**ACKNOWLEDGEMENT**

“Ability and ambition are not enough for success. Many able persons have failed to achieve anything worthwhile because of lack of guidance and direction. Success of any project depends greatly on support, guidance and encouragement received from the guide".

We have been fortunate to have more than one pillar of strength in our humble effort to make this project successful.

It give us great pleasure to express our deep sense of gratitude towards our project guide Dr. Archana Bhise for their resourceful and able guidance which lead to timely completion of this report. It was really insight and obsession for innovative ideas that motivated us to consider our idea seriously. We also managed to learn quite few things from them which will definitely help us in future. We sincerely thank them for this kind cooperation and extreme patience that they have shown.

We feel obliged for the amount of faith and confidence they showed in us and it was their presence that kept us going on. We are very thankful to Dr. Manoj Sankhe, Head of Department and Prof. Kanchan Bakade, Project coordinator for providing all the necessary facilities and support.

We would also like to thank the entire teaching and non-teaching staff of EXTC department, who extended their kind cooperation. Last but not the least; we would like to thank our family and friends for their constant support.

Sneha Ramchandra Potghan

M. Tech (EXTC)

Roll No.:- C006

SAP No : 70231016006

**ABSTRACT**

Cancer is a class of diseases characterized by uncontrolled growth of the tissues. There are over 100 different types of cancer, and each is classified by the type of cell that is initially affected. Cancer harms the body when altered cells divide uncontrollably to form lumps or masses of tissue called tumours.

Tumours can grow and interfere with the respiratory, digestive, nervous, and circulatory systems and they can release hormones that alter body function. There are various medical imaging modalities like X-ray, chest-radiography, CT, SPECT, NM, MRI, CT and PET. And, Optical Modalities like Endoscopy, Microscopy or Photography exists to identify the presence of disease. CT image based lung nodule detection is the most widely used and accepted method for detecting lung cancer. Automated Computer Aided Diagnosing (CAD) systems are more useful tool for advanced decision making in radiology.

The Proposed work is to detection and classification of lung tumor from DICOM CT images using machine learning approaches. The proposed system consists of many steps such as image database collection, normalization, segmentation, feature extraction, classification and 3D volume.

At first, lung DICOM CT images are sent as input to the system. In second stage, segmentation is performed to segment the lung CT image using thresholding, morphology, K-mean clustering, Contour and a strong feature extraction method has been introduced to extract the some important feature of segmented images i.e. ASM, Contrast, Mean, Standard deviation, Homogeneity, Dissimilarity . Extracted features are used to trained and tested by using K-Nearest Neighbor (KNN) and Multi-Layer Perceptron (MLP) classifier and finally the system is tested for both tumorous and non-tumorous images. The proposed approach achieved the classification accuracy of 98.30% using KNN and 98.31% using MLP. If CT is tumorous, 3D volume is generated

**Keywords:** Medical Images Processing, Feature Extraction, Multilayer Perceptron MLP, KNN, 3D volume.

# Table of Contents

[Table of Contents i](#_Toc511654618)

[List of Figures iii](#_Toc511654619)

[List of Tables iii](#_Toc511654620)

[Chapter 1 Introduction 1](#_Toc511654621)

[1.1. Overview 1](#_Toc511654622)

[1.2. Scope of the Project 2](#_Toc511654623)

[1.3. Report Organization 3](#_Toc511654624)

[1.4. About DICOM 3](#_Toc511654625)

[Chapter 2 Literature Survey 5](#_Toc511654626)

[2.1. Detailed Survey 5](#_Toc511654627)

[2.2. Summary 11](#_Toc511654628)

[Chapter 3 Problem Statement 14](#_Toc511654629)

[3.1. Problem Statement 14](#_Toc511654630)

[3.2. Objective of the Project 14](#_Toc511654631)

[Chapter 4 Methodology 15](#_Toc511654632)

[4.1. System Overview 15](#_Toc511654633)

[Chapter 5 Software Requirement 28](#_Toc511654634)

[5.1. DICOM Viewer 28](#_Toc511654635)

[5.2. Python 28](#_Toc511654636)

[Chapter 6 Implementation and Results 32](#_Toc511654637)

[6.1. Process flow 32](#_Toc511654638)

[6.2. Implementation Steps 33](#_Toc511654639)

[6.3. Experimental Results 33](#_Toc511654640)

[6.4. Graphics User Interface Result : 39](#_Toc511654641)

[Conclusion and Future Scope 41](#_Toc511654642)

[Conclusion 41](#_Toc511654643)

[Future Scope 41](#_Toc511654644)

[References 42](#_Toc511654645)

[Publication 45](#_Toc511654646)

[Appendix 46](#_Toc511654647)

# List of Figures

[Figure 4‑1 : System Flow 15](#_Toc511654648)

[Figure 4‑2 : Sample Image 16](#_Toc511654649)

[Figure 4‑3 : Segmentation Flow 17](#_Toc511654650)

[Figure 4‑4 : Layered diagram of Multi-Layer Perceptron NN 23](#_Toc511654651)

[Figure 4‑5 : KNN Basic Diagram 24](#_Toc511654652)

[Figure 6‑1 : Process Flow 32](#_Toc511654653)

[Figure 6‑2 : Sample Lung Original Image 34](#_Toc511654654)

[Figure 6‑3 : Original Image with Tumor 34](#_Toc511654655)

[Figure 6‑4 : Morphology Opening Image 34](#_Toc511654656)

[Figure 6‑5 : Threshold Image 35](#_Toc511654657)

[Figure 6‑6 : Lung Mask 35](#_Toc511654658)

[Figure 6‑7 : Mask \* Original Image 35](#_Toc511654659)

[Figure 6‑8 : K-mean clustering and Contour Detection 35](#_Toc511654660)

[Figure 6‑9 : 3-D view of lung tumor 37](#_Toc511654661)

[Figure 6‑10 : 3-D view of lung in Grayscale 37](#_Toc511654662)

[Figure 6‑11 : 3-D view of lung with tumor in blue scale 38](#_Toc511654663)

[Figure 6‑12 : GUI 39](#_Toc511654664)

[Figure 6‑13 Play Images in DICOM Directory 39](#_Toc511654665)

[Figure 6‑14 : GUI Segmentation Display 40](#_Toc511654666)

# List of Tables

[Table 2‑1 : Literature Survey Summary 13](#_Toc511653819)

[Table 4‑1 : Parameter used to train MLP 23](#_Toc511653820)

[Table 6‑1 : Database Distribution 33](#_Toc511653821)

[Table 6‑2 : Values of Feature Extraction 36](#_Toc511653822)

[Table 6‑3 : Classifier Accuracy 36](#_Toc511653823)

# Introduction

Overview

Oncology is a branch of medicine that deals with the prevention, diagnosis, and treatment of cancer. A medical professional who practices oncology is an oncologist. The name's etymological origin is the Greek word (onkos), meaning "tumor", "volume" or "mass" and the word (logos), meaning "speech".

The three components which have improved survival in tumor are:

1. Prevention by reduction of risk factors like tobacco and alcohol consumption.
2. Early diagnosis screening of common cancers and comprehensive diagnosis and staging.
3. Treatment multi-modality management by discussion in tumor board and treatment in a comprehensive cancer centre.

Cancer refers to the abnormal growth of cells anywhere in the body; which tends to proliferate in an uncontrolled way [1]. Many cancers and the abnormal cells which compose it are further identified by the name of the tissue that the abnormal cells originated from, for example, breast cancer, lung cancer, colon cancer, prostate cancer and so on. Lung cancer is a leading cause of death worldwide [2].

Lung cancer refers to the uncontrolled growth of abnormal cells in the lung. Typically, a computed tomography (CT) scan of the thorax is the most sensitive method for detecting lung nodules and the surrounding structures. A CT scan is a painless, non-invasive diagnostic imaging procedure which creates precise multiple images (slices) of the body structures, such as the lungs [3]. The cross-sectional images generated during a CT scan can be reformatted in multiple planes, and can generate 3D images. The national lung screening trial (NLST) has shown a relative risk reduction in lung cancer-specific mortality of 20% and 6.7% in all-cause mortality using low dose CT screening [4].

A lung nodule is a round lesion with a diameter smaller than 3 cm. It can be either benign (non-cancerous) or malignant (cancerous), and is found in 1 of each 100 CT scans of the chest. In a CT scan, the lung cancer is observed as round white shadow nodules, therefore it is important to detect and classify those nodules for the screening and diagnosis purposes.

The likelihood that a nodule can be cancerous is about 40%; however, the risk varies considerably depending upon several factors. For example, in people with age less than 35 years, the chance that a lung nodule can be cancerous is minor than 1%, whereas the half of lung nodules in people over 50 are malignant (cancerous) [5]. When a nodule is detected on a CT scan, the radiologists must compare the current CT scan with the previous ones. If the nodule on earlier CT scans has not changed in size, shape or appearance, it is probably non-cancerous. If a lung nodule is new or has changed in size, shape or appearance, then a bronchoscopy or tissue biopsy is recommended to determine if it is cancerous.

The possibility to obtain an accurate interpretation from CT scans demands a big effort by the specialists, due to the large number of scans that are often managed and analysed. The analysis becomes more complex when the progress of the disease is still not visually significant (early stage). For the radiologist, the process of examining a CT scan to detect lung nodules takes between 15 and 20 minutes. On the same day, the radiologist typically analyses, at least, 45 images and this might be a fatiguing process. Therefore, different diagnosis results can be obtained by different specialists for the same scan and 3D volume is calculated manually from the tumor. This gives the idea about volume of tumor which in turn helps for treatment planning system.

