Donal Jovian Nazareth, Dr. Savitha Hiremath
- Publisher Name: IJRASET, published in Volume 10 Issue VI June 2022
- Status: Accepted

GitHub link:

https://github.com/gayathri1462/Lung-Cancer-Analysis-and-Prediction