

An-Najah National University

Faculty of Graduate Studies

**Design and Implementation of Digital
System for Lung Cancer Early Detection
Using Neural Network and Image Processing**

Prepared By

Deema Sohrab Sawalha

Supervised

Dr. Adnan Salman

**This Thesis is Submitted in Partial Fulfillment of The Requirements
for the Degree of Master of Advanced Computing, Faculty Graduate
Studies, An-Najah National University, Nablus, Palestine.**

2021

**Design and Implementation of Digital System for Lung
Cancer Early Detection Using Neural Network and
Image Processing**

By

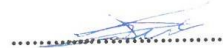
Deema Sohrab Sawalha

This thesis was defended successfully on 16/02/2021, and approved by:

Defense Committee Members

Signature

– Dr. Adnan Salman \ Supervisor



– Prof. Khalid Khanfar \ External Examiner



– Dr. Samir Matar \ Internal Examiner



Dedication

This work would not have been possible without the help of my family, who supported me in every way one could imagine....

My father and Mother I love you

My Husband for without him I wouldn't be standing here Today

My sisters and brother

My dearest friend Nisreen

And to my sons for they are my greatest motivators, Zaid and Omar.

Acknowledgement

I would like to express my sincere gratitude to staff of the Departments of Computer Sciences and Mathematics for their support.

To the committee for taking the time to review my work and give me their much-appreciated notes and remarks.

Also, I would express my deep gratitude to my supervisor Dr. Adnan Salman for his guidance and instructions and his dedication to help me through my study.

I would especially thank Dr. Baker Abdulhaq, Dr. Samir Matar, Dr. Amjad Hawwash, and Mr. Mohammed Adas for their encouragement and helpful suggestions.

الإقرار

أنا الموقعة أدناه مقدمة الرسالة التي تحمل العنوان:

**Design and Implementation of Digital System for Lung
Cancer Early Detection Using Neural Network and Image
Processing**

أقر بأن ما اشتملت عليه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه فيما ورد، وأن هذه الرسالة ككل، أو أي جزء منها لم يقدم من قبل لنيل أية درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

Student's Name:..... اسم الطالبة:.....

Signature:..... التوقيع:.....

Date:..... التاريخ:.....

Table of Contents

No.	Contents	Page
	Dedication	III
	Acknowledgement	IV
	Declaration	V
	Table of Contents	VI
	List of Tables	VIII
	List of Figures	IX
	Abstract	XI
1	Chapter One: Introduction	1
1.1	Thesis Objectives	5
2	Chapter Two: Related Works	6
3	Chapter Three: Methodology	23
3.1	Preprocessing Stage	25
3.1.1	Read DICOM files	25
3.1.2	Convert pixels' values to HU	25
3.2	Image Segmentation Stage	29
3.2.1	Image Analysis	29
3.2.1.1	Gabor Filters	29
3.2.1.2	Edge detection	33
3.2.2	Image Binarization	37
3.2.2.1	Histogram and Binarization	37
3.2.2.2	Morphological Operations	41
3.2.3	Watershed segmentation	47
3.3	Features Extraction Stage	49
3.4	Classification and Machine Learning Stage	56
3.4.1	Read Annotations Step	56
3.4.2	Classification Step using Neural Networks	66
3.4.3	Classification Step using Convolutional Neural Networks on HU matrix	72
4	Chapter Four: Results	78
4.1	Preprocessing stage	80
4.2	Image Segmentation stage	81
4.3	Features Extraction stage	89
4.4	Classification and machine learning stage	90
4.4.1	Reading annotations and extract features	91
4.4.2	Classification step	97
4.4.3	Classification Step using Convolutional Neural Networks on HU matrix	108
5	Chapter Five: Discussion and Conclusion	110