Scope of the Project

The research scope is geared towards the limits of the research such as:

1. The medical images are store in DICOM format, so the DICOM image format has been studied.

2. The texture feature extraction techniques are studied.

3. Python programming is used to get analysis and do comparison.

4. The evaluation is measured based on the performance parameter such as accuracy.

5. The findings and conclusion made in this study is based on experiments done on dataset. Clinical and standard LIDC database is used for experiment.

6. The proposed Neural Network classification system is developed in Python.

7. 3D volume of lung tumor is developed.

Report Organization

This research covers seven chapters, which are introduction, literature survey, problem statement, methodology, system analysis, software description, testing and results, discussion and conclusion.

Chapter 1 of this report consists of overview of the study, objective, scope of the project.

Chapter 2 of this report presents a review of the literature related to study. This literature covers various issues and challenges involved in the process of lung cancer classification. In this approach discussed the different techniques, algorithm and classifiers which classify the lung tumor CT.

Chapter 3 consists of problem statement and objective of the project.

Chapter 4 consists of wide description on research methodology, which provides a rich discussion about the flow of this research. This includes how the operational and experimental work has been carried out for the study.

Chapter 5 consists of Software Description, gives the idea about Python software.

Chapter 6 consists of proposed implementation of system and details about the process flow overall result and the analysis of proposed system is displayed and discussed.

About DICOM

DICOM means for Digital Imaging and Communications in Medicine. It is a backbone of modern image display. It represents imaging workflow in universal and fundamental standard of digital image examination. It is not just an impression or a tool. It provides multilane reconstruction for accurate diagnosis of medical imaging observations. It is a new digital modality to transfer and store data and to design a protocol to cover all the functional aspects of digital medicine that has highest diagnostic standards and best performance. DICOM in Radiology is a strong and efficient radiology practice in health care information technology domain.

DICOM also integrates medical imaging into a complete enterprise wide electronic patient record solution with interoperability and cost effectiveness. DICOM collects the original digital medical images. DICOM has the capability to share and store images regardless of machine or method. DICOM operates on image visibility and concentrates on changes. DICOM provides much ease and flexibility in work, with fewer efforts. It stands as universal standard of digital medicine having excellent Image quality. It supports up to 65,536 shades of grey for monochrome image display. DICOM provides full support for numerous images in a certain time.

# Literature Survey

Detailed Survey

As lung cancer is a major health issues in the 21st century, numerous studies are done on the lung cancer related to cause of lung cancer, its statistics, detection of lung cancer its classification etc. Some of those techniques are presented as below,

1. Gupta, O. Martens, Y. Le Moullec, and T. Saar (2015):

In paper “A Tool for Lung Nodules Analysis based on Segmentation and Morphological Operation” a novel, simple and efficient tool has been developed to support lung cancer oriented educational research, which allows the visualization and segmentation of marked nodules through Greyscale Thresholding and erosion morphological operations on the Lung Image Database Consortium (LIDC) based CT images.

The tool is also capable of providing mathematical calculations of nodules such as area, perimeter and nodule centroid; but further studies are still required for enhanced calculative analysis likewise roundness, internal structure, orientation and convex area. Eventually, this tool will benefit the lung cancer research and educational fields [6]. In future, the additional functionality of recognition of benign and malignant nodules along with 3D nodule segmentation will be introduced for accurate estimation of nodules.

1. M. B. A. Miah and M. A. Yousuf (2015):

“Detection of Lung Cancer from CT Image Using Image Processing and Neural Network” paper describes lung cancer as one kind of dangerous diseases, so it is necessary to detect it in early stages. But the detection of lung cancer is the most difficult task. From the literature review, many techniques are used for the detection of lung cancer but they have some limitations.

In this, the proposed method pursues approaches in which first step is Binary Thresholding and then feature extraction and then these features are used to train up the neural network and test the neural network. The proposed system successfully detects the lung cancer from CT scan images. At the end, the system can say that the system has achieved its desired expectation. The proposed system test 150 types of lung CT images and obtains the result where overall success rate of the system is 96.67%, which meet the expectation of system [7]. In future this technique can be used in the detection of brain tumor, breast cancer etc.

1. R. Sammouda (2016):

In “Segmentation and Analysis of CT Chest Images for Early Lung Cancer Detection” paper, they have achieved the purpose of developing an automatic CAD system for early detection of lung cancer, by analysing human lung CT images using several phases. The approach starts by extracting the lung regions from the CT image using classical image processing techniques, including bit-planes representation of raw 3D-CT images producing 2D slices. They have applied the procedures viz. Erosion, Median filter, Dilation, Outlining, Lung Border Extraction and Flood Fill algorithm, in sequence. The results are given as very sharp borders and filled lung regions with raw CT data without any modification to their pixel's values [8].

After the extraction step, the extracted lung regions are segmented using an unsupervised modified Hopfield Neural Network Classifier. The segmented lung regions are grouped into very similar clusters giving more understanding of the different tissues in the lung regions. Then, the initial lung candidate nodules, resulting from the segmentation process, are used in the nodules detection process. Finally, we apply three main filters to extract the true lung cancer regions with respect to the information formulated from experimented radiologists in the field. The detection process is very promising and gives candidate regions to be used as input for a next diagnosis process to distinguish between benign and malignant cancer cases as our future works.

1. M. Alam, G. Sankaranarayanan, and V. Devarajan (2016):

In “Lung Nodule Detection and Segmentation Using a Patch-Based Multi-Atlas Method” paper, an atlas based lung nodule selection process is designed and implemented. This work is the first to use a patch-based, multi-atlas method in lung nodule selection. In this process, a new size and shape based feature vector is designed and used for atlas selection [9]. This method is tested on 5 different patients and proved to work in all cases. This method is implemented for 2D images. The next step is to extend this method to 3D.

1. V. Kalpana and G. Rajini (2016):

In “Segmentation of Lung Lesion Nodules using DICOM with Structuring Elements and Noise” - A comparative study paper describes that the rate of lung cancer progression is due to the cell type variation and the deaths due to poor prognosis because of several diagnostic problems. For the detection of earliest stage tumours and effectiveness of screening the considerable interest lies on the recommendation of reliably improved modalities and techniques. Better screening interventions and diagnostic procedures for early lung cancer are developed by considering the DICOM image susceptible with Gaussian, salt & pepper, speckle and Poisson Noises using Disk and Diamond varying from sizes 1 to 4.

Diagnostic algorithms can also assess abnormality size, growth and descriptive characteristics. The analysis is made by taking correlation factor & PSNR [10]. Gaussian operator proves the best with Gaussian noise and Sobel operator proves the best with salt and pepper noise for disk and diamond with sizes 1 to 4. Sobel operator also exhibits exact results with all noises except for Diamond 4. Poisson Noise shows similar result with average detector for disk and diamond with size 2 & 4.Sobel can be partially considered when it is used with Poisson Noise.

The epidemiology of cancer can be significantly reported by comparing the results taken from these approaches. The examinations of patterns support the diagnosis and treatment without delay will control mortality. From this perspective, assessments can be further done using several operators and various structuring elements like square, Rectangle, Line, Pair, Ball, Octagon and Arbitrary and contribute to enhance the survival. These results will have high impact at the discretion of the radiologist to evaluate the indeterminate lesions. To ensure safety and quality treatment, the results obtained are very crucial, as the measures to be taken will generate the advancement of the disease for the oncologist. Environment with different structuring elements and sizes will be helpful to know extent of the content of the image affected by the noise.

1. G. P. Pratap and R. Chauhan (2016):

“Detection of Lung Cancer Cells using Image Processing Techniques” paper describes two stages:

1. Processing of distortion input image utilizing filter and segmentation
2. Morphological operations on CT picture.

The growth influenced lungs locale can be seen in the last algorithm process for particular CT information image [11]. The proposed strategy can likewise be connected to identify some other malignancy like breast cancer, skin malignancy and so forth. Also it finds place in medical research as well.

1. Ada, Rajneet Kaur (2013):

In this paper uses a computational procedure that sorts the images into groups according to their similarities. In this paper, Histogram Equalization is used for pre-processing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that we predict the survival rate of a patient by extracted features. Experimental analysis is made with dataset to evaluate the performance of the different classifiers.

The performance is based on the correct and incorrect classification of the classifier. In this paper, Neural Network Algorithm is implemented using open source and its performance is compared to other classification algorithms. It shows the best results with highest TP Rate and lowest FP Rate and in case of correctly classification, it gives the 96.04% result as compare to other classifiers. The second paper of this same author is based on Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images. In this paper, a hybrid technique based-on feature extraction and Principal Component Analysis (PCA) is mentioned [12].

1. Fatma Taher, Naoufel Werghi1, Hussain Al-Ahmad1, Rachid Sammouda (2012):

This paper presents two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. However, the extreme variation in the gray level and the relative contrast among the images make the segmentation results less accurate, thus we applied a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature.

The thresholding algorithm succeeded in extracting the nuclei and cytoplasm regions. Moreover, it succeeded in determining the best range of thresholding values. The HNN and FCM methods are designed to classify the image of N pixels among M classes. In this study, we used 1000 sputum color images to test both methods, and HNN has shown a better classification result than FCM, the HNN succeeded in extracting the nuclei and cytoplasm regions.

