LUNG CANCER PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for
the award of
Bachelor of Engineering Degree in Computer Science and Engineering

by

CHEKURI GOPI KRISHNA (38110104)
CHENNUPATI SURESH (38110106)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119.

MARCH 2022

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
Chennai - 600119
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICAT

This is to certify that this Project Report is the bonafide work of **CHEKURI GOPI KRISHNA(Reg.No. 38110104),CHENNUPATI SURESH(Reg.No. 38110106)** who carried out the project entitled **"LUNG CANCER PREDICTION USING MACHINE LEARNING"** under our supervision from Nov 2021 to April 2022.

Internal Guide

Dr. T. JUDGI., M.E., Ph.D.,

Head of the Department

Submitted for Viva voce Examination held on_____

Internal Examiner **External Examiner**

DECLARATION

I, **CHEKURI GOPI KRISHNA (Reg.No. 38110104),CHENNUPATI SURESH(Reg.No. 38110106**) hereby declare that the Project Report entitled **"LUNG CANCER PREDICTION USING MACHINE LEARNING"** done by me under the guidance of **Dr. T.JUDGI.,M.E.,Ph.D.,** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Scienceand Engineering.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

We convey our thanks to **Dr. T. SASIKALA M.E., Ph.D., Dean, School of Computing, Dr. S. Vigneshwari M.E., Ph.D. and Dr. L. Lakshmanan M.E., Ph.D., Heads of the Department**, Department of Computer Science and Engineering for providing us the necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Dr.T.JUDGI.,M.E.,Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of **COMPUTER SCIENCE AND ENGINEERING** who were helpful in many ways for the completion of the project.

ABSTRACT

Past years have experienced increasing mortality rate due to lung cancer and thus it becomes crucial to predict whether the tumor has transformed to cancer or not, if the prediction is made at an early stage then many lives can be saved and accurate prediction also can help the doctors start their treatment. Computed tomography plays a vital role in ensuring the condition of tumor that by checking the size of tumor, location of tumor, etc. In this paper, we have proposed a framework for prediction of cancer at an early stage so that many lives that are in an endangered situation could be revived. Basically, our focus is on two domains of computer science that is Digital Image Processing acronymed DIP and Machine Learning. Digital image processing is well-known for the phase of preprocessing the image. In the further stage, the pre-processed image is exposed to segmentation phase and then the segmented image is passed for feature extraction and finally the extracted features are trained using machine learning classification algorithms like SVM (Support Vector Machines), Random Forest, ANN (Artificial Neural Network). Based on the classification results obtained, prediction is made whether the tumor is benign or malignant. The inevitable parameters such as accuracy, Recall and precision are calculated for determining which algorithm has the highest predictive accuracy.

TABLE OF CONTENTS

CHAPTER NUMBER	TITLE	PAGE NUMBER
	ABSTRACT	**V**
	LIST OF FIGURES	**VIII**
	LIST OF TABLES	
1	**INTRODUCTION**	**1**
	1.1 OVERVIEW	1
	1.2 PURPOSE OF MACHINE LEARNING	2
	1.3 PROBLEM STATEMENT	2
	1.4 OBJECTIVES	3
2	**2.1 LITERATURE SURVEY**	4
	2.2 SYSTEM ANALYSIS	7
	2.2.1 DRAWBACKS OF EXIXTING SYSTEMS	7
	2.2.2 PROPOSED SYSTEM	7
	2.2.3 SYSTEM REQUIRMENTS	8
	2.2.4 HARDWARE REQUIRMENTS	8
	2.2.5 SOFTWARE REQUIRMENTS	8
3	**SYSTEM DESIGN**	**10**
	3.1 PROPOSED WORK	10
	3.2 ARCHITECHTURE DIAGRAMM	12
	3.3 MODILES	12
	3.3.1 PRE PROCESSING LAYER	12
	3.3.2 SEGMENTATION LAYER	13
	3.3.3 FEATURE EXTRACTION LAYER	13
	3.3.4 CLASSIFICATION LAYER	14

4	**SYSTEM IMPLEMENTATION**	**15**
	4.1 UML DIAGRAMS	15
	4.2 DATASET RESEARCH	22
	4.3 PROPOSED ALGORITHMS	23
5	**TESTING METHODS AND RESULTS**	**24**
	5.1 UNIT TESTING	24
	5.2 INTEGRATION TESTING	24
	5.3 ACCEPTANCE TESTING	24
	5.4 ACCURACY TESTING	26
6	**CONCLUSION**	**28**
	6.1 FUTURE SCOPE AND CONCLUSION	28
	6.2 REFERENCES	29
	APPENDIX	**33**
	A. SOURCE CODE	33
	B. OUTPUT SCREENSOTS	36

LIST OG FIGURES

FIGURE NO	FIGURE NAME	PAGE NO
3.1	FLOW DIAGRAM OF PROPOSED WORK	11
3.2	IMPLEMENTATION MODEL	12
4.1	USE CASE DIAGRAMS	16
4.1.1	UPLOAD CT SCANS	17
4.1.2	VIEW DETECTION RESULTS	17
4.1.3	MAKE PREDICTIONS	18
4.1.4	VIEW PREDICTIONS	19
4.1.5	CT SCAN SLICES	20
4.1.6	CANCER MASKS	21
5.4	REGION TABLE FEATURE	27
5.4.1	ACCURACY GRAPH	27

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Machine learning, as a powerful approach to achieve Artificial Intelligence, has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives; it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition.

Machine Learning is a scientific research of statistical procedures and methods which they are used by computer systems designed to perform such functions without specific instructions, rather than trusting in the models and conclusions. This is believed to be part of an artificial intelligence. Machine Learning algorithms sets up a mathematical model based on data examples called "training data" to make predictions without the completion of a task being explicitly programmed.

1.2 PURPOSE OF THE MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

1.3 PROBLEM STATEMENT

With the rapid increase in population rate, the rate of diseases like cancer, chikungunya, cholera etc., are also increasing. Among all of them, cancer is becoming a common cause of death. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow older or become damaged, they die, and new cells take their place. When cancer cells develop, however, this orderly process breaks down. As cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumor. This tumor starts spreading to different of body. Tumors are of two types benign and malignant where benign (non-cancerous) is the mass of cell which lack in ability to spread to other part of the body and malignant (cancerous) is the growth of cell which has ability to spread in other part

of body this spreading of infection is called metastasis. There is various type of cancer like Lung cancer, leukemia, and colon cancer etc. The incidence of lung cancer has significantly increased from the early 19th century. There is various cause of lung cancer like smoking, exposure to radon gas, secondhand smoking, and exposure to asbestos etc.

1.4 OBJECTIVE

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

.

CHAPTER2

2.1 LITERATURE SURVEY

In the 21st century, cancer is still considered a serious disease as the mortality rates are high. Among all cancer types, lung cancer ranks first regarding morbidity and mortality [1, 2]. There are two main categories of lung cancer: non-small-cell lung cancer (NSCLC) and small cell lung cancer (SCLC). For non-small-cell lung cancer, a subcategorization into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) is further used. These types of cancers account for approximately 85% of lung cancer cases [3]. Compared with the diagnosis of benign and malignant, further fine-grained classification of lung cancers such as LUSC, LUAD, and SCLC is of great significance for the prognosis of lung cancer. Accurately determining the category of lung cancer in the early diagnosis directly influences the effect of the treatment and thus the patients' survival rate [1, 4]. Positron emission tomography (PET) and computed tomography (CT) are both widely used noninvasive diagnostic imaging techniques for clinical diagnosis in general and for the diagnosis of lung cancer in particular [4]. Immunohistochemical evaluation is considered the gold standard for lung cancer classification. However, this procedure requires a tissue biopsy, an invasive procedure with the inherent risk of a delayed diagnosis and thus exacerbation of the patient's pain.

Advances in artificial intelligence research enabled numerous studies on the automatic diagnosis of lung cancer. The use of data in lung cancer-type classification is roughly divided into three categories: CT and PET image data as well as pathological images [5]. The well-known data science community Kaggle provides high-quality CT images for participants with the task to distinguish malignant or benign nodules from pulmonary nodules. Kaggle competitions repeatedly produce excellent deep learning approaches for these tasks [6, 7]. With the progresses in the research of automatic lung cancer diagnosis, studies are no longer limited to the classification of benign and malignant nodules and data sets are no longer limited to CT images [8–12]. Wu et al. [9] use quantitative imaging characteristics such as statistical, histogram-related, morphological, and textural features from PET images to predict the distance metastasis of NSCLC, which shows that quantitative features based on PET images can effectively characterize intratumor heterogeneity and complexity. Two recent publications propose the application of deep learning to pathological images to classify NSCLC and SCLC [10] and to classify transcriptome subtypes of LUAD [11]. The complexity of the clinical diagnosis of lung cancer is also characterized by the wide range of imaging modality, which is employed in the diagnosis [13, 14].

Previous research already proved that deep learning approaches can not only use the feature distribution patterns from different pulmonary imaging modalities but even merging different features to achieve the computer-aided diagnosis. Liang et al. [15] employ multichannel techniques to predict the IDH genotype from PET/CT data using a convolutional neural network (CNN), while other approaches use a parallel CNN architecture to extract several features of different imaging modalities [16, 17].

Compared with the classification of the benign and malignant, the classification of the three types of lung cancer from medical images are more suitable to constitute a fine-grained image recognition problem as diverse distributions of features and potential pathological features need to be considered. Because the fine-grained features which need to extract in images, and meanwhile the lesion region is a small part of the whole image, the deep learning framework is susceptible to feature noise. At present, most methods based on various deep learning frameworks have proved to have certain bottleneck in fine-grained problems. In order to solve this problem, the previous research mainly implements the attention mechanism from the two dimensions (channel and spatial) of the feature representation. The channel attention mechanism models the relationship between feature channels [18], while the spatial attention mechanism ensures that noise is suppressed by weighting feature representation spatially [19–21]. So far, spatial attention mechanism has been used in medical image processing to enhance extracted features [20, 21]. The channel attention mechanism has been used in the detection and classification of pulmonary disease [22, 23]. The presentation of these attention mechanisms illustrates the source of characteristic noise from different perspectives. There are few related studies on how to use the attention mechanism more effectively on images with different imaging modalities, so the deep learning model based on the multimodality dataset still has problems in fine-grained problems.

Many works has already been proposed for prediction of cancer by various researchers among then Palani et al., [5] has proposed IoT based predictive modeling by using fuzzy C mean clustering for segmentation and incremental classification algorithm using association rule mining and decision tree for classification for classifying the tumor sets and based on the output generated by incremental classification model convolutional neural network has been applied with other features for predicting benign or malignant.

Lynch et al., [6] Various machine learning algorithm are implemented for predicting the survivability rate of person, performance is measured based on root mean square error. Each model is trained using 10-fold cross validation, as the parameters are preprocessed by assigning default value so cross

validation is used for avoiding over fitting.

FENWA et al., [3] proposed a model whether feature like contrast, brightness from the image dataset is extracted using texture based feature extraction and on that two type of machine learning algorithm are applied one is artificial neural network another one is support vector machine and then performance has been evaluated on both the algorithm to compare which algorithm is giving more accuracy.