No.	Contents	Page
	References	116
	الملخص	ب

List of Tables

No.	Caption	Page
Table.3.1	Section of Properties' Output	52
Table.3.2	Section of MAT file for ROI Intensity values (HU matrix)	53
Table.3.3	Section of MAT file for Original DICOM values in ROI (DICOM matrix)	54
Table.3.4	Summary of MATLAB functions used in image processing stages	55
Table.3.5	Features extracted for classification step	61
Table.3.6	All Features extracted at once	62
Table.4.1	The 6 features extracted for scan LIDC-IDRI-0004	95
Table.4.2	All the features extracted at once as row vector	95
Table.4.3	Sample of the output CSV file	97
Table.4.4	A sample of the input dataset	98
Table.4.5	Comparison Table of the used classification algorithms with the dataset	100
Table.4.6	SVM Confusion Matrix	101
Table.4.7	SVM Classification Report	101
Table.4.8	CART Confusion Matrix	102
Table.4.9	CART Classification Report	102
Table.4.10	KNN Confusion Matrix	102
Table.4.11	KNN Classification Report	102
Table.4.12	LR Confusion Matrix	103
Table.4.13	LR Classification Report	103
Table.4.14	LDA Confusion Matrix	103
Table.4.15	LDA Classification Report	103
Table.4.16	NB Confusion Matrix	103
Table.4.17	NB Classification Report	104
Table.4.18	NN Confusion Matrix	104
Table.4.19	NN Classification Report	104
Table.4.20	CNN Confusion Matrix	109
Table.4.21	CNN Classification Report	109

List of Figures

No.	Caption	Page
Fig. (3.1)	16-bit DICOM image	25
Fig. (3.2)	HU Image	28
Fig. (3.3)	Speckle noise	30
Fig. (3.4)	Set of 16 Gabor Filters	31
Fig. (3.5)	Image after applying Gabor Filters	32
Fig. (3.6.a)	Canny Edge	36
Fig. (3.6.b)	Edge smoothing using Sobel	36
Fig. (3.7)	Image Histogram	38
Fig. (3.8)	THRESH_TOOL interface from MATLAB Central File Exchange GUI	39
Fig. (3.9)	Binarization using Otsu's Method	40
Fig. (3.10.a)	Lung Mask	40
Fig. (3.10.b)	Lungs from Image in (Fig.3.2)	40
Fig. (3.11)	Mask and lungs after applying Morphological operations.	42
Fig. (3.12)	Disk-shaped SE with Radius =1	43
Fig. (3.13)	Opening and closing by Reconstruction	44
Fig. (3.14)	Regional Maxima after Reconstruction	45
Fig. (3.15)	Regional Maxima superimposed on Original Image	45
Fig. (3.16)	Clear Borders and Extract ROI	46
Fig. (3.17)	Binarized ROI after taking image complement of Fig (3.16)	46
Fig. (3.18)	Colored Watershed Label Matrix	48
Fig. (3.19)	Watershed Labels imposed on Original Image	48
Fig. (3.20)	Extracted ROI	49
Fig. (3.21)	An example of annotated lobulation CT image, where the red rectangle indicates the lobulation region annotated by the radiologist	59
Fig. (3.22)	Scan Information and its total number of annotations	63
Fig. (3.23)	Nodule's Contour	64
Fig. (3.24)	Nodule's Centroid	64
Fig. (3.25)	Boolean mask of the nodule	65
Fig. (3.26)	Sigmoid Function	70
Fig. (3.27)	CNN architecture made of several 3D layers	74
Fig. (4.1.a)	DICOM image	81
Fig. (4.1.b)	HU Image	81