In this paper authors uses a rule based thresholding classifier as a pre-processing step. The thresholding classifier is succeeded in solving the problem of intensity variation and in detecting the nuclei and cytoplasm regions, it has the ability to mask all the debris cells and to determine the best rang of threshold values. Overall, the thresholding classifier has achieved a good accuracy of 98%, with a high value of sensitivity and specificity of 83% and 99% respectively [13].

1. Jennifer Cabrera et.al. (2015)

Developed lung cancer classification tools that rely on Support Vector Machines to classify lung cancer data based on oligonucleotide microarrays. It is provides pre-processing methods to remove possible non biological factors and biases in the data. It also provides gene selection techniques to select the most relevant genes and reduced the extremely large size of the microarray dataset. It computes which genes are markers or most attributable to a certain class of lung cancer. Finally, it presents its output and results in a tabular form and provides as much information about the data as possible. The SNR statistics used in this tool only involves binary identification and comparison of genes between samples, hence the more scattered results [14].

1. S. kanitkar et.al.(2015)

The proposed a system which is having stages such as pre-processing stage, segmentation stage, feature extraction stage and classification. For smoothing, Gaussian filter is applied on the input image because Gaussian smoothing is very effective for removing noise, it removes high frequency components from the image. Gabor function is used for image enhancement. Watershed segmentation is used to extract the region minimum value from an image. It determines the corresponding to the dividing line with the least value. Watershed gives 90 % accuracy compared to the thresholding algorithm. So it is efficient for segmentation [15].

1. Elmar Rendon-Gonzalez et al. (2016)

The proposed a CAD system for lung cancer detection. This system is based on four main steps: pre-processing, lungs parenchyma segmentation, nodule detection and reduction of False Positives (FP). In pre-processing steps, masks are created by thresholding technique and enhanced by morphological operations. A priori information and Hounsfield Units (HU) is used to calculate suspicious region of interest. The four features are extracted from lungs ROI: Area, eccentricity, circularity and fractal dimensions and extracted features are classified using SVM classifier. This approach achieved 78.08% accuracy. In future, this approach can be extending for 3D feature extraction to obtained better features [16].

1. Ramandeep Kaur.(2013)

Proposed a new classifier which utilizes MLP approach i.e. Multi-Layer Perceptron Nearest Neighbor (MLP-NN). Multi-Layer Perceptron nearest neighbor perform better in detecting the correct disease. MLP-NN gives 87 % accuracy in 100 different lung image databases. This approach handles noisy data and reduced complexity. Proposed approach feature extraction is conducted by MAD with Gabor filter [17].

1. Harikumar Rajaguru.(2016)

The presents a comparison of oral cancer classification with Multi-Layer perceptron and Gaussian mixture measure. Multi-Layer perceptron feed forward neural network. Feature for the classifier is tumor size, Node, Metasis. This paper compares the classification accuracy of the TNM (Tumor, Node, Metasis) staging system with the aid of Multi-Layer Perceptron (MLP) and Gaussian Mixture Model (GMM) classifiers. MLP gives 89.5 % accuracy for all stages [18].

Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr.**  **No.** | **Author** | **Publications/**  **Year** | **Methodology** | **Remarks** |
|  | Anindya Gupta, Olev Martens, Yannick Le Moullec & Tonis Saar | Intelligent Signal Processing (WISP), IEEE 9th International Symposium 2015 | Greyscale Thresholding and morphological operation. | 1. Introduced 3D visualization. |
|  | Md. Badrul Alam Miah & Mohammad Abu Yousuf | Electrical Engineering and Information Communication Technology (ICEEICT), International Conference on. IEEE 2015 | Segmentation and neural network detection. | 1. Tested cancerous non-cancerous images. |
|  | Rachid Sammouda | Computer & Information Technology (GSCIT), Global Summit on. IEEE 2016 | Hopfield artificial neural network. | 1. Got very sharp lung borders |
|  | Mustafa Alam, Ganesh Sankaranarayanan  & Venkat Devarajan | Computational Science and Computational Intelligence (CSCI), International Conference on IEEE 2016 | Patch-based multi-atlas method. | 1. K-NN classification for atlas selection 2. LOG for post-processing 3. 95% sensitivity |
|  | V. Kalpana & G. K. Rajini | Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference | Morphological analysis, contour detection and watershed algorithm. | 1. Used DICOM CT images 2. Analysis made by correlation factor and PSNR |
|  | G. P. Pratap and R. Chauhan | Power Electronics, Intelligent Control and Energy Systems (ICPEICES), IEEE International Conference 2016 | Classify into 2 stages malignant and benin cancer. | 1. Segmentation is done by using morphology operator. |
|  | Ada, Rajneet Kaur | (IJAIEM), Issue 6th June 2013 | Histogram equalization is used for pre-processing purpose. Different classifier is used for comparison purpose. | 1. PCA is used Accuracy is 96.03 %. 2. TP, FP is calculated. |
|  | Fatma Taher, Naoufel Werghi1 | GCC Conference and Exhibition (GCC), 2011 IEEE | Segmentation algorithm -Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm. Thresholding classifier is used for classification purpose. | 1. Good accuracy of 98%, with a high value of sensitivity and specificity of 83% and 99% respectively. |
|  | Cabrera, Jennifer | In Information, Intelligence, Systems and Applications (IISA), 6th International Conference on, IEEE, 2015. | Microarray and SVM classification is used | 1. SNR,PSNR is calculated . |
|  | S. kanitkar | In Pervasive Computing (ICPC), International Conference on, IEEE, 2015. | Gabour filter and Watershed Segmentation | 1. Classification is used 90% accuracy. |
|  | Elmar Rendon-Gonzalez et al. | 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW), Kharkiv, 2016. | Proposed a CAD system. The four features are extracted from lungs ROI: Area, eccentricity, circularity and fractal dimensions and extracted features are classified using SVM classifier. | 1. Accuracy of SVM classifier is 78.08 %. |
|  | Ramandeep Kaur. | International Journal of Computer Applications. 2013 | Proposed a new classifier which is a combination of K Nearest Neighbour and Multi-Layer perceptron (MLP-NN).  For Feature extraction Gabor filter is used. | 1. MLP-NN gives accuracy of 87 %. |
|  | Harikumar Rajaguru. | [The 16th International Conference on Biomedical Engineering](https://link.springer.com/book/10.1007/978-981-10-4220-1), 2016 | Used Multi-Layer Perceptron based feed forward back propagation classifier. Feature for the classifier is tumor size, Node, Metasis. This paper compares the classification accuracy of the TNM (Tumor, Node, Metasis) | 1. Accuracy of MLP is 89.5 % for all stages. |

Table ‑ : Literature Survey Summary

# Problem Statement

Problem Statement

* Detection of lung tumor from DICOM CT images using image processing technology.
* Design a Neural Network to classify tumorous and non-tumorous modules.
* Generate 3D volume of tumor.

Objective of the Project

The Objective of the Products is as below,

1. To explore, collect and select the appropriate lung CT dataset for lung tumor detection.
2. To pre-process the data to make it suitable for further processing.
3. To implement feature extraction algorithm like GLCM.
4. To implement the existing algorithms to detect and classify the lung tumor.
5. To use an effective classifier, to improve the performance.
6. To calculate, validate and verify the performance of the effective classifier.
7. Generate 3D volume of tumor.

For the radiologist, lung nodules are usually, accidentally, detected in a CT scan, because they are not big enough to be seen easily. In this project, the nodules are characterized by the computation of the texture features obtained from the gray level co-occurrence matrix (GLCM) in the wavelet domain and are classified using a neural network in order to classify CT images into two categories: with tumorous lung nodules and without lung nodules.

# Methodology

System Overview

The System is designed to perform a series of activities. The activities are as 1:- Image Acquisition, takes a single image or series of image 2:- Pre-Processing, cleans the image from noise and various other factors 3:- Image Segmentation, partitions the image into multiple meaningful segments 4:- Feature Extraction, measures data and builds required derived values also called as Features 5:- Classification, bases on the features extracted the image and segments are classified 6:- 3D Volume, volume detection based on the slices and the tumor volume. below image denotes the process flow of the system.

Figure ‑ : System Flow

1. Image Acquisition

First step is to acquire the CT scan image of lung cancer patient. The lung CT images are having low noise when compared to X-ray and MRI images; hence they are considered for developing the technique. The main advantage of using computed tomography images is that, it gives better clarity and less distortion. For research work, the CT images are acquired from NIH/NCI Lung Image Database Consortium (LIDC) dataset. DICOM (Digital Imaging and Communications in Medicine) has become a standard for medical Imaging. Figure 4.2 shows a typical CT image of lung cancer patient used for analysis. The acquired images are in raw form. In the acquired images lot of noise is observed. To improve the contrast, clarity, separate the background noise, it is required to pre-process the images. Hence, various techniques like smoothing, enhancement are applied to get image in required form. DICOM image is 16 bit image.

In this project LIDC database from TCIA collection is used.