Öztürk et al., [7] proposed a model where a five type of feature extraction techniques were used in individual classification algorithm to predict at which features extraction technique which machine learning algorithm is giving more accuracy.

Jin et al., [8] proposed a model where the original image is first converted into binary image the erosion and dilution has been operated on that image after that image has been segmented on the segmented image region of interest extraction is applied to identify volume or size of the tumor and after extraction convolutional neural network is applied with softmax classification layer to recognize the tumor is cancerous or not.

Sumathipala et al., [9] proposed a model where the image data are taken from LIDC-IDRI, after collecting the image data image filtration has been implemented, filtration is done based on the patient who went through biopsy and module level is equal to 30 and then images whose module level is equal to 30 is segmented and then Logistic regression and random forest has been applied for prediction.

2.2 SYSTEM ANALYSIS

2.2.1 DRAWBACKS OF EXISTING SYSTEMS

In some cases, the application still does not have accurate results. Further optimization is needed.

Priority information is needed for segmentation. Database extension is required for greater accuracy.

Only a few diseases are covered. Therefore, the work must be expanded to cover more diseases.

Possible causes that can cause misclassification can be: Symptoms of the disease vary from cigarettes, optimizing the characteristics needed, more training patterns are needed to cover and predict more cases - the actual disease.

2.2.2 PROPOSED SYSTEM

The main theme of this project, is to detect the lung cancer and to take precautions to avoid or clear that diseases. It overcomes the drawbacks of existing system.

The implementation phase begins with smoking as input and do the following steps:

- Pre-Processing layer.
- Segmentation layer.
- Feature Extraction of layer.
- Machine learning classifier.

After performing all the above steps, we can detect the disease of lung cancer.

2.2.3 SYSTEM REQUIREMENTS

Requirement analysis determines the requirements of a new system. This project analyses on product and resource requirement, which is required for this successful system. The product requirement includes input and output requirements it gives the wants in term of input to produce the required output. The resource requirements give in brief about the software and hardware that are needed to achieve the required functionality.

2.2.4 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It shows what the systems do and not how it should be implemented.

- i5 intel 8th Gen Processor
- 12 GB RAM
- 1 TB Hard Disk
- 4 GB Nvidia GPU
- Monitor
- Web camera

2.2.5 SOFTWARE REQUIREMENTS

The software requirements are the specification of the system. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the team's and tracking the team's progress throughout the development activity.

- Windows 10
- Web Browser
- Python Package Manager
- IDE
- Linux

CHAPTER 3

SYSTEM DESIGN

3.1 PROPOSED WORK

Based on the literature survey a novel model has been proposed which consist of pre-processing block, segmentation block, feature extraction block and then classification block. In prediction of cancer CT scan report is basically used. But CT scan report is full of noise which cannot be seen by human eye for that reason various digital image processing plays a important role to get a noise free image. Digital image processing is the process where the analysis and manipulation of image is used to extract some useful information from the image. Digital image processing involve various step like image pre-processing where we can enhance the image using histogram equalization, spatial filter etc. Then image restoration can be done where various kind of noise like salt and pepper noise, Gaussian noise etc are applied and filter like median filter, mean filter can be applied on the pre-processed image. After that color conversions is applied only if the image is colored image then convert it to gray level. Fig. shows the proposed novel framework. Image segmentation is a process which divides the image into several segment based on the pixel, once the image segmentation is over the feature extraction can be applied. Feature extraction is a type of dimensionality reduction where a set of raw data is reduced to more manageable group image data for extracting the feature like region and texture. After extracting the feature different machine learning technique is used to classify the image.

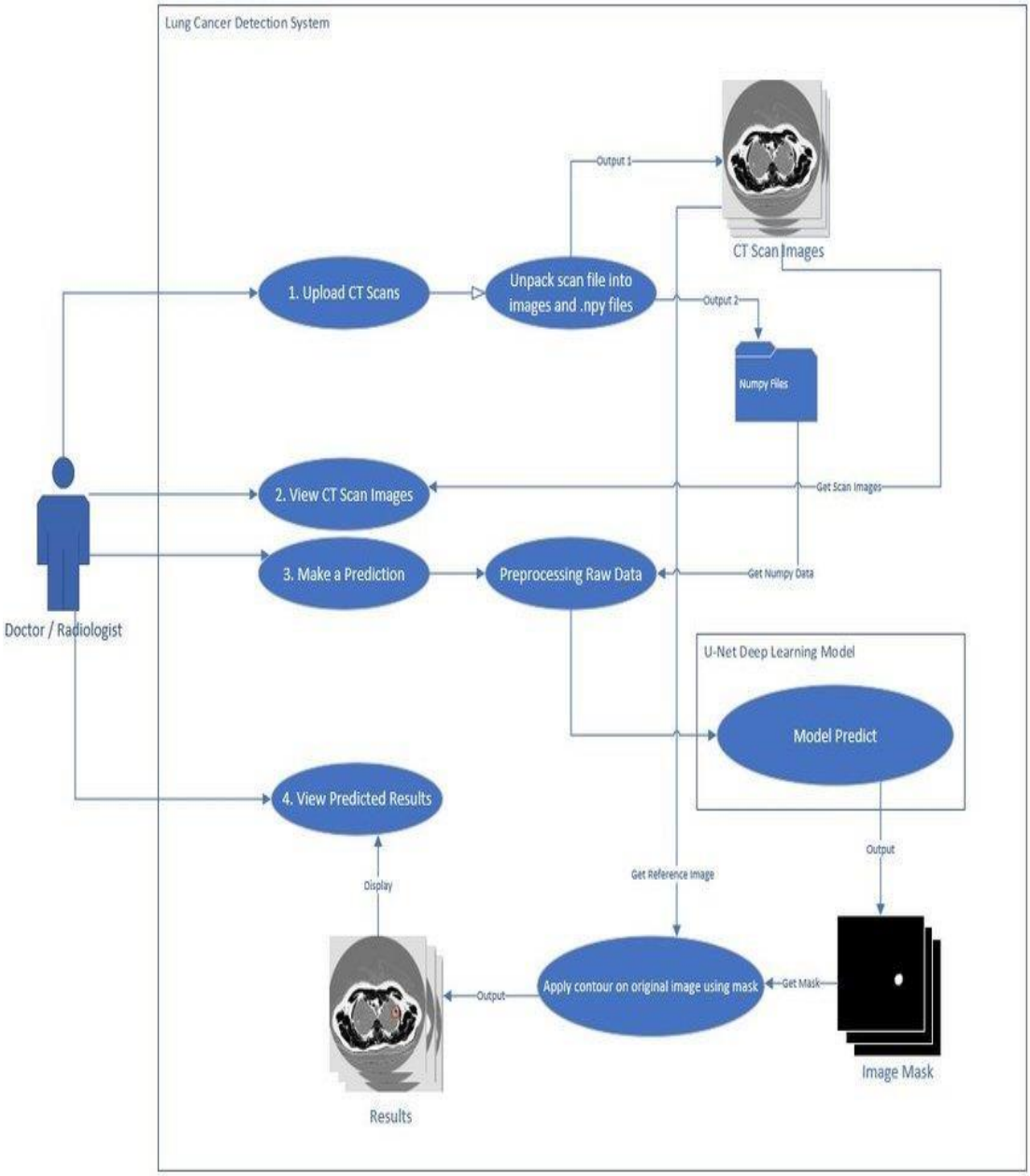


Fig 3.1 Flow Diagram of Proposed work

3.2 ARCHITECTURE DIAGRAM

The following algorithm describes the step by step approach for the proposed model.

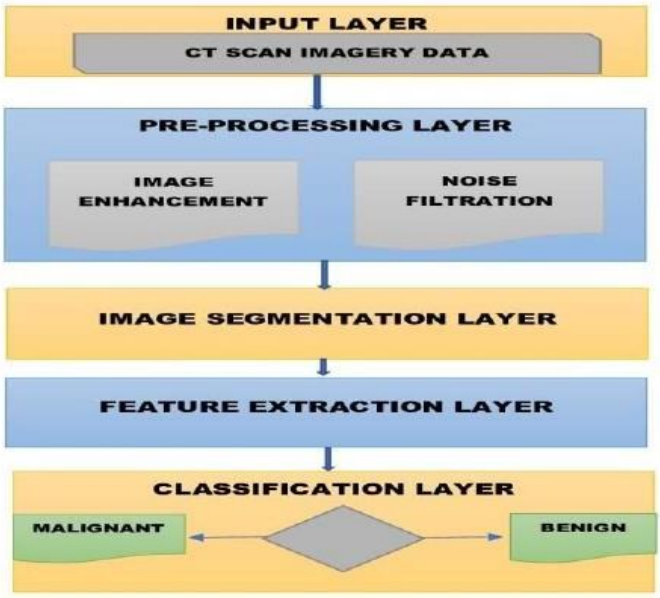


Fig 3.2Implementation Model

.

3.3 Modules

3.3.1 Pre-Processing Layer

Image has been collected from LIDC-LDRI. The original image was full of noise and for that first we have applied histogram equalization on the image to enhance the image and then on the equalized image median filter has been applied to remove the noise which was already present in the image after getting the noise free image we have applied some more noise in the image yield more clearer picture then again noise has been removed using median filter. Generally median filter is non linear digital filtering technique and it is also used as smoothing

of images as it don't blur the edges completely as compare to other filtration technique like Gaussian filter or average filter.

3.3.2. Segmentation Layer

Image segmentation is a method of partitioning the image into various parts. After pre-processing the image on the pre-processed image segmentation is applied to acquire the information from the image. For image segmentation first we have applied edge detection technique through edge detection we can segment the boundary of the image for edge detection prewitt operator has been used, on that operator threshold has been applied so that after edge detection the intensity value which is less than threshold is removed and the intensity value which is higher than or equal to threshold will consider for further segmentation after getting the segmented image by edge detection we will apply watershed segmentation on the output image. Watershed segmentation takes the concept topographical landscape with ridge and valley which is defined by a gray level with respective pixel or gradient magnitude. There exist various ways to segment using watershed segmentation here we have used watershed segmentation using gradient. The gradient magnitude is used to preprocess the gray scale image; it has high pixel value along the object edge and low pixel value in another left region. And through this we can get the final segmented image through which we can extract features.

3.3.3. Feature Extraction Layer

The output generated by segmentation is used for feature extraction. By doing feature extraction we have extracted two types of feature one is region based another is texture based region based we have extracted feature like area in context to image means pixel of the image, perimeter in context image mean vector containing the distance around the boundary of each region in the image, centroid means the centre of mass of the region and it is in 1 X 2 vector form, image and based on texture we have extracted feature like mean is used to find average intensity, standard deviation is used to measure average contrast, smoothness used to measure relative smoothness of the intensity in the region, entropy is used to measure randomness using statistical approach of texture based.

3.3.4. Classification Layer

After feature extraction we will apply classification technique on both the feature to compare at which feature extraction which machine learning algorithm is giving more accuracy. Machine learning algorithm which has been used is support vector machine, artificial neural network and Random forest. After applying classification technique, it can be predicted that the tumour is cancerous or not and at which feature we are getting more accurate prediction.