No.	Caption	Page
Fig. (4.2)	Result after applying 26-Gabor filters	81
Fig. (4.3)	Edge Detection Methods	82
Fig. (4.3.a)	Canny	82
Fig. (4.3.b)	Prewitt	82
Fig. (4.3.c)	Roberts	82
Fig. (4.3.d)	Sobel	82
Fig. (4.4)	Image Histogram	83
Fig. (4.5)	Binarization using	84
Fig. (4.5.a)	Global Thresholding	84
Fig. (4.5.b)	Adaptive Thresholding	84
Fig. (4.5.c)	Otsu's Thresholding	84
Fig. (4.6)	Lung Mask	85
Fig. (4.7)	Lungs from multiplying Lung mask with the original input image	85
Fig. (4.8)	Applying Morphological operations to create a mask and multiply it with the original input image	86
Fig. (4.9)	Final output image after Morphological Operations	86
Fig. (4.10)	Regional maxima extraction	87
Fig. (4.11)	Regional Maxima superimposed on the original image	87
Fig. (4.12)	Clear Borders and eliminate unwanted ROIs	87
Fig. (4.13)	final ROI in the binary image	87
Fig. (4.14.a)	Watershed labeled regions	88
Fig. (4.14.b)	Labels imposed on the original input image	88
Fig. (4.15)	Final ROI labeled	88
Fig. (4.16)	Original DICOM Image	92
Fig. (4.17)	Nodule centroid marked in red	93
Fig. (4.18)	Scan information and nodule contours shown in red	94
Fig. (4.19)	Boolean mask on right, original nodule on left	96
Fig. (4.20)	The box and whisker plots of the dataset's variables	98
Fig. (4.21)	The Histograms of the variables in the dataset	99
Fig. (4.22)	Classification Algorithms Comparison Box Plot	101
Fig. (4.23)	Model accuracy	105
Fig. (4.24)	Model Loss	105

Design and Implementation of Digital System for Lung Cancer Early Detection Using Neural Network and Image Processing

By

Deema Sohrab Sawalha

Supervisor

Dr. Adnan Salman

Abstract

Lung cancer is the most common type of cancer tumors among males worldwide. It accounts for 1 in 5 cancer death and occurs often between the ages of 55 and 65. This is also the case in Palestine, where lung cancer accounts for 22.8% among males. Early detection of Lung cancer at initial stages is a crucial step in the treatment process, where survival rate can significantly increase. In this thesis, we applied image processing techniques, using MATLAB and parallel processing on Computed Tomography (CT) images of lung cancer for several patients to identify cancer regions' location and size. The steps included analyzing and segmenting the images, extracting features and isolate candidates with size less than 3 mm as separate regions of interests (ROI). A comparative study between different image processing algorithms was also conducted on several images to identify the most accurate algorithms to be used in the lung cancer screening process. Machine learning algorithms and Neural Networks were used to classify cancer from the candidates ROIs by the extraction of attributes used for classification of the pathologic features for the marked annotations using Python. The tested algorithms were compared to determine the most suitable algorithm for detecting cancer in certain ROI.

Chapter One

Introduction

Chapter One

Introduction

Lung cancer is a disease characterized by uncontrollable growth of abnormal cells into a tumor. They start off in one or both lungs, in the cells around the air passages. These abnormal cells divide and multiply rapidly to form a tumor. As the size of the tumor becomes larger, it prevents the lung from doing its function of providing the bloodstream with oxygen. Tumors that do not spread and stay isolated in one place are considered benign. When a tumor spreads to other parts of the body through the bloodstream or the lymphatic system and destroys healthy tissues, the result is a severe condition and generally is difficult to treat. According to the World Health Organization, cancer accounts for 14% of all global deaths and lung cancer is the number one killer among all types of cancer. It has the smallest survival rate after diagnosis [13].

Depending on their cellular characteristics, there are two types of lung cancer, small-cell lung cancer and non-small-cell lung cancer. Small-cell lung cancer accounts for 20% of lung cancer and its main cause is tobacco smoking. The main characteristic of this type is its rapid spread to other parts of the body through the bloodstream and lymph nodes [12].