Figure ‑ : Sample Image

1. Image Segmentation

Image segmentation is a crucial process for most image analysis consequent tasks. Especially, most of the existing techniques for image description and recognition are highly depend on the segmentation results. Segmentation splits the image into its constituent regions or objects. Segmentation of medical images in 2D has many beneficial applications for the medical professional such as: visualization and volume estimation of objects of concern, detection of oddities, tissue quantification and organization and many more. The main objective of segmentation is to simplify and change the representation of the image into something that is more significant and easier to examine. Image segmentation is usually used to trace objects and borders such as lines, curves, etc. in images. More accurately, image segmentation is the process of allocating a label to every pixel in an image such that pixels with the same label share certain pictorial features. The outcome of image segmentation is a set of segments that collectively cover the entire image, or a set of edges extracted from the image i.e. edge detection. In a given region all pixels are similar relating to some distinctive or computed property, such as texture, intensity or color. With respect to the same characteristics adjacent regions are significantly different. One of two basic properties of intensity values Segmentation algorithms are based on: discontinuity and similarity. In the first group we partition the image based on abrupt changes in intensity, such as edges in an image. The next group is based on segregating the image into regions that are alike according to a predefined criterion.

Thresholding Image

 Morphology operator (Mask)

 Mask \*original

K-Mean Clustering

Apply Contour

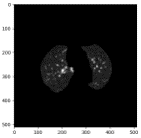
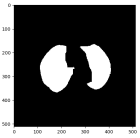
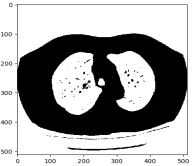


Figure ‑ : Segmentation Flow

* 1. **Thresholding Image:**

Thresholding is a process of converting a grayscale input image to a bi-level image by using an optimal threshold. For a thresholding algorithm to be really effective, it should preserve logical and semantic content. There are two types of thresholding algorithms

1. Global thresholding algorithms
2. Local or adaptive thresholding algorithms

In global thresholding, a single threshold for all the image pixels is used. When the pixel values of the components and that of background are fairly consistent in their respective values over the entire image, global thresholding could be used. Here, Global Thresholding Method is used. Threshold value is 625.

* 1. **Morphology Operation:**

Dilation and erosion are two elementary morphological operations used for Lung masking. An opening is erosion followed by dilation with structuring element. Morphology operation is applied on Otsu thresholding image.

Erosion and Dilation: Morphology is a broad set of image processing operations that process images based on shapes. Morphological operations apply a structuring element to an input image, creating an output image of the same size. In a Morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbours. By choosing the size and shape of the neighbourhood, you can construct a morphological operation that is sensitive to specific shapes in the input image. The most basic morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image. In the morphological dilation and erosion operations, the state of any given pixel in the output image is determined by applying a rule to the corresponding pixel and its neighbours in the input image. The rule used to process the pixels defines the operation as dilation or erosion.

Morphology operator is used to create mask. This mask is get multiply with original image to find out abnormality in image

* 1. **K-mean Clustering**

K-means clustering is the unsupervised machine learning algorithm. The main idea behind this algorithm is to classify the given set of data into a number of clusters. The location of the centers of each cluster plays an important role for good classification. The center of each cluster should be as far as possible for effective. The next step is to include each belonging point nearest to the center [19]. When all the points from the dataset are classified then the first step is completed. Again, recalculate the new centroid resulting from the first step. The process is repeated for new data points and nearest new centers. At every iteration, the position of center is changing to get the optimized center point which is represented by the objective function given by,

(1)

Where ‘||xi - vj||’ is the Euclidean distance between xi and cj

‘xi’ is the number of data points in the ith cluster.

‘cj’ is the number of cluster centers.

Morphological operations are applied on the segmented image to get proper boundaries of the lungs template. Erosion and dilation operation is used to erode the small unwanted binary objects and dilate the useful binary object. This binary mask is convolved with the original image to get original lungs volumes. These lungs volumes are used to extract the features [20]. K mean clustering is applied to each CT slice.

* 1. **Contour**

Contour can be explained simply as a curve joining all the continuous points (along the boundary), having same color or intensity. The contours are a useful tool for shape analysis and object detection and recognition.

* For better accuracy, use binary images. So before finding contours, apply threshold or canny edge detection.
* Find Contours function modifies the source image. So if you want source image even after finding contours, already store it to some other variables.
* In Open CV, finding contours is like finding white object from black background. So remember, object to be found should be white and background should be black.

Contour is used to highlight the tumor part from CT images.

1. Feature Extraction

In Machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Features are an important measurement for image understanding, especially the feature representation of the segmented region that used for object classification and analysis. Features are gray-level histogram, GLCM (Gray-Level Co-occurrence Matrix) texture, shape invariant moment and FFT frequency features. Feature extraction calculates features on the basis of which image can be easily classified as normal or abnormal one. The feature extraction is the process to represent a raw image to facilitate decision making such as pattern classification. [21] Features will be extracted from the tumor regions from CT images. Feature extraction involves reducing the amount of data required to describe a large set of data accurately. Features are used as inputs to classifiers that assign them to the class that they represent. The intention of feature extraction is to reduce the original data by measuring positive properties, or features, that discriminate one input sample from another sample

* 1. **Grey Level Co-occurrence Matrix (GLCM)**

The gray level co-occurrence matrix (GLCM) is a pattern matrix used to find the

texture pattern in an image by modelling texture as a 2-dimensional array gray level variation. This array is called the gray level co-occurrence matrix. The GLCM feature is

considered among the set of features to describe the pixel contrast and the energy of the

region of interest. These features clearly differentiate normal tissue from the abnormal tissue of the lung based on the extracted contrast and energy.

The computed features were: auto-correlation, entropy, sum average, sum variance, sum entropy, difference variance, and difference entropy, information measure of correlation, contrast, dissimilarity, energy, cluster prominence, cluster shade, variance, inverse difference moment, homogeneity and maximum probability.

Contrast, Dissimilarity, Homogeneity, Angular Second Moment (ASM), Mean, Energy, Standard deviation this feature is used to train a classifier.

1. Contrast : It is the measures of local variations in the GLCM

(2)

1. Dissimilarity: It measures the variation in the gray levels of the grayscale image. (3)
2. Homogeneity: It measures of the closeness of the element distribution in GLCM. *Homogeneity =* (4)
3. Angular Second Moment (ASM): The rotational acceleration measurement is carried out using ASM. (5)
4. Mean: It is average of all pixel values. (6)
5. Energy: The energy is the measure of uniformity between the pixels. (7)
6. Standard Deviation: Standard deviation measures the brightness within the image region. (8)

Where n is the number of pixels, P (i, j) is the pixel information,

μx, μy = mean of the Px and Py

This feature extraction is give as input to a classifier to detect tumorous and Non tumorous module.

1. Classification

The features and their combination in various scales are used for medical image classification. Medical image classification procedure has two parts: training (learning) and classification (testing). There are many classic classifiers, such as C4.5, SVM (Support Vector Machine), Naive Bayes, KNN (K-Nearest Neighbour) and ANN (Artificial Neural Network). However, the classic classifiers can’t support multi-feature vectors well. An ensemble learning based classification framework is proposed and applied to medical image classification task with feature extracted. The classic classifiers are used as base classifier to construct a stronger classifier with a decision maker. Here, Multilayer Perceptron based classifier is used.

**Supervised Classification**

Supervised learning based classification is depends on data which is created from the knowledge of domain. In supervised learning labelled data points are used. To determine accurate categorization of an image in supervised classification pre-labelled samples are required. In this method, training is required or expertise knowledge is required so this technique becomes time consuming. This is the reason; in some areas this technique is not suitable. In order to determine a decision rule for classification, it is necessary to know the spectral characteristics or feature with respect to the population of each class.

**Advantages**

* Errors can be detected by operators and they often remedy them.
* Expertise knowledge required, so this method will give the accurate result.

**Disadvantages**

* Not suitable to deal with big data, because for each area it requires area experts.
* Very Time consuming. It takes so much time to identify pre-labelled samples.

**Unsupervised Classification**

Some situation requires little information about the area to be classified, only image properties are used as a

* Randomly sampled data’s several groups will be divided mechanically into the same classes by using clustering techniques.
* These clustered classes later used for determining population statistics. This kind of Classification is called the unsupervised classification.

**Advantages**

* Scientist spends less time to classify the domain. As a result only required images are classified.
* This approach is very a suitable to classify large data.

**Disadvantages**

* Any kind of training is not given in this method, so it requires great knowledge about the area or about the method which is suitable for the desired area.
* With large data sets computation time is large and it creates useless classifier.

**Neural Network**

ANN A[rtificial neural networks](https://en.wikipedia.org/wiki/Artificial_neural_network) are [computational models](https://en.wikipedia.org/wiki/Computational_model) inspired by [biological neural networks](https://en.wikipedia.org/wiki/Biological_neural_network), and are used to [approximate](https://en.wikipedia.org/wiki/Universal_approximation_theorem) [functions](https://en.wikipedia.org/wiki/Function_(mathematics)) that are generally unknown. Particularly, they are inspired by the behaviour of [neurons](https://en.wikipedia.org/wiki/Neuron) and the electrical signals they convey between input, processing, and output from the brain. The way neurons semantically communicate is an area of on-going research. Most artificial neural networks bear only some resemblance to their more complex biological counterparts, but are very effective at their intended tasks (e.g. classification or segmentation).

Here Multi-Layer Perceptron classifier is used.

* 1. **Multi-Layer Perceptron Neural network**

Multi-Layer Perceptron algorithm of the artificial neural network is used as a classification method. This algorithm consists of three layers, i.e. Input layer, a hidden layer and an output layer. In each layer, there are several neurons. MLP produces different classes using direct learning process and calculates the optimal weights by backpropagation training process [22]. The layered block diagram of the Multi-Layer Perceptron is as shown in Figure 4-4.