CHAPTER 4

SYSTEM IMPLEMENTATION

Proposed Algorithm can be viewed as follows:
Input: Image Data (ID)
Output: Classification as benign or malignant
Step 1: Input the image data (ID)
Step 2: Pre-Process the image
Step 2.1: If the image is noise free
Go to step 3
Else
Go to step 2.1
Step 2.2: Apply image Enhancement Method
Step 2.3: Apply filter to enhanced image to reduce noise
Step 3: Segment the image
Step 3.1: Segment the boundary of the output image generated at step 2.3 using Edge Detection
Step 3.2: After edge detection segment apply watershed gradient segmentation.
Step 4: Feature Extraction
Step 4.1: Region based feature are extracted like area, perimeter, centroid.
Step 4.2: Statistical based feature are extracted like mean, standard deviation, smoothness.
Step 5: Apply classification algorithm for training and prediction of tumour as benign or malignant.
Step 6: Evaluate the parameter like accuracy, precision, Recall.
Step 7: End

4.1 UML Diagrams

Use case diagram

The following Figure Shows an overview of the lung cancer detection system. Users will be able to upload a CT Scan, view the detection results and view cancer diagnostics, at this current moment we are not yet sure if we are be able to deliver the cancer diagnostics part of this project so it is optional.

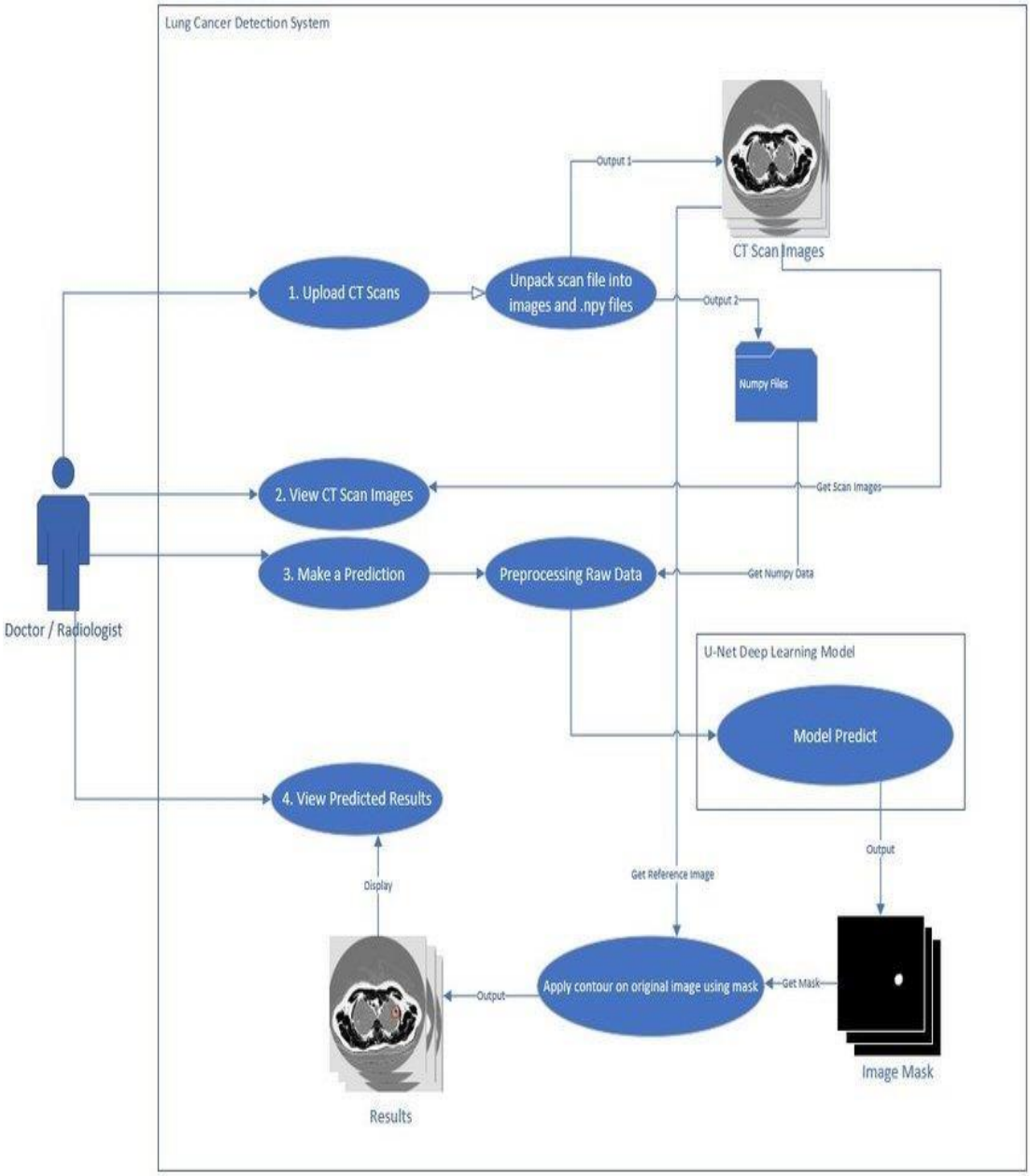


Fig: Use Case Diagram of the System

The following figure shows the first use case. The user uploads a metadata file (.mhd) and a raw file (.raw) to the back end system via POST request. The system takes the metadata and uses it to unpack the raw files which contain the images. The system then takes the image data and saves it into image files (.png) and image data array (.npy) using OpenCV and Numpy. The reason for this is because the images generated by OpenCV is used to show to the users in the gallery. Saving the images into (.png) alters the image arrays so the model does not react well to the changes in the data. Instead we use create Numpy files for each image to give to the model

and use the filenames to reference the image and the numpy files.

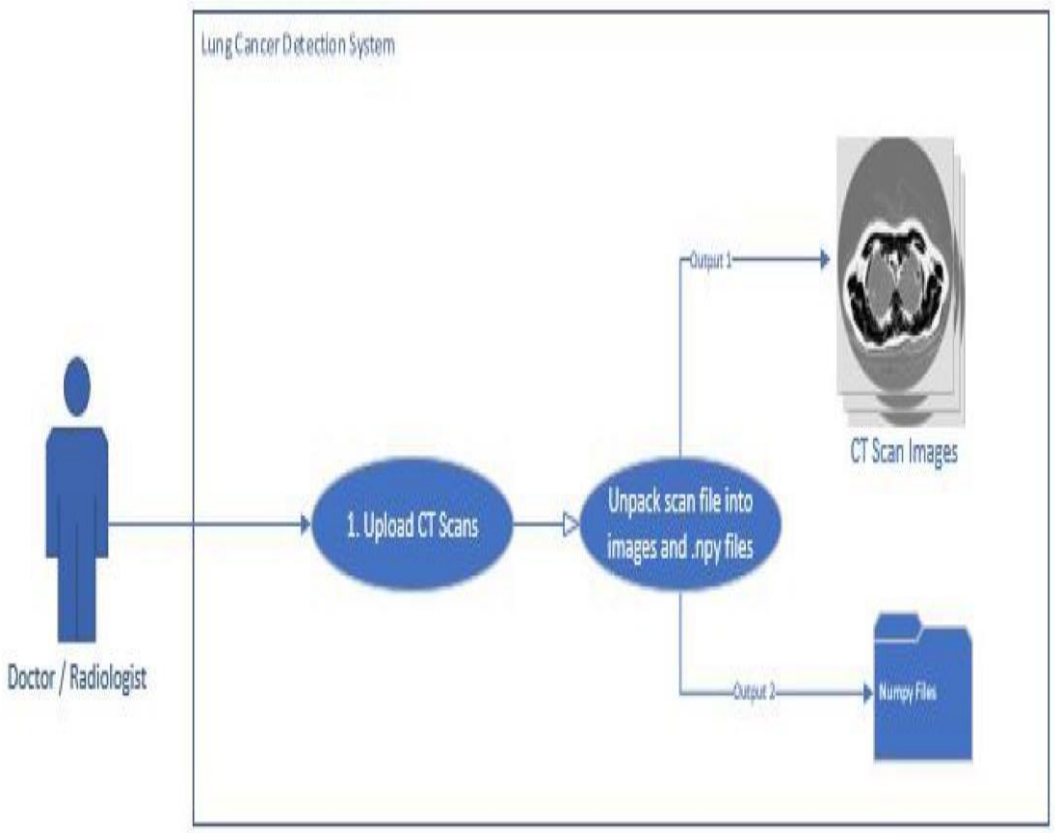


Fig: Upload CT scans

The following Figure shows the second use case. The system pulls the image files from the back end and displays it in the front end. The system does this by sending GET requests for each image. The images on the front end can be displayed via a carousel image or a gallery style.

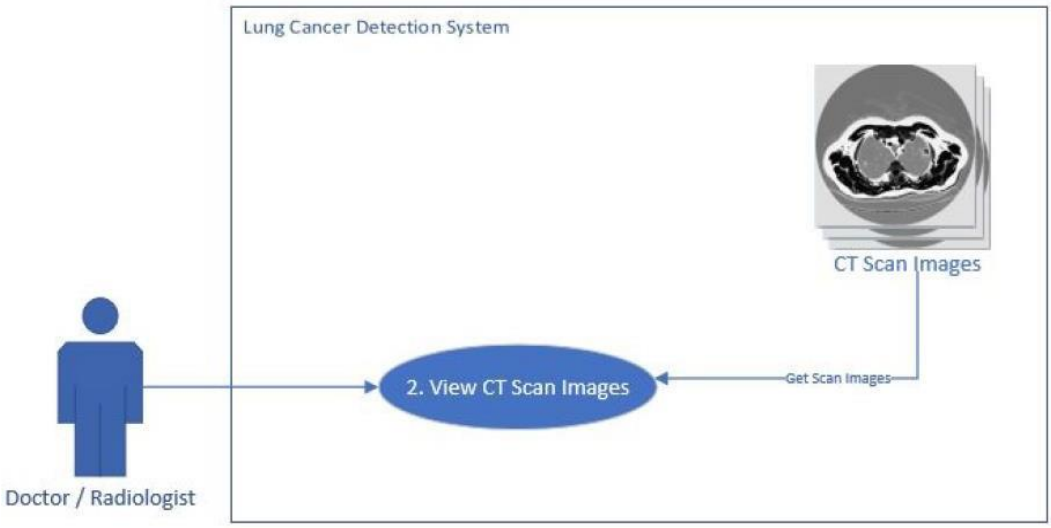


Fig: View detection results

The following Figure shows the third use case. When a user selects and image that he or she wants to make predictions. The system takes the filenames from the user during selection and uses this filename to reference a numpy file. This numpy file is then preprocessed before its fed to the deep learning model. The model then outputs an image mask as seen on the diagram.