Once the lung cancer is diagnosed, medical centers develop a treatment process based on the cancer type and its stage. Knowing the stage is extremely important in the development of the appropriate treatment

procedure. Staging the lung cancer uses the TNM system. In this system, the letter T (Tumor) describes the size of the tumor and where it is located in the lung or the body. The letter N (lymph Node) indicates whether the cancer has spread to lymph nodes, and the letter M (Metastasis) indicates whether the cancer has spread to other parts of the body. A number from 0 to 4 is assigned to each factor. A higher number indicates the severity of the cancer [12].

After using the TNM system, each type is staged differently. Small-cell lung cancer staged to be either in limited stage or in extensive stage. The limited stage indicates that cancer is found in one lung and possibly nearby lymph nodes, but it hasn't spread past the lung. The extensive stage indicates that cancer already spread to the other lung or to other parts of the body. The stages of non-small cell lung cancer can be in one of the following stages: 1) Occult stage: cancer cells are found but no tumor is found in the lungs by imaging tests, 2) Stage 0: in this stage the cancer is very small and hasn't spread, 3) Stage 1: Cancer presents in the lung tissue but lymph nodes are not affected, 4) Stage 2: Cancer has spread to nearby lymph nodes or into the chest, 5) Stage 3: Cancer continue to spread from the lungs to the lymph nodes and nearby organs, and 6) Stage 4: cancer has spread throughout the body and may affect the liver and bones[12].

Lung cancer is by far a quite disease in its early stages. There are no symptoms or warning signs, which makes it harder to detect before it actually develops to advanced stages. The 5-years survival rate is 14% for

lung cancer detected at any stage. However, if detected at early stages, before it has spread to other parts of the body, the 5-years survival rate become 55%. Therefore, early detection of lung cancer is extremely important in the treatment process and it can save many lives [12].

In this thesis, we propose a system to detect lung cancer from CT images. The proposed system consists of four main stages, preprocessing stage, segmentation stage, features extraction stage, and machine learning stage. In this system, CT images are passed through the preprocessing stage for image enhancement and noise removal. In the second stage, image is segmented to define boundaries between different tissues. In the third stage, important features are extracted. In the fourth stage, a machine learning procedure is applied to the extracted features to identify cancer regions.

lung cancer in early stages. However, the proposed system involves the application of several procedures. In each procedure several different algorithms can be applied, and for each algorithm there are several parameters that affect the accuracy and precision of the algorithm. Another goal of this study is to evaluate the accuracy and quality of the application of different algorithms to obtain the most accurate results.

CT images are considered the most suitable among imaging techniques to detect cancer in early stages, because of its ability to form three-dimensional images of the chest, resulting in greater resolution of nodules. [28][29]

By comparing the results obtained by the system with those obtained from online datasets [1], the cancer tumor can be identified and detected. In this study we use several CT images for patients with lung cancer. These images are taken from the Cancer Imaging Archive database [1].

1.1. Thesis Objectives:

1) General Objective:

The general objective of this research is to design and implement a system that can detect the presence of lung cancer in CT images enhance the performance and increase the detection accuracy.

2) Specific objectives:

- Design and implement a system using image processing techniques to detect the presence of lung cancer in CT images.
- Design and implement a neural network system to classify whether there exists lung cancer or not, feed the network with labeled dataset to learn the features that help classify cancer tumors.
- Feed the output of image processing techniques into CNN to train the system to classify the output into two categories; cancer and not-cancer.

3) Research questions and identified problems:

- How to analyze, segment, extract features of images?
- What data to extract and use for classification?
- Which classification algorithms can be used?

Chapter Two

Related Works

Chapter Two

Related Works

Several authors have used image processing techniques for lung cancer early detection.

Sharma and Jindal (ICCTAI' 2011) [7] developed a Computer Aided Diagnosing (CAD) system for early detection of lung cancer based on automatic diagnosis of the lung regions included in Chest CT Images. The methods they used were denoising images using Wiener filters, extracting lung region of interests by applying segmentation algorithms such as image slicing algorithm to DICOM CT images, binarization, and Erosion filter. Segmenting extracted lung region to simplify the representation of image into something easier to analyze, performing edge detection steps using Sobel Methods. Isolated various desired shapes of an image with feature extraction and small thresholding values. They focus on the size of cluster detected and allowed the diagnosis of 3mm in diameter nodules, with 90% sensitivity and with 0.05 false positive per image. The overall accuracy achieved was 80%.