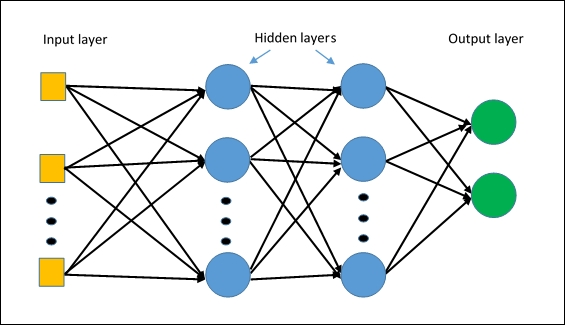


Figure ‑ : Layered diagram of Multi-Layer Perceptron NN

The parameter used to train the MLP model are as shown in Table4-1

|  |  |
| --- | --- |
| Number of hidden layers | 15 |
| Alpha | 0.0001 |
| Learning rate | 0.001 |
| Batch size | 200 |
| Batch size | 1 |
| Solver | lbfgs (Backpropagation) |

Table ‑ : Parameter used to train MLP

The solver is used for weight optimization. Solver ‘lbfgs’ is an optimizer of quasi –newton method i.e. a back propagation neural network. Back propagation neural network improve the accuracy of the neural network.

* 1. **K-Nearest Neighbor (KNN)**

KNN is one of the simplest of classification algorithms available for supervised learning. The idea is to search for the closest match of the test data in feature space. We will look into it with the above image.

****

Figure ‑ : KNN Basic Diagram

In the image, there are two families, Blue Squares and Red Triangles. Consider each family as a Class. Their houses are shown in their town map which is called as feature space. Now a new member comes into the feature space and creates a new place, which is shown as a green circle. It should be added to one of these Blue/Red families. This process is called as Classification. Since here is dealing with KNN, apply this algorithm. One method is to check who his nearest neighbor is from the image, it is clear it is the Red Triangle family. So it is also added into Red Triangle. This method is called simply Nearest Neighbour, because classification depends only on the nearest neighbor. But there is a problem with that. Red Triangle may be the nearest. But what if there are lots of the Blue Squares near to it? Then Blue Squares have more strength in that locality than Red Triangle. So just checking nearest one is not sufficient. Instead, it checks some k nearest families. Then whoever is the majority of them, the new one belongs to that family. In this image, let’s take K=3, i.e. 3 nearest families. It has two Red and one Blue (there are two Blues equidistant, but since K=3, it take only one of them), so again it should be added to Red family. But what if consider the value of k is 7 i.e. K=7.Then it has 5 Blue families and 2 Red families. Now it should be added to Blue family. So it all changes with value of k .If K = 4, it has 2 Red and 2 Blue neighbors. It is a tie. So better take k as an odd number. So this method is called k-Nearest Neighbor since classification depends on k-nearest neighbors. In this way the k Nearest Neighbor algorithm worked.

* 1. **Assumptions in KNN**

KNN assumes that the data is in a feature space. More exactly, the data points are in a metric space. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. Usually Euclidean distance is used as the distance metric. Here in present work, along with the Euclidean distance, city block, cosine and correlation distances are also used to classified samples.

* 1. **Parameter Selection**

The best choice of K depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when K = 1) is called the nearest neighbor (NN) algorithm. Here K=3 is used.

The K- nearest neighbor algorithm is as follow:

1. Determine parameter K= number of nearest neighbors.
2. Calculate the distance between the query-instance and all the training samples

The Euclidean distance between X = (x1, x2, x3,..xn) and Y = (y1,y2, y3,yn) is

defined as:

1. Sort the distance and determine nearest neighbors based on the k-th minimum

distance

1. Gather the category of the nearest neighbors.
2. Use simple majority of the category of nearest neighbors as the prediction value of

the query instance.

1. 3-D Volume

3D imaging diagnostic equipment such as CT scanners, MRI, 3D, 4D ultrasound is used in many countries. This equipment has been alternatively assisted by information technology that requires high-end computers with dedicated software and programs. Nowadays, such high-tech equipment are equipped with domestic technology. On the other hand, a medical diagnosis system is not structured clearly. The medical units in a national health care system are not correlative with any standard process to operate image diagnostic equipment or to control the patient data. 3D medical image reconstruction is an element of a field called visualization that represents data as pictures to assist the users for a better understanding. Recent medical 3D image reconstruction techniques construct the 3D images from the sets of 2D slices. It can be recorded by different equipment like CT, Ultrasound, MRI, etc. 3D reconstruction from multiple images is the formation or the generation of 3D models from a set of images. It is the inversion process of obtaining 2D images from 3D sights. The core of the image is a projection from a 3D view onto a 2D plane, during the progression the depth is lost. The 3D point consistent with a particular image point is considered to be a Line Of Sight (LOS). It is impossible to determine the point on this line relates to the image point from a distinct image. If two images are accessible, then the position of the 3D point can be viewed as the intersection of the two prediction rays. The aforementioned process is termed as triangulation. The basic idea for this process is the association between several views that express the information of equivalent sets of points. The sets of points should contain some arrangement and this arrangement belongs to the calibration of the camera. In recent times, there is a significant demand for 3D content for communication, computer graphics and virtual reality, activating stress in the requirements.

* 1. **3D Rendering Techniques**

3D rendering is the 3D computer graphics process that spontaneously translates 3D wire frame models into 2D images with 3D photorealistic effects or non-photorealistic rendering. Rendering is the visualization of an image using a model with the help of a software program. Two types of rendering techniques are widely used.

They are,

1. Surface Rendering (SR)

2. Volume Rendering (VR)

Volume Rendering(VR) is a visualization of data in 3D by computing the 2D projection of a volume in any desired direction without any prior geometric information. A classic 3D dataset is a set of 2D slice images attained by a CT, MRI or Micro-CT scanner. These are attained in a regular pattern and have a regular number of pixels in a consistent pattern. Usually, this method is used to visualize the complete volume transparency of data. Along every ray, color and opacity need to be intended at each voxel. Consequently, information calculated along each ray will be combined to a pixel on the image plane (Hussain & Prasad 2012). This method helps to see a complete structure of the object. The drawback of this method is the huge amount of calculations that involves high-end design computers. This technique is suitable for low contrast data. Once a 3D volume is reconstructed, it can be spotted from different directions. In general, the color can be assigned to each voxel for rendering. This color detail is extracted from the types of camera views, where the particular vessel is visible. To judge occlusions of multiple voxels, a z-buffer can be included for rendering. Every voxel can be projected onto the image plane and the pixel color is changed if it is nearer to the virtual camera then the current value is stored in the z-buffer. Typically, various approaches can be classified by the direction of the volume traversal like arbitrary traversal, front-to-back and back-to-front.

Volume= area of each tumour slice \* slice thickness

# Software Requirement

DICOM Viewer

DICOM is acronym for Digital Imaging and Communications in Medicine. DICOM is used worldwide to store, exchange, and transmit medical images. DICOM has been central to the development of modern radiological imaging: DICOM incorporates standards for imaging modalities such as radiography, ultrasonography, computed tomography (CT), magnetic resonance imaging (MRI), and radiation therapy. DICOM includes protocols for image exchange (e.g., via portable media such as DVDs), image compression, 3-D visualization, image presentation, and results reporting.

DICOM files can be exchanged between two entities that are capable of receiving image and patient data in DICOM format (.dcm). The different devices come with DICOM Conformance Statements which clearly state which DICOM classes they support, and the standard includes a file format definition and a network communication that uses TCP/IP to communicate between systems.

Free, open source software package for image analysis and scientific visualization, with the integrated support of components of DICOM standard.

Python

Python is a widely used high-level programming language for general-purpose programming. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation. Python is having thousands of libraries. Libraries such as NumPy, SciPy and Matplotlib allow the effective use of Python in scientific computing, with specialized libraries such as Biopython and Astropy providing domain-specific functionality.

Python has been used in artificial intelligence projects. As a scripting language with modular architecture, simple syntax and rich text processing tools, Python is often used for natural language processing.

**Advantages**

* **Easy Syntax**

Python's syntax is easy to learn, so both non-programmers and programmers can start programming right away.

* **Readability**

Python's syntax is very clear, so it is easy to understand program code. (Python is often referred to as "executable pseudo-code" because its syntax mostly follows the conventions used by programmers to outline their ideas without the formal verbosity of code in most programming languages; in other words syntax of Python is almost identical to the simplified "pseudo-code" used by many programmers to prototype and describe their solution to other programmers. Thus Python can be used to prototype and test code which is later to be implemented in other programming languages).

* **High-Level Language**

Python looks more like a readable, human language than like a low-level language. This gives you the ability to program at a faster rate than a low-level language will allow you.

* **Object oriented programming**

Object-oriented programming allows you to create data structures that can be re-used, which reduces the amount of repetitive work that you'll need to do. Programming languages usually define objects with namespaces, like class or def, and objects can edit themselves by using keyword, like this or self. Most modern programming languages are object-oriented (such as Java, C++, and C#) or have support for OOP features (such as Perl version 5 and later). Additionally object-oriented techniques can be used in the design of almost any non-trivial software and implemented in almost any programming or scripting language. (For example a number of Linux kernel features are "objects" which implement their own [encapsulation](https://en.wikipedia.org/wiki/Encapsulation_(object-oriented_programming)) of behaviour and data structure via pointers, specifically pointers to functions, in the C programming language)Python's support for object-oriented programming is one of its greatest benefits to new programmers because they will be encountering the same concepts and terminology in their work environment. If you ever decide to switch languages, or use any other for that fact, you'll have a significant chance that you'll be working with object-oriented programming.