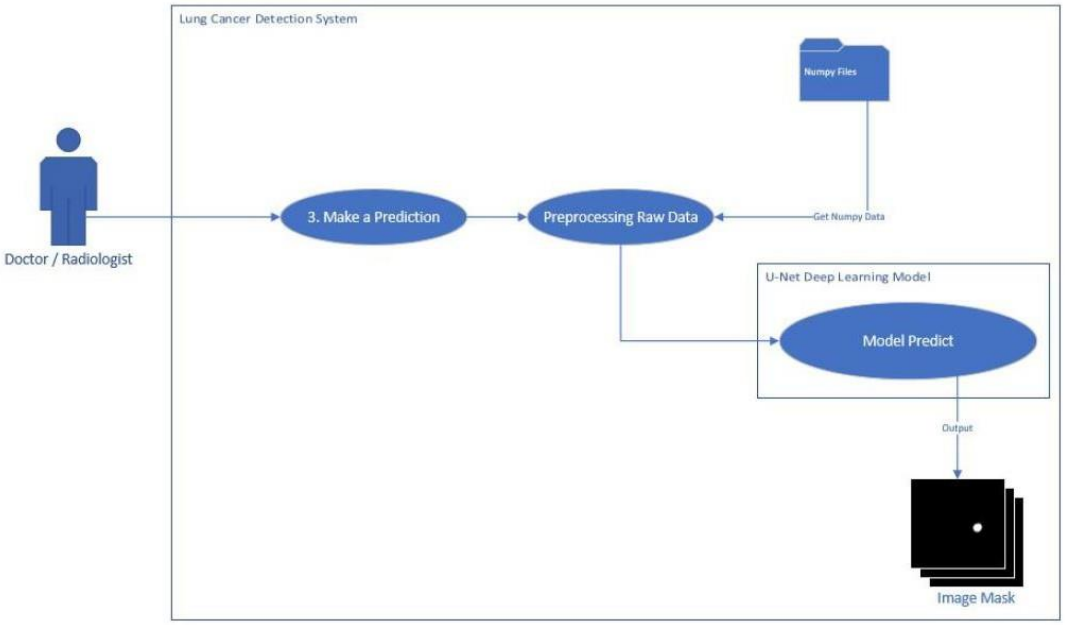


Fig Make Predictions

The following Figure shows the last use case. The system takes the original CT scan reference image and the associated mask and applies an image contour on the original image.

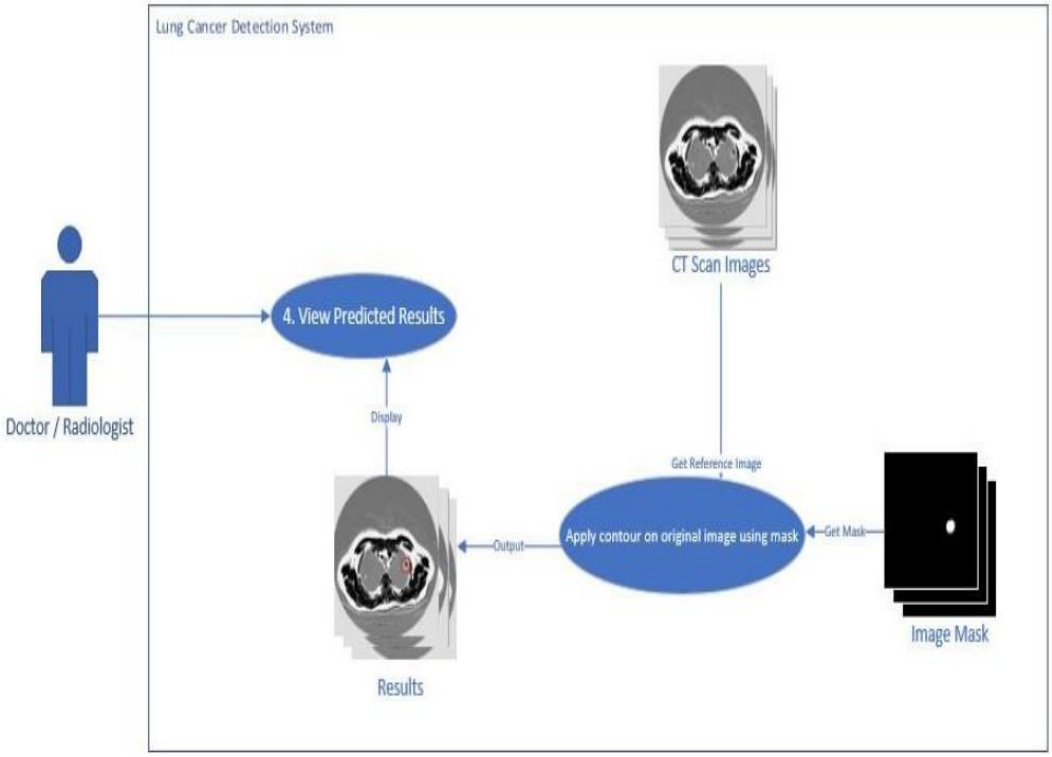
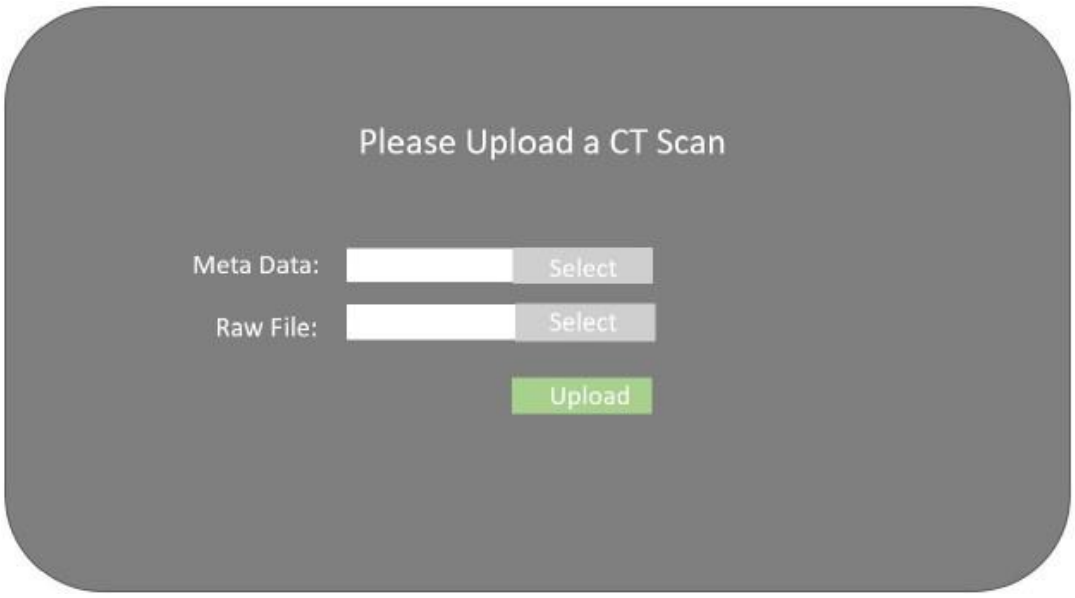


Fig: View Predictions

Home page



CT Scan slices

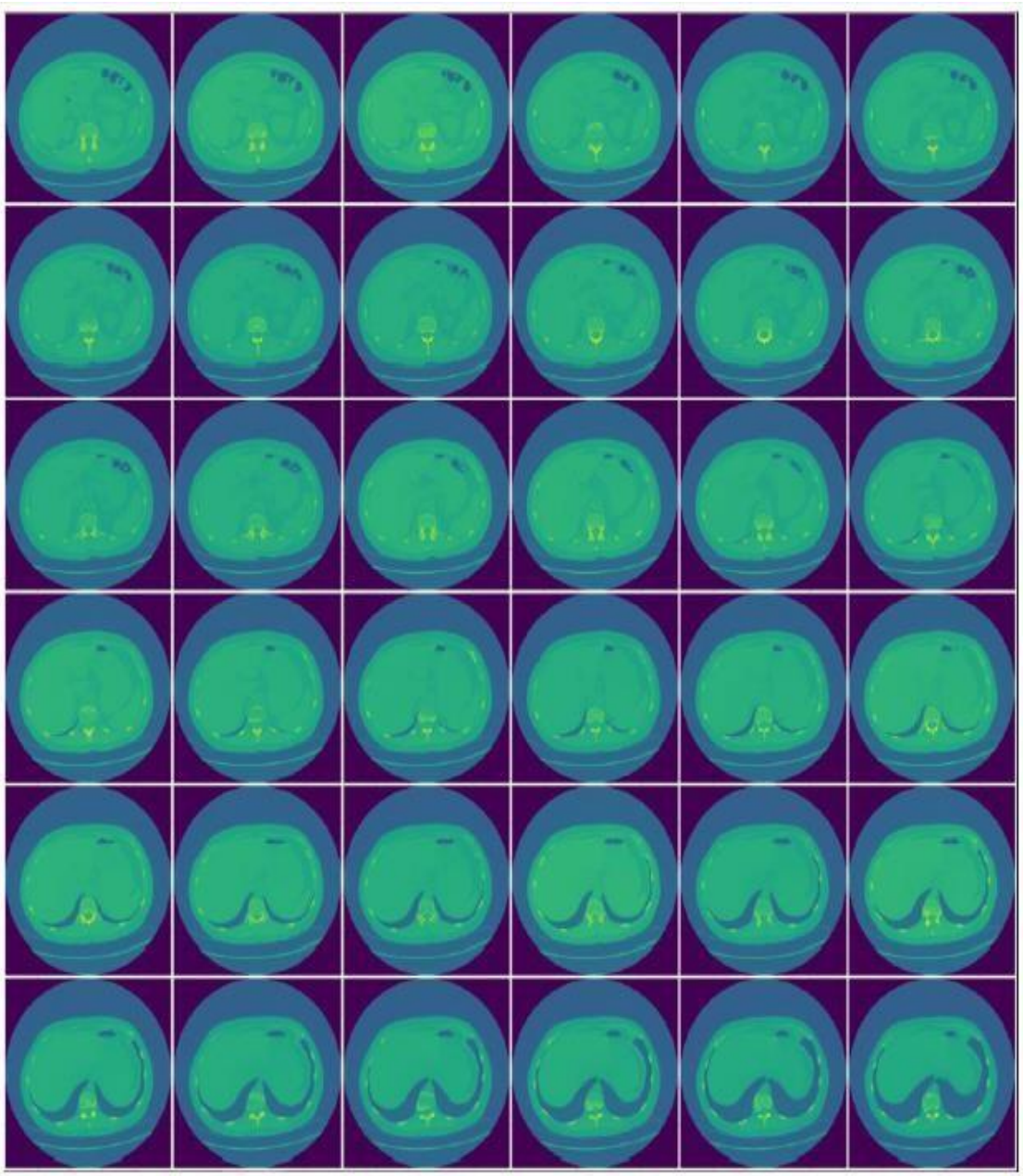


Fig CT Scan slices

The above Figure shows a set of images for a single CT scan. The dimensions for the CT scan is about 512 height, 512 width and approximately 200 images, although there also exists scans which are over 300 images.

Cancer Masks

With the use of the annotations and Mulholland et al's makemask algorithm. The author was able to extract a boundary around cancer nodules. A short explanation of masks and the makemask algorithm used is shown in the appendix. The following Figure shows sample images of cancer masks, the majority of which is small and some are large.