Nunzio, Tommasi, and colleagues (2011) [19] proposed a completely automated segmentation algorithm, which is a fully 3D to ensure coherence between adjacent CT slices. They calculated an appropriate gray-value threshold for segmentation, by analyzing the image histogram. A simple-threshold 3D Region Growing (RG) was applied to the CT volume to give a binary mask. They extracted the external airways by wavelet simulation

model resulting in a mask containing only the lungs, and thus, handling the Hilar Region extraction problem. They separated the lungs from each other using the Region Growing method, resulting in a left mask and a right mask. These masks do not enclose nodules and vessels because their density is large compared to the lung parenchyma. So, to ensure the inclusion of lung nodules, they chose to adopt 3D binary morphological closing with a fixed size of 30mm diameter spherical structure element so the resulting mask would cover all pulmonary parenchyma and all lung nodule types, in each lung. After this step, the closing mask was applied to the original CT image and the difference gave the ROI of nodules that can be searched.

They achieved 96% with accurate segmentation of the sample, 100% sensitivity for nodule inclusion in the segmentation algorithm.

Al-Tarawneh (Leonardo Electronic Journal of Practices and Technologies Issue 20, January-June 2012) [3] took image quality and accuracy as the core factors of his research. The main detected features for accurate images' comparison were pixel percentage and mask-labelling. He followed a four-stages system, starting with image Acquisition, applied several techniques for image Enhancement, then applied image Segmentation algorithms, and fourth stage was obtaining general features from enhanced segmented image to indicate normality or abnormality of an image. For image Enhancement, he used and compared between three algorithms, Gabor filter (the best), Auto Enhancement algorithm, and FFT.

For image Segmentation, he used thresholding and Marker-Controlled Watershed segmentation approach (gave better results). And for the Feature Extraction stage, he used Binarization and Masking approach, combining those two approaches led to take a decision whether the case is normal or abnormal counting the black and white pixels in the ROI in the final image.

Kuruvilla, Gunavathi, (2013) [22] presented a CAD classification system in CT images of lungs developed using Artificial Neural Network. The lung was segmented using morphological operations, the grayscale is converted to binary, the threshold level was calculated by Otsu's method, then a morphological opening operation was performed with a structuring shape of 'periodic line'. The image was edge cleared with 98% accuracy for segmenting images correctly.

Extracting statistical parameters from ROI, like, mean, standard deviation, skewness, kurtosis, fifth and sixth central moments and using them as features for classification in the artificial neural network. They used a feed-forward neural network and feed-forward back propagation neural network to train the data set. It consisted of a layer of an input unit and a single output unit, with one hidden layer. They used thirteen training algorithms for the backpropagation neural network to update the weights and biases values.

Then they proposed two new training algorithms; training function 1 adjusts momentum factor and the learning rate according to the Gradient Descent which resulted in increasing the classification accuracy up to

93.3% compared to the 'traingdx' with 91.11% maximum accuracy among the thirteen tested algorithms, and training function 2 modifying function 1 to reduce the mean square error also by adjusting the variables according to the Gradient Descent achieved 0.0942.

The statistical parameter skewness of the features used for classification gave the maximum classification accuracy with an increase of 5 – 8%. The training function 1 gave an accuracy of 93.3% with a specificity of 100% and a sensitivity of 91.4%. the performance of the proposed CAD system was good with a sensitivity of 82%.

Ada and Kaur, (2013) [24] developed a solution for the detection of lung cancer using image processing algorithms. Image quality was the core of the research. In this paper, features were extracted using principal component analysis (PCA) and Histogram Equalization was used for image preprocessing.