* **It's Free**

Python is both free and open-source. The Python Software Foundation distributes pre-made binaries that are freely available for use on all major operating systems called CPython. You can get CPython's source-code, too. Plus, you can modify the source code and distribute as allowed by CPython's license.

* **Cross-platform**

Python runs on all major operating systems like Microsoft Windows, Linux, and Mac OS X.

* **Widely Supported**

Python has an active support community with many web sites, mailing lists, and USENET "netnews" groups that attract a large number of knowledgeable and helpful contributes.

* **It's Safe**

Python doesn't have pointers like other C-based languages, making it much more reliable. Along with that, errors never pass silently unless they're explicitly silenced. This allows you to see and read why the program crashed and where to correct your error.

* **Batteries Included**

Python is famous for being the "batteries are included" language. There are over 300 standard library modules which contain modules and classes for a wide variety of programming tasks. For example the standard library contains modules for safely creating temporary files (named or anonymous), mapping files into memory (including use of shared and anonymous memory mappings), spawning and controlling sub-processes, compressing and decompressing files (compatible with gzip or PK-zip) and archives files (such as Unix/Linux "tar"), accessing indexed "DBM" (database) files, interfacing to various graphical user interfaces (such as the TK toolkit and the popular Windows multi-platform windowing system), parsing and maintaining CSV (comma-separated values) and ".cfg" or ".ini" configuration files (similar in syntax to the venerable WIN.INI files from MS-DOS and MS-Windows), for sending e-mail, fetching and parsing web pages, etc. It's possible, for example, to create a custom web server in Python using less than a dozen lines of code, and one of the standard libraries, of course.

* **Extensible**

In addition to the standard libraries there are extensive collections of freely available add-on modules, libraries, frameworks, and tool-kits. These generally conform to similar standards and conventions; for example almost all of the database adapters (to talk to almost any client-server RDBMS engine such as MySQL, Postgress, Oracle, etc.) conform to the Python DBAPI and thus can mostly be accessed using the same code. So it's usually easy to modify a Python program to support any database engine.

**Disadvantages**

* **Speed**

Python is executed by an interpreter instead of compilation, which causes it to be slower than if it was compiled and then executed. However, for most applications, it is by far fast enough. One Python idiom is "Speed isn't a problem until it's a problem."

# Implementation and Results

Process flow

The process flow for Lung tumor classification is as shown in below figure 6-1.

Training Phase

Testing Phase

CT Scan

CT Scan

Segmentation

Feature Extraction

Feature Extraction

Segmentation

Classifier

Learning

Tumorous

Non-Tumorous

Figure ‑ : Process Flow

The flow of proposed approach is an explained and shown in figure 6.1. Proposed system incorporate two phases: Training phase and testing phase. The training phase involves the construction of labelled data set and feature vector. This dataset is saved or store in structured way. The process of storing dataset is a part of classifier training. The labelled dataset is tagged to either tumorous or non-tumorous classification.

While in testing phase, when dataset is extracted using feature extraction, the dataset is sent as a input to the classifier, which ten uses the labelled dataset already stored to classify the dataset as tumorous or non-tumorous. The feature extracted from the segmented image by using GLCM techniques. The extracted feature was trained Classifier separately and test images according to that trained model.

Learning Module is common to both Training Phase and Testing Phase. When Training is being done Learning Module stores the Classified Image as either tumorous or non-tumorous. Also, different features that are extracted for both tumorous and non-tumorous are stored for individual images. While in testing phase, the segmented images along with the features are sent to the learning module, where the learning module compares the features of already stored segments for either tumorous or non-tumorous. If a Valid Match (with approximation) is found, the image or segment of the image is classified as tumorous. And, when displaying the image with tumorous nodule, a contour is drawn across the segment for visual identification and plotting.

Implementation Steps

Below are the Implementation Steps which will be used for the Project flow.

1. Collection of tumor CT database
2. Apply Segmentation techniques
3. Extract GLCM features
4. Classify the tumor CT images
5. Generate 3D volume of tumor

Experimental Results

The system of lung tumor classification using image processing and soft computing techniques is implemented by using Python programming in pycharm editor 3.6 version.

Total 25 CT people images in DICOM format are used for experimentation. Firstly the 6000 images are divided into training and testing dataset. Here, 5500 images are used for training while 100 images are used for testing purpose. The image processing techniques are implemented on these training and testing images. Then their features are extracted. These extracted features are stored in an excel sheets which are further classified to MLP and KNN classifier.

|  |  |  |
| --- | --- | --- |
| Database | Number of images | |
| Training | Testing |
| LIDC | 4000 | 80 |
| NSLCS | 2000 | 20 |

Table ‑ : Database Distribution

The original Image of Lung CT images are shown in figure, One CT consist of many number of slices. In below figure is a sample 9 slice of one patient.

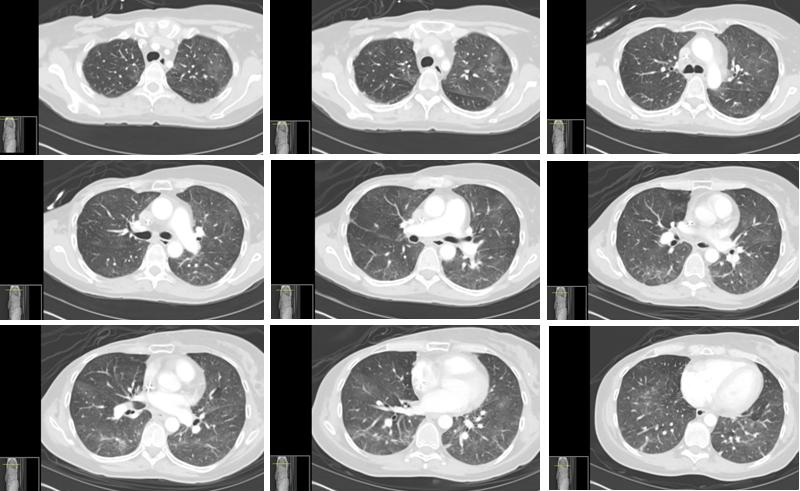


Figure ‑ : Sample Lung Original Image

|  |  |
| --- | --- |
| I:\op\original.png  Figure ‑ : Original Image with Tumor | I:\op\morph-open.png  Figure ‑ : Morphology Opening Image |
| Figure ‑ : Threshold Image | Figure ‑ : Lung Mask |
| Figure ‑ : Mask \* Original Image | I:\op\segmentation.png  Figure ‑ : K-mean clustering and Contour Detection |

In segmentation Morphological operations are applied on the segmented image to get proper boundaries of the lungs template (Mask). Erosion and dilation operation is used to erode the small unwanted binary objects and dilate the useful binary object. This binary mask is convolved with the original image to get original lungs mask. After multiplication K-mean Clustering is used to identify tumor part from image. Then contour detection is applied on K-mean output to highlights the boundaries of Lung tumor.

For machine learning algorithm, the features are taken as an input hence feature extraction is the important step. In the proposed algorithm, the GLCM is used to extract the features. The features are extracted from the segmented grayscale lung volume. Contrast, Dissimilarity, Homogeneity, Angular Second Moment (ASM), Mean, Energy, Standard deviation this feature is used to train a classifier. Table 6-2 gives the idea about feature values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sr. No | Contrast | Dissimilarity | Homogeneity | ASM | Energy | Mean | Standard Deviation |
| 1 | 146 | 96 | 261589 | 3.97E+09 | 63001.98 | 272.8054 | 1216.792 |
| 2 | 144 | 90 | 261592.4 | 3.97E+09 | 63001.98 | 273.4708 | 1216.988 |
| 3 | 148 | 98 | 261588 | 3.96E+09 | 62931.38 | 273.9084 | 1217.169 |
| 4 | 144 | 96 | 261588.8 | 3.97E+09 | 62985.37 | 274.227 | 1217.208 |
| 5 | 138 | 90 | 261591.8 | 3.97E+09 | 63022.73 | 274.5451 | 1217.172 |
| 6 | 130 | 76 | 261599.4 | 3.99E+09 | 63130.53 | 274.9871 | 1217.274 |
| 7 | 148 | 88 | 261594 | 3.99E+09 | 63147.1 | 275.3596 | 1217.357 |
| 8 | 102 | 54 | 261609.8 | 4.01E+09 | 63300.17 | 274.4824 | 1216.461 |
| 9 | 70 | 38 | 261616.2 | 4.01E+09 | 63329.09 | 272.2862 | 1214.532 |
| 10 | 74 | 42 | 261614.2 | 4.01E+09 | 63320.83 | 270.2262 | 1212.762 |
| 11 | 80 | 44 | 261613.6 | 4.01E+09 | 63333.22 | 269.1934 | 1211.826 |
| 12 | 72 | 42 | 261614 | 4.01E+09 | 63312.57 | 268.8703 | 1211.536 |
| 13 | 106 | 64 | 261604.2 | 3.99E+09 | 63205.06 | 269.4121 | 1212.053 |
| 14 | 148 | 88 | 261594 | 3.98E+09 | 63072.5 | 270.302 | 1212.861 |
| 15 | 192 | 110 | 261585.2 | 3.97E+09 | 63022.73 | 270.3346 | 1213.025 |
| 16 | 168 | 94 | 261592.4 | 3.98E+09 | 63097.38 | 269.8783 | 1212.768 |
| 17 | 300 | 154 | 261569.3 | 3.96E+09 | 62914.75 | 269.5974 | 1212.704 |
| 18 | 314 | 176 | 261557.7 | 3.93E+09 | 62706.62 | 269.1178 | 1212.555 |
| ….. | …... | ….….. | ….. | …. | …… | …….. | ……. |
| 6207 | 372 | 216 | 261539.6 | 3.92E+09 | 62635.71 | 267.9825 | 1212.181 |
| 6208 | 642 | 368 | 261475.4 | 3.82E+09 | 61833.78 | 267.0551 | 1211.769 |