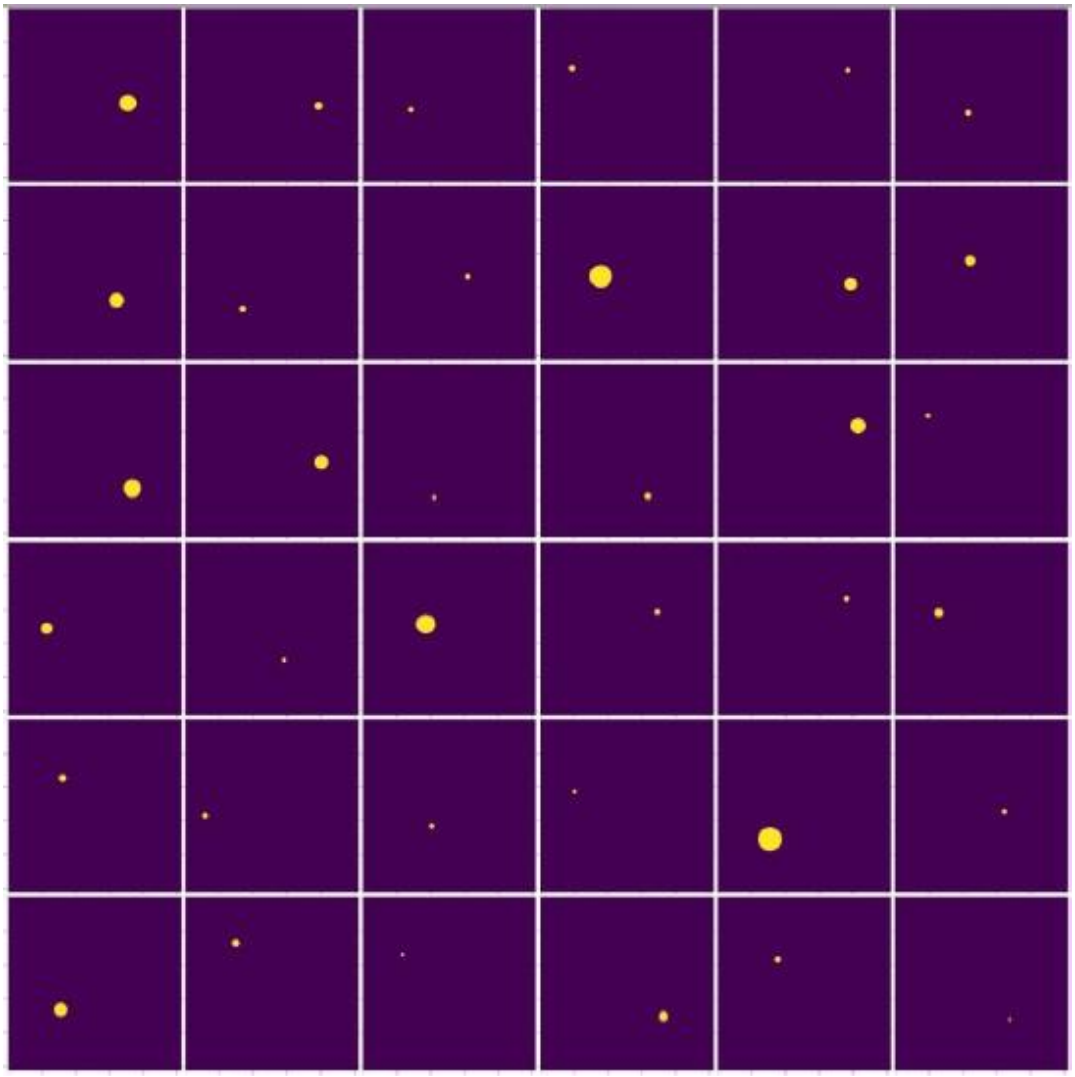


Fig: Cancer Masks(Labelled)

Creating Lung Images:

The next step is to create our lung images segmented from our original image. This is also done via Mulholland et al's algorithm shown in the appendix section. The following Figure shows a sample images of segmented lungs with cancer, we can see some of the cancer is very small and hard to determine visually but some are very large and are clearly malignant.

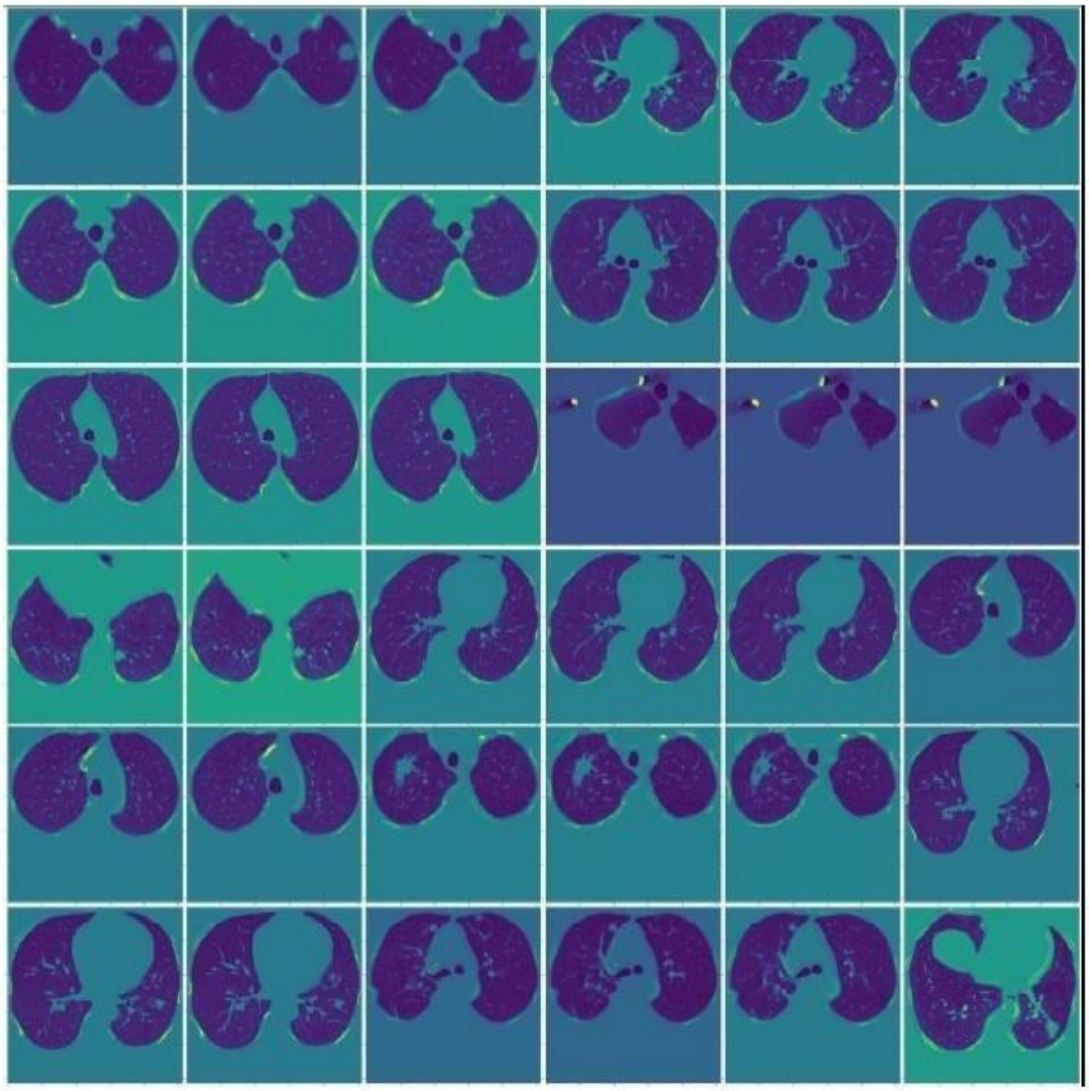


Figure: Lungs Segmented (Feature Set)

4.2 DATASET RESEARCH

Building deep learning models require a lot of data.For this project datasets has been researched and identified before any real work has begun. Since there is a heavy emphasis on building models for this project,a key part of the project relies on a Dataset. Prior to coding,we had to ensure that we had a great dataset to work with to build a model. From doing research there are 2large datasets that we could work with. **Kaggle 3D Unlabeled Dataset:Data Science Bowl 2017:** This dataset was part of the Kaggle competition Data Science Bowl2017[4]. The topic of the competition wasaboutlungcancerdetection.The dataset was provided by the National Lung Cancer Screening Trial, The Cancer Imaging Archive, Diagnostic Image Analysis Group (Radboud University), Lahey Hospital and Medical Center and Copenhagen University Hospital. The dataset contains full CT scan

images of a patients lungs. The dimension for this is (512, 512, 200) which is (Height, Width, No. of Images). For this project the dataset has been used to segment different parts of the CT scans as part of feature engineering and visualizations. This dataset was what we originally wanted to work with as the data was labeled as desired and useful for the project. However the largest challenge that hindered from continuing using this data is the size of the entire dataset. The entire dataset is about 100GB zipped which could not fit on the laptop. Preprocessing the entire dataset would also be too computationally heavy.

LUNA16 Labeled 3D- Lung Nodule Analysis Dataset 2016: The LUNA16 dataset is also 3D CT scans of lung cancer annotated by radiologists. The dataset contains 3D images and a CSV file containing annotations. This dataset was part of the LUNA16 Grand Challenge in 2016[38]. The dataset is still publicly available for research. Like the Kaggle dataset the format of the 3D Image is a 3-dimensional array (512, 512, 200) (height, width, no. of images). The main advantage of this dataset is that the dataset is broken down into10 subsets which makes it much easier to work with.

4.3 PROPOSED ALGORITHMS

- **SVM (Suppoot Vector Machine**) :One of the simple and useful approaches in supervised learning is support vector classification. User defined support vector classifier can be framed using various kernel function to improve the accuracy. Support vector classifier is well suited for both structured and unstructured data. Support vector classifier is not affected with over fitting problem and makes it more reliable.

- **Random forest:** Combination of classifier trees represents random forest classifier. One of the finest approaches to represent input variables in form of trees that makes a forest like structure. Input Data are represented in trees and each tree specifies a class label. Random forest depends on its error rate. Error rate signifies in to two directions. First one is the correlation between trees and second one is the strength of the tree.

- **ANN (Artificial Neural Network):** Neural network are the basic block of machine learning approach in which the learning process is carried in between neuron. Artificial neural network (ANN) comprises of input layer, intermediate layer having hidden neurons and output layer. Every input neuron is connected to hidden neuron through appropriate weight and similarly weight is connected between hidden unit to output unit. Neuron presented in hidden neuron and output neuron are processed with some known threshold functional value. Depending on the requirement the activation will be used to process the neuron.

CHAPTER 5

TESTING METHODS

5.1 UNIT TESTING

Unit testing is typically conducted as part of the software life cycle 's combined code and unit test process while coding and unit testing is not unusual to be done as two separate stages.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Check goals

• All field entries need to work correctly.

• Pages have to be allowed from the connection found.

• Don't miss the entry screen, notifications and answers.

Attributes to check

• Make sure submissions are of the right format

• Don't authorize duplicate entries

• All links will take the user to the page in question.

5.2 INTEGRATION TESTING

Software integration testing is the gradual integration testing on a single platform of two or more compatible software components to create errors triggered by device defects. The purpose of the integration check is to verify whether components or software applications communicate without error, e.g. components in a software system or-one step up-software applications at the company level.

Test Results: All of the above test cases passed successfully. No defects found.

5.3 ACCEPTANCE TESTING

User Acceptance Testing is a critical phase of any project and requires substantial end user involvement. This also ensures the framework meets the practical demands.