The system consisted of four stages: image preprocessing and enhancement stage to improve the interpretability or perception of information included, so they used Histogram Equalization. The feature extraction stage to detect and isolate different desired shapes of a given image using the binarization approach. The threshold was chosen based on the Gray Level Co-occurrence Method (GLCM). The following features were extracted: contrast energy, entropy, homogeneity, and maximum probability determining whether the case is normal or abnormal.

The Principal Component Analysis (PCA) used to standardize the data in an image, and the features extracted were passed through the PCA for better classification. The proposed technique gave very promising results compared with other used algorithms.

Sowjanya, Bharath, and Yadav, (2013) [26] proposed a system to detect lung cancer nodules focusing on the steps used to detect the nodules to get accurate results. There were three main steps: image enhancement to provide better input to other automated image processing techniques. Gabor filter was used to extract spatially localized spectral features, followed by auto enhancement techniques and FFT to adjust the image's color, brightness, and contrast to optimum levels depending on statistical calculations such as mean and variance. Next was the image segmentation to assign labels to pixels sharing certain visual characteristics, they used Thresholding approach to convert the image into binary image, then applied Marker-Controlled Watershed approach to segment unique boundaries from the image. The last step was features extraction to determine normality or abnormality of an image by using two approaches; the binarization approach which depends on the number of black and white pixels, and the second approach is Masking; which depends on the fact that white connected areas inside ROI increase the percent of cancer presence. The final image appears either as a solid blue color indicating normal case, while the appearance of many RGB masses indicates presence of cancer.

They used the peak signal to noise ratio (PSNR) to determine which approach was better for enhancing the image and the result came up with Gabor filter as the best method to use, and they used it again to check which segmentation technique is better and it was the Watershed method that returned higher values of PSNR.

The proposed model has the following stages: image pre-processing stage, using the Median filter and Gaussian filter. Segmentation stage using the Watershed Segmentation. Features extraction stage which increased the features extracted like, Centroid, Diameter, and pixel Mean Intensity, increasing the accuracy of detecting cancer nodules.

As for classification stage, to determine whether the cancer is benign or malignant they applied the Support Vector Machine (SVM) algorithm. The extracted features were used as the training features to generate the model. The detection accuracy is 92% and the classifier accuracy is 86.6% which was not performed in other models studied by the authors. There are still some limitations on the model; the accuracy is still not 100%, and it can't specify the degree of cancer.

Patil and Jain (Vol. 3 Issue 4 March 2014) [2] used image processing techniques on CT images to detect lung cancer cells to get accurate results in early stages using various enhancement and segmentation techniques such as Thresholding and Watershed Transform. They followed many steps including Image Acquisition, Image Enhancement, Image

Segmentation, and Feature Extraction. They compared two methods for solving the problem; watershed segmentation and thresholding approach.

They used gray scale converted CT images in MATLAB environment, removing noises and enhancing the quality of CT image, for that, Gabor filter enhancement technique was used. In segmentation stage, thresholding approach used to get better recognition of regions of interest separating background from foreground and detecting edges and boundaries, they used Watershed segmentation. For feature extraction stage, the binarization and masking approaches were used to determine the abnormal masses in lungs which appeared as RGB masses in the final image. The Achieved quality using Watershed was 85.27% and better than Thresholding approach which resulted in 81.24%.

Gajdhane and Deshpande (IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 5, Ver. III (Sep – Oct. 2014)) [4] aimed to get more accurate results from CT images using MATLAB, by using various enhancement and segmentation techniques, to help early detection of lung cancer cells. They used Denoising and Wiener filtering in the Enhancement stage, segmentation and pixel grouping into objects with similar characteristics by Watershed algorithm. Thresholding is used for the feature extraction stage using rule-based steps. They evaluated cancer's size and its penetration into surrounding tissues to determine the best treatment. They used Median filter for denoising image, Gabor filter for enhancing an image, Marker-Controlled Watershed algorithm, and

extracted the features: area, perimeter, and eccentricity measured in a scalar. These values help determine at which stage the cancer is. For classification, the SVM approach was used. They concluded that following this system on CT images and with some modification, the same steps can use X-Ray and MRI images as well to determine which best helps to identify cancer cells at early stages.