Table ‑ : Values of Feature Extraction

Classifier is used to classify the tumorous and non tumorous image. Table 6-3 gives the accuracy of both classifer in training and testing phase.

|  |  |
| --- | --- |
| Classifier | Accuracy |
| MLP Training | 98.20 |
| MLP Testing | 98.30 |
| KNN Cross-Validation | 0.9792 |
| KNN Testing | 98.3091 |

Table ‑ : Classifier Accuracy

The 3D view of tumor are shown as figure 6.8

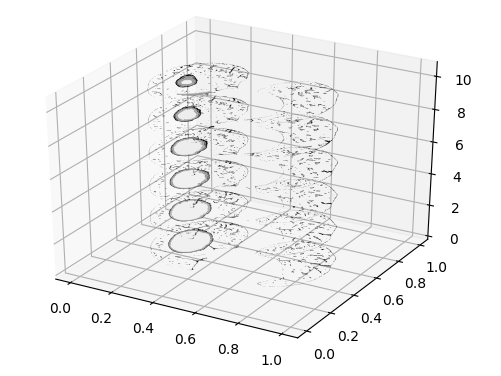


Figure ‑ : 3-D view of lung tumor

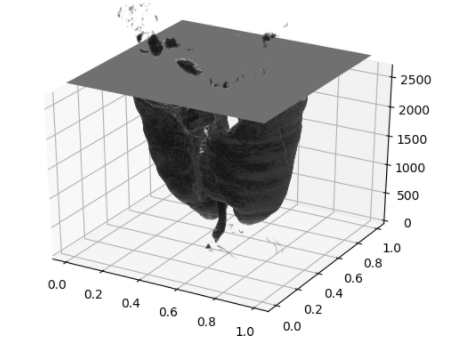


Figure ‑ : 3-D view of lung in Grayscale

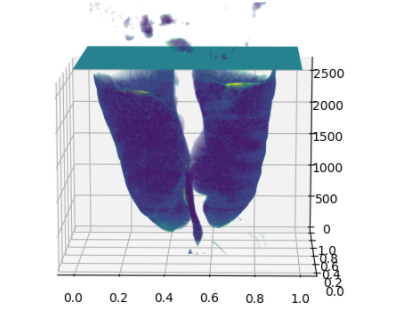


Figure ‑ : 3-D view of lung with tumor in blue scale

Graphics User Interface Result :

|  |  |
| --- | --- |
| Figure ‑ : GUI | Figure ‑ Play Images in DICOM Directory |
| Figure ‑ : GUI Segmentation Display | |

# Conclusion and Future Scope

Conclusion

Lung cancer is the second most cause of death from all type of cancer. According to survey done by American cancer society its occurrence is high in both male and females. In this system detection and classification of lung cancer nodule is done using CT images in DICOM format. Total 6000 images are used in the system implementation. The CT image segmentation is performed using thresholding, morphology, K-mean clustering, and contour detection algorithm. Features are contrast, energy, dissimilarity, mean, ASM, homogeneity are calculated using Python commands. The database of feature sets is generated in excel sheet. Total 6000 images are used for training set and 100 for testing. The feature database is provided to KNN and MLP classifier. The accuracy of MLP training is 98.20 %, MLP testing accuracy is 98.30. KNN Cross validation accuracy is 0.9792 % and KNN testing accuracy is 98.30 %. Out of these two classifier it is observed that the performance of KNN is better.3D view or volume gives the idea to the radiologist to identify the tumor region. This system will help radiologist for diagnosis of cancer.

Future Scope

* In future work we will try to extract more features so that the performance parameter of system will enhance their performance further.
* This system may be used for different type of tumor classification like Mammogram, Lung cancer etc.

# References

1. M. Iranifam, "Analytical applications of chemiluminescence methods for cancer detection and therapy," TrAC Trends in Analytical Chemistry, vol. 59, pp. 156-183, 2014.
2. R. Siegel, J. Ma, Z. Zou, and A. Jemal," Cancer statistics, 2014," CA: a cancer journal for clinicians, vol. 64, no.1, pp. 9-29, 2014.
3. D. C. Levin, L. Parker, E. J. Halpern, and V. M. Rao, "Are combined CT scans of the thorax being overused?" Journal of the American College of Radiology, vol. 11, no. 8, pp. 788-790, 2014.
4. J. F. Rasmussen, V. Siersma, J. H. Pedersen, B. Heleno, Z. Saghir, and J. Brodersen, "Healthcare costs in the Danish randomised controlled lung cancer CT-screening trial: a registry study," Lung Cancer, vol. 83, no. 3, pp. 347-355,2014.
5. H. Jiang, H. Ma, W. Qian, M. Gao, and Y. Li, "An automatic detection system of lung nodule based on multi-group patch-based deep learning network," IEEE Journal of Biomedical and Health Informatics, 2017.
6. Gupta, O. Martens, Y. Le Moullec, and T. Saar, "A tool for lung nodules analysis based on segmentation and morphological operation," in Intelligent Signal Processing (WISP), 2015 IEEE 9th International Symposium on. IEEE, 2015, pp. 1-5.
7. M. B. A. Miah and M. A. Yousuf, "Detection of lung cancer from CT image using image processing and neural network," in Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on. IEEE, 2015, pp. 1-6.
8. R. Sammouda, "Segmentation and analysis of CT chest images for early lung cancer detection," in Computer & Information Technology (GSCIT), 2016 Global Summit on. IEEE, 2016, pp. 120-126.
9. M. Alam, G. Sankaranarayanan, and V. Devarajan, "Lung nodule detection and segmentation using a patch-based multi-atlas method," in Computational Science and Computational Intelligence (CSCI), 2016 International Conference on IEEE, 2016, pp. 23-28
10. V. Kalpana and G. Rajini, "Segmentation of lung lesion nodules using DICOM with structuring elements and noise-a comparative study," in Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on. IEEE, 2016, pp. 252-257.
11. G. P. Pratap and R. Chauhan, "Detection of lung cancer cells using image processing techniques," in Power Electronics, Intelligent Control and Energy Systems (ICPEICES), IEEE International Conference on. IEEE, 2016, pp. 1-6.
12. Ada, Rajneet Kaur “Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier”, (IJAIEM)Volume 2, Issue 6, June 2013
13. Taher, Fatma, and Rachid Sammouda. "Lung cancer detection by using artificial neural network and fuzzy clustering methods." GCC Conference and Exhibition (GCC), 2011 IEEE. IEEE, 2011.
14. Cabrera, Jennifer, Abigaile Dionisio, and Geoffrey Solano. "Lung cancer classification tool using microarray data and support vector machines." In Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on, pp. 1-6. IEEE, 2015.
15. Kanitkar, Sayali Satish, N. D. Thombare, and S. S. Lokhande. "Detection of lung cancer using marker-controlled watershed transform." In Pervasive Computing (ICPC), 2015 International Conference on, pp. 1-6. IEEE, 2015.
16. E. Rendon-Gonzalez and V. Ponomaryov, "Automatic Lung nodule segmentation and classification in CT images based on SVM," 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW), Kharkiv, 2016, pp. 1-4.
17. Kaur, R., & Verma, P. (2015). “Improved MLP-NN based approach for Lung Diseases Classification”. International Journal of Computer Applications, 131(6), 22-26.
18. Harikumar Rajaguru, Harikumar Rajaguru,“Performance Comparison of Oral Cancer Classification with Gaussian Mixture Measures and Multi Layer Perceptron” [The 16th International Conference on Biomedical Engineering](https://link.springer.com/book/10.1007/978-981-10-4220-1) pp 123-129.
19. P. Sarker, M. M. H. Shuvo, Z. Hossain and S. Hasan, "Segmentation and classification of lung tumor from 3D CT image using K-means clustering algorithm," 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, 2017, pp. 731-736.
20. Anindya Gupta, Olev Martens, Yannick Le Moullec, Tönis Saar, "Methods for increased sensitivity and scope in automatic segmentation and detection of lung nodules in CT images", Signal Processing and Information Technology (ISSPIT) 2015 IEEE International Symposium on, pp. 375-380, 2015
21. Li, Z., R. Hayward, Y. Liu, and R. Walker. 2011. Spectral-Texture Feature Ex-traction Using Statistical Moments with Application to Object-Based Vegetation Species Classification. International Journal of Image and Data Fusion 2 (4): 347361.doi:10.1080 /19479832. 2010. 546372
22. M. Azehoun-Pazou, Geraud & M. Assogba, Kokou & Adegbidi, “A novel approach of black skin lesion images segmentation based on MLP Neural Network. ” 1-4. 10.1109/BIOSMART.2016.7835598