Test Results: All of the above test cases passed successfully. No defects found.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.
Invalid Input : identified classes of invalid input must be rejected.
Functions : identified functions must be exercised.
Output : identified classes of application outputs must be exercised.
Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most

other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Alpha Testing

Combination of three testing methods (unit, integration and system testing) forms the alpha testing. Alpha testing is conducted by the development team and quality assurance team. The following things in the application are tested during the alpha testing:

- Spelling Mistakes
- Broken Links
- Cloudy Directions
- Measuring the loading time in the minimum system specifications for optimization.

Beta Testing

Once the alpha testing is completed successfully, the product development team moves onto beta testing. In beta testing, a prerelease version of the software is provided to sample user base over the internet. These early users will use the software and provide a timely feedback about the software working. Following steps are performed in this testing:

- Installation of the software by the users and feedback submission to the development team at regular intervals.
- Noting down the visual errors, generating crash reports.
- Once the feedback is received, development team fixes the issues and once all the known bugs are fixed, a final version is released to the users.
- The nature of the software determines the consumer satisfaction. So, as more and more issues are fixed, nature of the product is increased which results in higher consumer satisfaction.

5.4 ACCURACY TESTING

A. Confusion Matrix

Confusion matrix gives a detail description of classification or misclassification in a form of matrix. It consists of true positive (correctly predict the positive class), true negative (correctly predict the negative class), false positive (incorrectly predict the positive class), false negative (incorrectly predict the negative class).

B. Clasification Accuracy

It is used to measure the performance of our prediction. It can be measure by correct prediction by overall prediction made.

C. Recall

It measures the proportion of actual positive that are correctly identified.

D. Precision

It measure the proposition of positive identification is actually correct.

E. F1 score

F1 score is the average of both precision and recall. In the proposed model for classification of tumour begin malignant or benign the machine learning algorithm used is artificial neural network, Random forest and Support vector machine. In both the feature that is region based and texture based artificial neural network is giving more accuracy. And comparing the accuracy with the proposed model, then it can be seen that accuracy has been increased whereas recall was less. For digital image processing was implemented in matlab R2017a and for classification using machine learning was implemented in jupyter notebook. A comparison between both the features is shown below.

	Accuracy	**Precision**	**Recall**	**F1-Score**
Random Forest	79%	100%	50%	67%
SVM	86%	100%	67%	80%
ANN	92%	100%	69%	81%

TABLE I. REGION BASED FEATURE

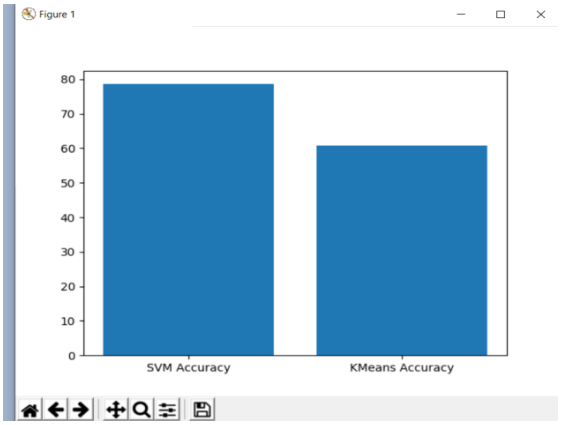


***Fig**. Accuracy Graph*

CHAPTER 6

CONCLUSION

6.1 FUTURE SCOPE AND CONCLUSION:

The proposed model shows the overview of prediction of lung cancer at an early stage. After prediction of the tumour begins malignant or benign, we generate a confusion matrix for each machine learning technique and based on the confusion matrix we calculate accuracy, Recall, precision and F1 score. From the result we can say that our proposed model can distinguish between benign and malignant, and it can be seen that artificial neural network is providing more accuracy in both texture and region based, as well as from the recall value we can say that it has correctly indentified maximum number of malignant tumour In near future deep learning shall outperform machine learning in the field of image classification, object recognition and feature extraction. CNN networks are well known for its features in providing accuracy with higher number of hidden layers in it.

6.2 REFERENCES:

[1] Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." *International Journal of Computer Science and Information Technologies* 4.1 (2013): 39-45.

[2] Zhang, Junjie, et al. "Pulmonary nodule detection in medical images: a survey." *Biomedical Signal Processing and Control* 43 (2018): 138- 147.

[3] Fenwa, Olusayo D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." *Int. J. Comput. Technol.* 15.1 (2016): 6418-6426.

[4] Daoud, Maisa, and Michael Mayo. "A survey of neural network-based cancer prediction models from microarray data." *Artificial intelligence in medicine* (2019).

[5] Palani, D., and K. Venkatalakshmi. "An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification." *Journal of medical systems* 43.2 (2019): 21.

[6] Lynch, Chip M., et al. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." *International journal of medical informatics* 108 (2017): 1-8.

[7] Öztürk, Şaban, and Bayram Akdemir. "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA." *Procedia computer science* 132 (2018): 40-46.

[8] Jin, Xin-Yu, Yu-Chen Zhang, and Qi-Liang Jin. "Pulmonary nodule detection based on CT images using convolution neural network." *2016 9th International symposium on computational intelligence and design (ISCID)*. Vol. 1. IEEE, 2016.

[9] Sumathipala, Yohan, et al. "Machine learning to predict lung nodule biopsy method using CT image features: A pilot study." *Computerized Medical Imaging and Graphics* 71 (2019): 1-8.

[10] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2013.

[11] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[12] L. A. Jemal, R. L. Siegel, and A. Jemal, “Lung cancer statistics,” in *Lung Cancer and Personalized Medicine*, vol. 893, pp. 1–19, Springer, Berlin, Germany, 2016.

[13] C. I. Henschke, D. I. Mccauley, D. F. Yankelevitz et al., “Early lung cancer action project: overall design and findings from baseline screening,” *The Lancet*, vol. 354, no. 9173, pp. 99–105, 1999.

[14] A. K. Alzubaidi, F. B. Sideseq, A. Faeq, and M. Basil, “Computer aided diagnosis in digital pathology application: review and perspective approach in lung cancer classification,” in *Proceedings of the New Trends in Information & Communications Technology Applications*, pp. 219–224, IEEE, Baghdad, Iraq, March 2017.

[15] W. Sun, B. Zheng, and Q. Wei, “Computer aided lung cancer diagnosis with deep learning algorithms,” in *Proceedings of the Medical Imaging: Computer-Aided Diagnosis*, vol. 9785, p. 97850Z, San Diego, CA, USA, March 2016.

[16] K. Kuan, M. Ravaut, G. Manek et al., “Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge,” 2017, https://arxiv.org/abs/1705.09435.

[17] G. Litjens, C. I. Sánchez, N. Timofeeva et al., “Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis,” *Scientific Reports*, vol. 6, no. 1, Article ID 26286, 2016.

[18] J. Wu, T. Aguilera, D. Shultz et al., “Early-stage non-small cell lung cancer: quantitative imaging characteristics of 18F fluorodeoxyglucose PET/CT allow prediction of distant metastasis,” *Radiology*, vol. 281, no. 1, pp. 270–278, 2016.

[19] A. Teramoto, T. Tsukamoto, Y. Kiriyama, and H. Fujita, “Automated classification of lung cancer types from cytological images using deep convolutional neural networks,” *BioMed Research International*, vol. 2017, Article ID 4956063, 9 pages, 2017.

[20] V. A. A. Antonio, N. Ono, A. Saito, T. Sato, M. Altaf-Ul-Amin, and M. Kanaya, “Classification of lung adenocarcinoma transcriptome subtypes from pathological images using deep convolutional networks,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1905–1913, 2018.

[21] S. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez, “Optimal deep learning model for classification of lung cancer on CT images,” *Future Generation Computer Systems*, vol. 92, pp. 374–382, 2019.

[22] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, “Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a

convolutional neural network technique," *Medical Physics*, vol. 43, no. 6, pp. 2821–2827, 2016.

[23] A. Teramoto, M. Tsujimoto, T. Inoue et al., "Automated classification of pulmonary nodules through a retrospective analysis of conventional CT and two-phase PET images in patients undergoing biopsy," *Asia Oceania Journal of Nuclear Medicine Biology*, vol. 7, no. 1, pp. 29–37, 2019.

[24] S. Liang, R. Zhang, D. Liang et al., "Multimodal 3D DenseNet for IDH genotype prediction in gliomas," *Genes*, vol. 9, no. 8, p. 382, 2018.

[25] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*, pp. 212–220, Springer, Berlin, Germany, 2016.

[26] F. Ye, P. Jian, J. Wang, Y. Li, and H. Zha, "Glioma grading based on 3D multimodal convolutional neural network and privileged learning," in *Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine*, pp. 759–763, IEEE, Kansas City, MO, USA, November 2017.

[27] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[28] W. Fei, M. Jiang, Q. Chen et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Honolulu, HI, USA, July 2017.

[29] J. Schlemper, O. Oktay, C. Liang et al., "Attention-gated networks for improving ultrasound scan plane detection," 2018, https://arxiv.org/abs/1804.05338.

[30] M. Al-Shabi, B. L. Lan, W. Y. Chan, K.-H. Ng, and M. Tan, "Lung nodule classification using deep local-global networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 10, pp. 1815–1819, 2019.

[31] L. Gong, S. Jiang, Z. Yang, G. Zhang, and L. Wang, "Automated pulmonary nodule detection in CT images using 3D deep squeeze-and-excitation networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 11, pp. 1969–1979, 2019.

[32] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly supervised deep learning for thoracic

disease classification and localization on chest X-rays," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 103–110, ACM, Washington, DC, USA, 2018.

[33] J. Shuiwang, Y. Ming, and Y. Kai, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, pp. 221–231, 2013.

[34] T. Jin, C. Hui, Z. Shan, and X. Wang, "Learning deep spatial lung features by 3D convolutional neural network for early cancer detection," in *Proceedings of the International Conference on Digital Image Computing: Techniques & Applications*, pp. 1–6, Sydney, Australia, November 2017.

[35] R. Dey, Z. Lu, and Y. Hong, "Diagnostic classification of lung nodules using 3D neural networks," in *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 774–778, IEEE, Washington, DC, USA, April 2018.

[36] H. Gao, L. Zhuang, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[37] J. D. Fauw, J. R. Ledsam, B. Romeraparedes et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[38] W. Guo, Z. Xu, and H. Zhang, "Interstitial lung disease classification using improved DenseNet," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30615–30626, 2019.

[39] J. Arevalo, T. Solorio, M. Montes-Y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proceedings of the ICLR (Workshop)*, Toulon, France, April 2017.

[40] C. Y. Lee, S. Xie, P. Gallagher et al., "Deeply-supervised nets," in *Proceedings of the Artificial Intelligence and Statistics*, pp. 562–570, San Diego, CA, USA, May 2015.

[41] X. Zhang, Y. Zou, and S. Wei, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proceedings of the International Conference on Digital Signal Processing*, pp. 1–5, IEEE, London, UK, August 2017.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, Seoul, South Korea,

October 2019.

APPENDIX

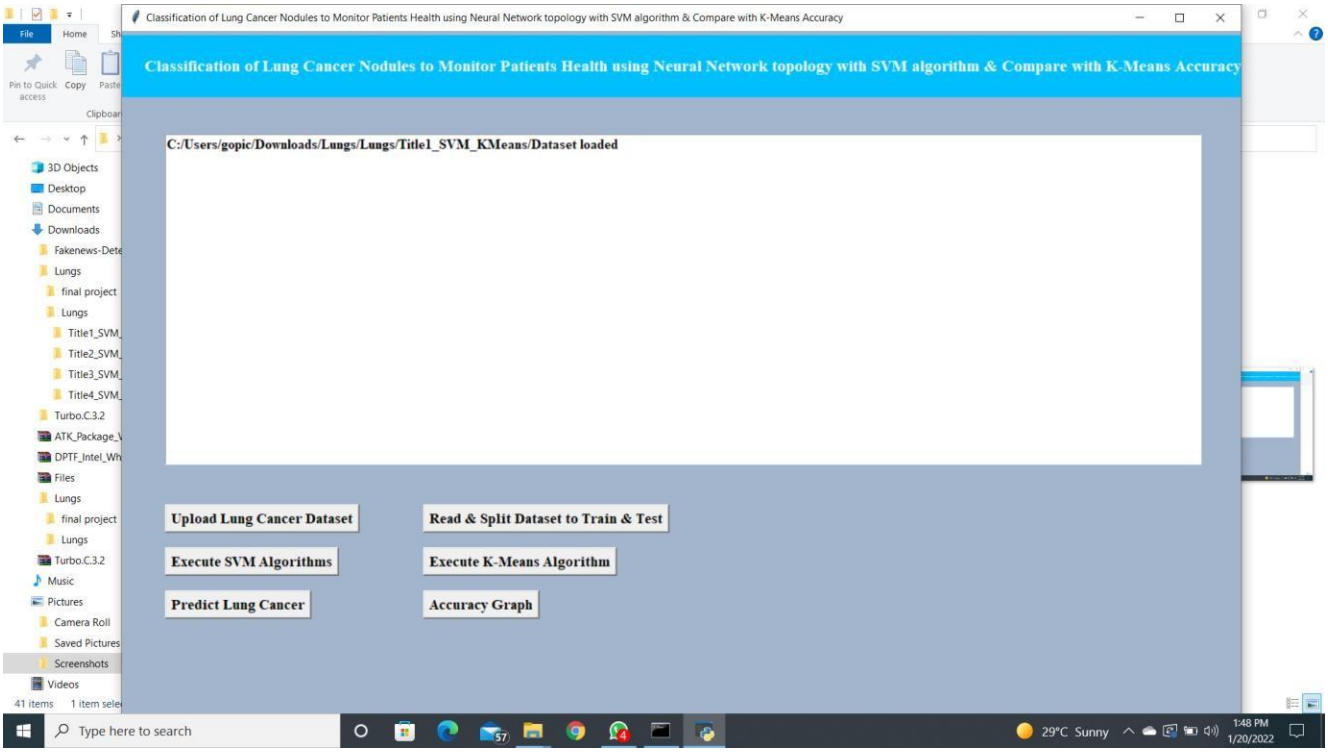
A.SOURCE CODE:

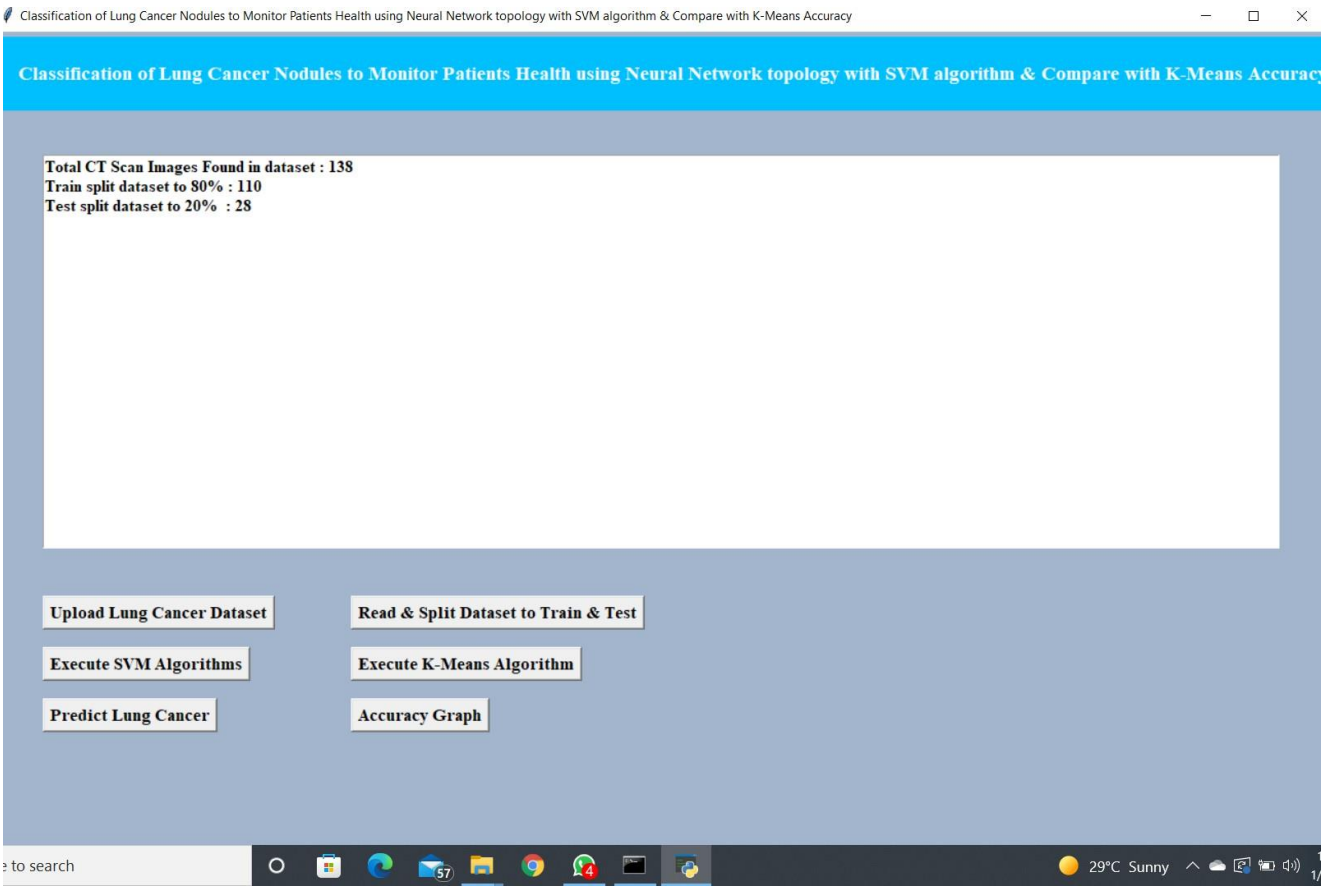
```
#!/usr/bin/env python
# coding: utf-8
# In[65]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# In[66]:
df=pd.read_csv('heart.csv')
# In[67]:
df
# In[68]:
df.shape
# In[69]:
df.columns
# In[70]:
df.describe()
# In[71]:
sns.heatmap(df.isnull())
#
# In[72]:
df.columns
# In[74]:
sns.countplot(df['target'])
# In[63]:
for i in df.columns:
    print(i,df[i].skew())
# In[64]:
df['target'].skew()
```

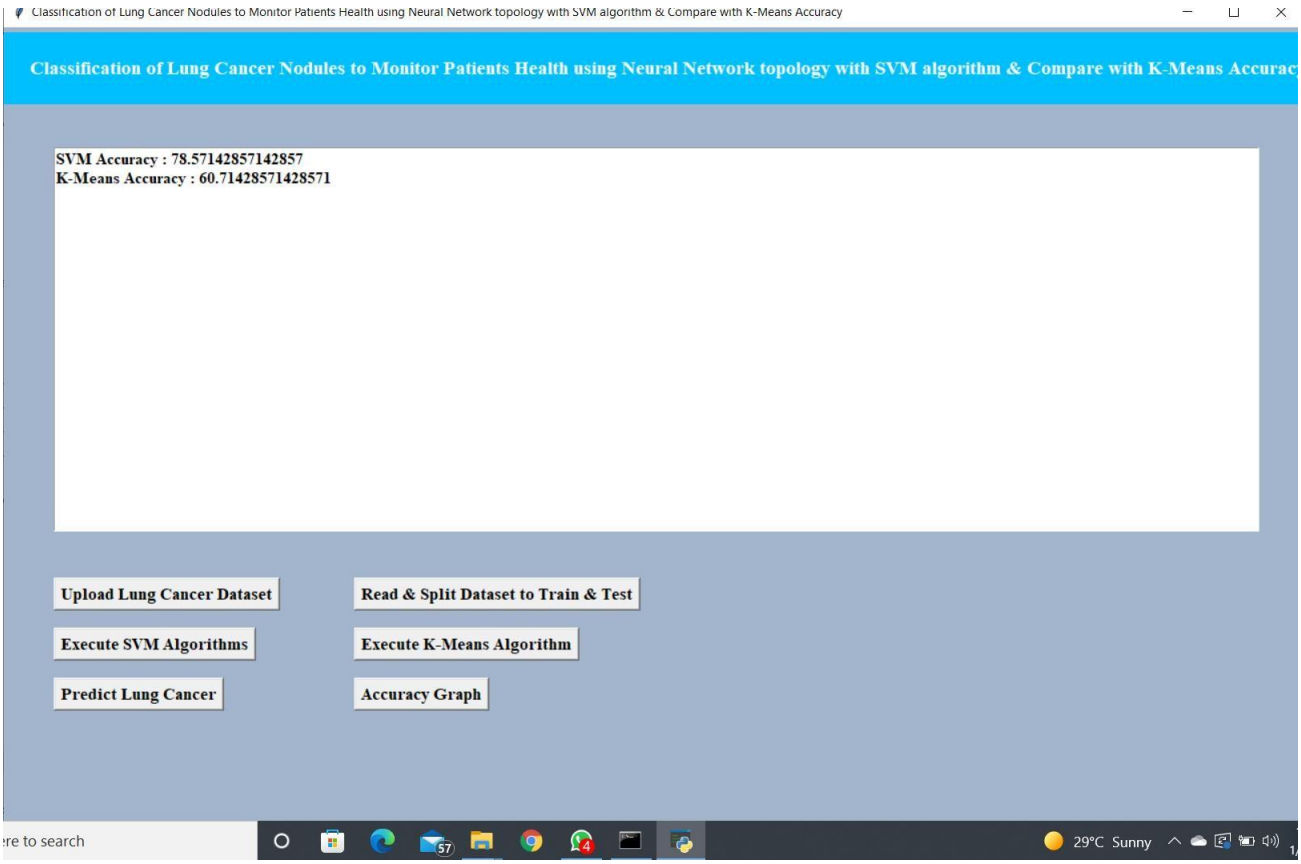
```
# In[8]:
df.info()
# In[9]:
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
scaler.fit_transform(df)
# In[45]:
df.corr()
# In[10]:
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
# In[11]:
from sklearn.model_selection import train_test_split
target=df['target']
df=df.drop(['target'],axis=1)
df.columns
# In[12]:
x_train,x_test,y_train,y_test=train_test_split(df,target,test_size=0.2)
print(y_test)
# In[13]:
rfc.fit(x_train,y_train)
# In[14]:
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
prediction=rfc.predict(x_test)
# In[34]:
accuracy_score(y_test,prediction)
# In[35]:
#Randomized Search Cv
# no of trees in random forest
n_estimators=[int(x) for x in np.linspace(100,1200,12)]
#no of features to consider at every split
max_features=['auto','sqrt']
#maximum no of levels in a tree
max_depth=[int(x) for x in np.linspace(5,30,6)]
#minimum no of samples to split at every node
```