Publication

* Sneha Potghan, R.Rajamenakshi, Dr.Archana Bhise “Multilayer Perceptron based Lung tumor classification” 2nd IEEE conference on Electronics, Communication and Aerospace Technology (ICECA2018) at Hotel Arcadia, Coimbatore, Tamilnadu, India presented on 30th March 2018

# Appendix

**Python Main file**

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import KNeighborsClassifier

import pandas as pd

from functions import \*

from skimage import morphology, measure, filters

from skimage.measure import label, regionprops

from sklearn.cluster import KMeans

import matplotlib.pyplot as pyplot

from skimage.feature import greycomatrix, greycoprops

from sklearn.neural\_network import MLPClassifier

from sklearn import cross\_validation

import pickle

path = "featuresdicom.xlsx"

INPUT\_SCAN\_FOLDER = "G:\\final\\dicom final\\database\\malignant\\LIDC-IDRI-0072\\"

fileDICOMFeatureList = pd.read\_excel(path, header=None)

matrixFeatures = np.array((fileDICOMFeatureList.as\_matrix())[1:, :])

yMatrixFeatures = matrixFeatures[:, 7]

xMatrifFeatures = matrixFeatures[:, 0:7]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(xMatrifFeatures, yMatrixFeatures, test\_size=0.2, random\_state=10)

y\_train = y\_train.astype('int')

y\_test = y\_test.astype('int')

# clf = MLPClassifier(hidden\_layer\_sizes=1000,solver='lbfgs')

# clf.fit(X\_train, y\_train)

modelFileMLP = 'mlpmodel.sav'

# pickle.dump(clf,open(filename,'wb'))

modelMLP = pickle.load(open(modelFileMLP, 'rb'))

# print(clf)

MLPscore = modelMLP.score(X\_train, y\_train)

MLPtest = modelMLP.predict(X\_test)

print('MLP training=', MLPscore \* 100)

print("MLP testing accuracy=", np.mean(MLPtest == y\_test) \* 100)

KNNmodel = KNeighborsClassifier()

kfold = cross\_validation.KFold(n=len(X\_train), n\_folds=10, random\_state=10)

cv\_results = cross\_validation.cross\_val\_score(KNNmodel, X\_train, y\_train, cv=kfold, scoring='accuracy')

message = "%s: %f " % ("KNN cross validation accuracy", cv\_results.mean())

print(message)

K\_value = 3

neigh = KNeighborsClassifier(n\_neighbors=K\_value, weights='uniform', algorithm='auto')

neigh.fit(X\_train, y\_train)

KNNpredictValue = neigh.predict(X\_test)

print("KNN testing accuracy=", np.mean(KNNpredictValue == y\_test) \* 100)

listProperties = ['contrast', 'dissimilarity', 'homogeneity', 'ASM', 'energy']

listFeatures = ['contrast', 'dissimilarity', 'homogeneity', 'ASM', 'energy', 'mean', 'stddev', 'label']

properties = np.zeros(6)

# glcmMatrix = []

final = []

arrayOriginalImages = dicomRead(INPUT\_SCAN\_FOLDER)

#test3D(ArrayDicom = arrayOriginalImages, i= len(arrayOriginalImages))

tumorArea = []

arrayTumorContour = []

arrayTempImage=[]

for z in range(0,len(arrayOriginalImages)):

tempImageSlice = arrayOriginalImages[z][:][:]

# img=img.pixel\_array

#imgg = tempImageSlice

tempImageMask = segment(tempImageSlice)

tempImageMask = np.where(tempImageMask == 255, 1, 0)

# pyplot.imshow(tempImageMask, cmap='gray')

# pyplot.show()

tempImageConvMask = tempImageMask \* tempImageSlice

arrayTempImage.append(tempImageConvMask)

tempImageConvMask = (tempImageConvMask / 256).astype('uint8')

ImageConvMask = tempImageConvMask

tempImageSliceMean = arrayOriginalImages[z][:][:].mean()

tempImageSliceStdDev = arrayOriginalImages[z][:][:].std()

glcmMatrix = (greycomatrix(tempImageConvMask, [1], [0], levels=2 \*\* 8))

for j in range(0, len(listProperties)):

properties[j] = (greycoprops(glcmMatrix, prop=listProperties[j]))

arrayFeatureValues = np.array([[properties[0], properties[1], properties[2], properties[3], properties[4], tempImageSliceMean, tempImageSliceStdDev]])

# pyplot.imshow(imgg,cmap='gray')

# pyplot.show()

# df = pd.DataFrame(final, columns=listFeatures)

y\_pred = neigh.predict(arrayFeatureValues)

tempSegmentedImage = tempImageConvMask

print(y\_pred)

if (y\_pred == 2 or y\_pred == 1):

segmented1 = tempSegmentedImage

tempSegmentedImageMean = np.mean(tempSegmentedImage)

tempSegmentedImageStdDev = np.std(tempSegmentedImage)

segmentedImage = tempSegmentedImage - tempSegmentedImageMean

segmentedImage = tempSegmentedImage / (tempSegmentedImageStdDev + 0.00001)

# pyplot.imshow(imgg,cmap='gray')

# pyplot.show()

# hist = pyplot.hist(segmented.flatten(), bins=200)

ROI = segmentedImage[100:400, 100:400]

ROImean = np.mean(ROI)

ROImaxv = np.max(tempSegmentedImage)

ROIminv = np.min(tempSegmentedImage)

tempSegmentedImage[tempSegmentedImage==ROImaxv]=tempSegmentedImageMean

tempSegmentedImage[tempSegmentedImage == ROIminv] = tempSegmentedImageMean

ROIkmeans = KMeans(n\_clusters=3).fit(np.reshape(ROI, [np.prod(ROI.shape), 1]))

ROIkmeanscenters = sorted(ROIkmeans.cluster\_centers\_.flatten())

ROIkmeansthreshold = np.mean(ROIkmeanscenters)

threshROIImg = np.where(segmentedImage >= ROIkmeansthreshold, 1.0, 0.0)

threshROIImg = morphology.erosion(threshROIImg, np.ones([9, 9]))

threshROIImg = morphology.dilation(threshROIImg, np.ones([9, 9]))

#pyplot.imshow(threshROIImg, cmap='gray')

#pyplot.show()

tumorContours = measure.find\_contours(threshROIImg, 0.8)

# Display the image and plot all contours found

tempTumorArea = []

if (tumorContours):

contourLabels = label(threshROIImg)

contourRegions = regionprops(contourLabels, threshROIImg)

arrayTumorContour.append(segmentedImage)

tempTumorArea = (tempTumorArea.append(contourRegions[i].area) for i in range(len(contourRegions)))

tempTumorArea = (contourRegions[0].area)

tumorArea.append(tempTumorArea)

fig, ax = pyplot.subplots()

ax.imshow(tempImageSlice, interpolation='nearest', cmap=pyplot.cm.gray)

for n, singleContour in enumerate(tumorContours):

ax.plot(singleContour[:, 1], singleContour[:, 0], linewidth=2)

ax.axis('image')

ax.set\_xticks([])

ax.set\_yticks([])

# threshROIImg = threshROIImg \* imgg

pyplot.imshow(tempImageSlice, cmap='gray')

#pyplot.show()

if (y\_pred == 1):

print(str(z) + ' Image is tumorous')

print(tempTumorArea.max())

# if(tempTumorArea<Put area here):

# elif(areaa<Put area here):

# elif(areaa < Put area here):

elif (y\_pred == 2):

print(str(z) + ' Image is tumorous')

print(tempTumorArea.max())

# if(areaa<Put area here):

# elif(areaa<Put area here):

# elif(areaa < Put area here):

else:

print(str(z) + ' Image is non tumorous')

else:

print(str(z)+' Image is non tumorous')

if (len(tumorArea)):

volume = 0;

for i in range(0, len(tumorArea) - 1):

if (i == 0):

volume = volume + (((tumorArea[i] + 0) \* 1.25) / 2)

else:

volume = volume + (((tumorArea[i] + tumorArea[i - 1]) \* 1.25) / 2)

print(volume)

print(arrayTumorContour)

from pylab import \*

from mpl\_toolkits.mplot3d import Axes3D

xx, yy = np.meshgrid(np.linspace(0, 1, 512), np.linspace(0, 1, 512))

X = xx

Y = yy

Z = len(arrayTempImage)

print('Z : '+str(Z))

ax2 = gca(projection='3d')

off = 4000 / (Z - 1)

print('off : ' + str(off))

for i in range(0,Z):

tempImage = arrayTempImage[i]

print("i : " + str(i))

print("max : " + str((tempImage).max()))

print("min : " + str((tempImage).min()))

print("tempImage.shape[0] " + str(tempImage.shape[0]))

print("tempImage.shape[1] " + str(tempImage.shape[1]))

tempImage = np.ma.masked\_where(tempImage < np.mean(tempImage), tempImage)

x, y = ogrid[0:tempImage.shape[0], 0:tempImage.shape[1]]

Z = 10 \* np.ones(X.shape)

ax = gca(projection='3d')

ax2.contourf(X, Y, tempImage, zdir='z', offset=i \* off, antialiased=True)

show()