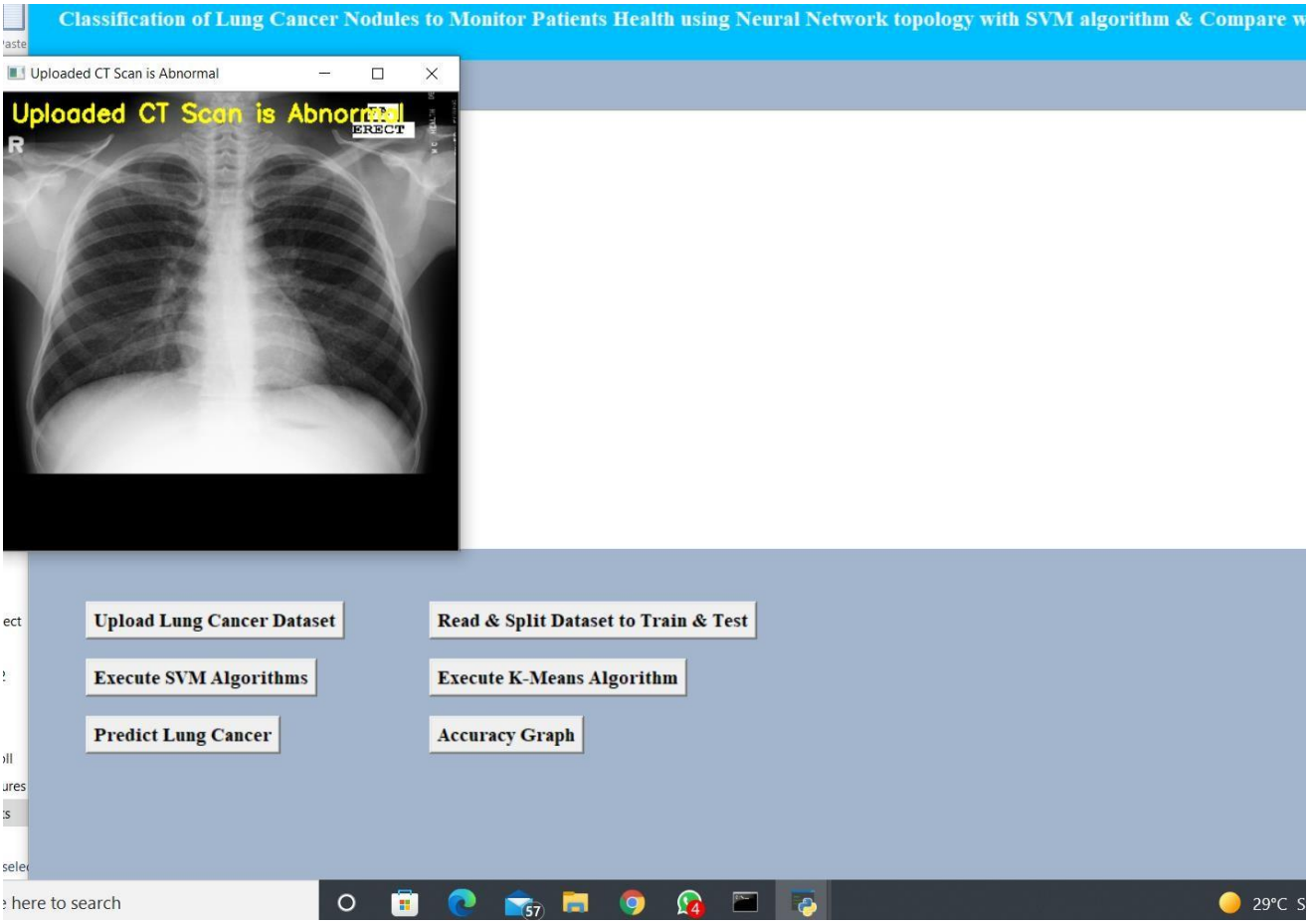
```
min_samples_split=[2,5,10,15,100]
#minimum no of samples required at each leaf node
min_samples_leaf=[1,2,5,10]
# In[36]:
from sklearn.model_selection import RandomizedSearchCV
# In[37]:
random_grid={'n_estimators':n_estimators,'max_features':max_features,'max_depth':max_depth,'
min_samples_split':min_samples_split,'min_samples_leaf':min_samples_leaf}
# In[39]:
rf=RandomizedSearchCV(estimator=rfc,param_distributions=random_grid,scoring='neg_mean_sq
uared_error',n_iter=10,cv=5,verbose=2,random_state=42,n_jobs=1)
# In[40]:
rf.fit(x_train,y_train)
# In[41]:
predictions=rf.predict(x_test)
# In[75]:
sns.countplot(predictions)
# In[42]:
accuracy_score(y_test,predictions)
# In[48]:
from sklearn.ensemble import GradientBoostingClassifier
gbr=GradientBoostingClassifier(n_estimators=3000,learning_rate=0.05)
# In[49]:
gbr.fit(x_train,y_train)
predictions1=gbr.predict(x_test)
# In[50]:
accuracy_score(y_test,predictions1)
```

B.OUTPUT SCREENSHOTS




Classification of Lung Cancer Nodules to Monitor Patients Health using Neural Network topology with SVM algorithm & Compare with K-Means Accuracy
Classification of Lung Cancer Nodules to Monitor Patients Health using Neural Network topology with SVM algorithm & Compare with K-Means Accuracy
Total CT Scan Images Found in dataset : 138
Train split dataset to 80% : 110
Test split dataset to 20% : 28
Upload Lung Cancer Dataset
Read & Split Dataset to Train & Test
Execute SVM Algorithms
Execute K-Means Algorithm
Predict Lung Cancer
Accuracy Graph
to search
29°C Sunny


Classification of Lung Cancer Nodules to Monitor Patients Health using Neural Network topology with SVM algorithm & Compare with K-Means Accuracy
Classification of Lung Cancer Nodules to Monitor Patients Health using Neural Network topology with SVM algorithm & Compare with K-Means Accurac
SVM Accuracy : 78.57142857142857
K-Means Accuracy : 60.71428571428571
Upload Lung Cancer Dataset
Read & Split Dataset to Train & Test
Execute SVM Algorithms
Execute K-Means Algorithm
Predict Lung Cancer
Accuracy Graph
re to search
29°C Sunny


Classification of Lung Cancer Nodules to Monitor Patients Health using Neural Network topology with SVM algorithm & Compare w
Uploaded CT Scan is Abnormal
Uploaded CT Scan is Abnormal
ERECT
Upload Lung Cancer Dataset
Read & Split Dataset to Train & Test
Execute SVM Algorithms
Execute K-Means Algorithm
Predict Lung Cancer
Accuracy Graph
here to search
29°C

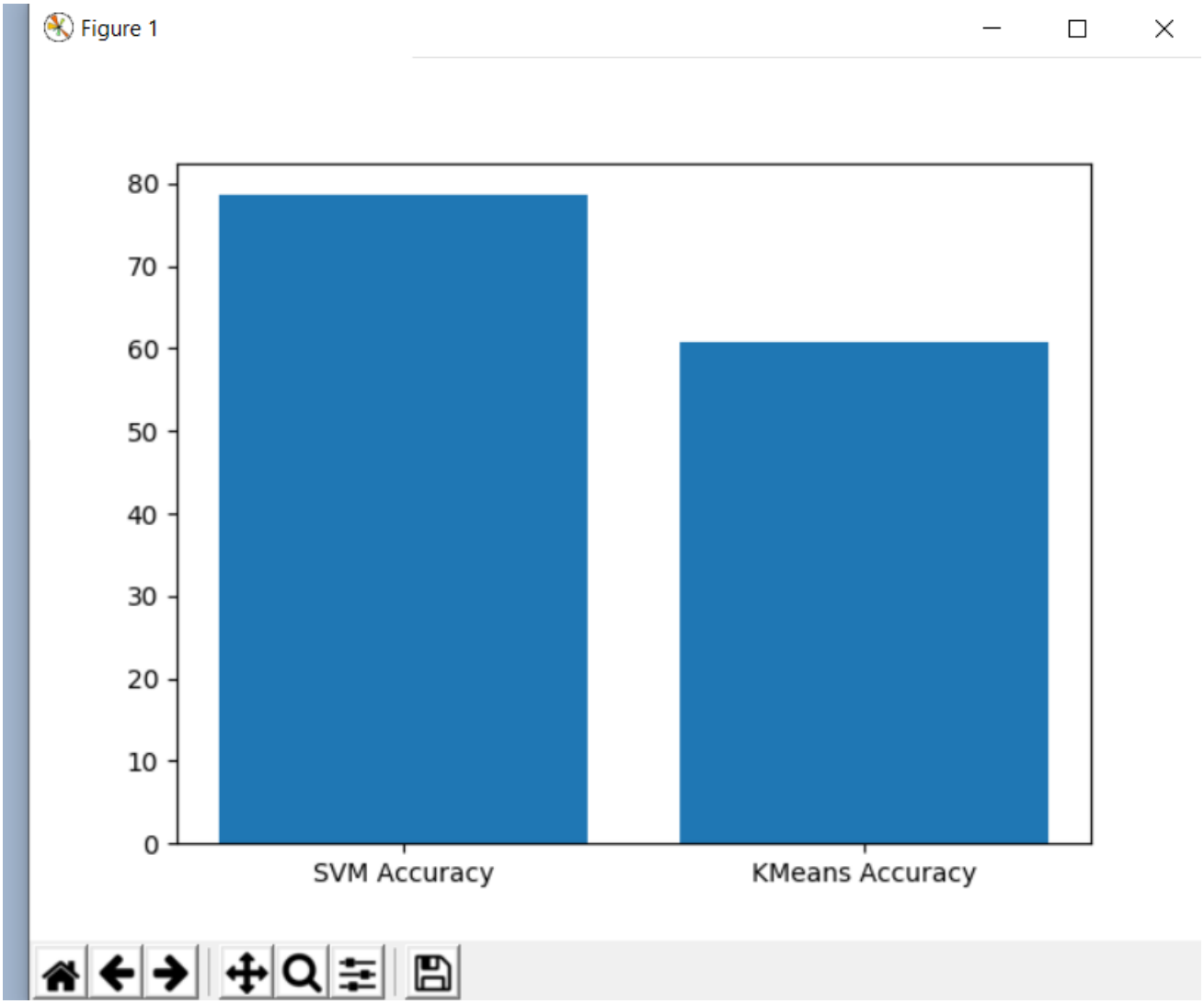


Figure 1
80
70
60
50
40
30
20
10
0
SVM Accuracy
KMeans Accuracy