Kumar and Kumar, (2014) [23] presented an accurate lung parenchyma segmentation system by using Region Growing Algorithm for early detection of any parenchymal disease. The segmentation algorithm consisted of seven main steps: image preprocessing, histogram analysis threshold, region growing with threshold, edge detection, morphological filling, multiply the mask with the original image.

The region growing algorithm starts with a seed pixel, as the initial beginning of regions of interests, then the regions are grown from the seed points to adjacent pixels with the same intensity value as the seed points. It separated the background from the lung area. The algorithm was tested on 20 patients' CT images and all were accurately segmented.

Miah and Abu Yousef (May 2015) [8] took as a basis for their work the Lung Cancer Detection and Classification using Machine Learning and Multinomial Bayesian, estimating of lung cancer using image segmentation and back propagation. For image acquisition, they took JPEG image format. Preprocessing steps of gray scale conversion, normalization, denoising, binarization, and removing unwanted regions of image. For

image segmentation, performed edge detection technique and converted the image into dilated one. Thresholding methods taking three threshold values. In Feature extraction stage, they checked how black pixels are distributed in the image and the distances between these pixels. The neural network detection features are for classification and detection purposes. The resulted system achieved an accuracy of 99% by binarization technique and 96.67% using Neural Network Detection techniques.

Magdy, Zayed, and Fakhr, (2015) [20] talked about a CAD system to analyze and automatically segment the lungs and classify each lung into cancer or normal. The system consisted of five main steps. For preprocessing, Wiener Filter is applied, then they combined histogram analysis with thresholding and morphological operations to segment the lung regions. Amplitude-Modulation Frequency-Modulation (AM-FM) method is used to extract ROI and used Partial Least Squares Regression (PLSR) used to select from extracted features for the classification step. AM-FM methods provide pixel-based transformation in terms of instantaneous amplitude (IA), instantaneous frequency (IF), and instantaneous phase (IP) unlike the Fourier Transforms that provide the frequency content of the signal.

They used and compared four classifiers (K-Nearest Neighbor, Linear, Support Vector Machine, and Naïve Bayes) to obtain the best accuracy, sensitivity, and specificity. The Linear Classifier returned 95% for accuracy, 94% for sensitivity, and 97% for specificity.

Orozco, Villegas, Sánchez, Domínguez, and Alfaro, (2015) [25] focused on the design of a computer-aided diagnosis (CADx) system that would be helpful for assisting radiologist as a second opinion to classify lung nodules and to reduce the time of the CT scan evaluation. the nodules were characterized by the computation of the texture features obtained from the gray level co-occurrence matrix (GLCM) in the wavelet domain and were classified using an SVM with radial basis function in order to classify CT images into two categories: with and without cancerous lung nodules. The novelty of the paper is the elimination of the typical structure segmentation stage, this is because the detection of candidate lung nodules is carried out by means of a wavelet transform. Another novelty of the system is the use of wavelet features to describe the lung nodules and that the only preprocessing stage performed is the extraction of a ROI.

They made a performance comparison of 11 recent works including their work related to CADx systems and different methodologies were used to create these CADx system. The stages of the proposed method to design CADx are: extraction of ROI using Hough transform, wavelet transform was used instead of image segmentation to transform the ROI in images from spatial domain to transform domain with Discrete Wavelet Transform (DWT), they chose the Daubechies db1, db2, and db4, the third stage is feature extraction where the Grey Level Co-occurrence Matrix (GLCM) was used to extract second order statistical texture features of each wavelet sub-band in order to characterize the nodules. A set of 19 texture features

were extracted from four different angles of GLCM. The next stage was attribute and sub-band selection; it allows to automatically search for the best subset of attributes in the features vector, so it reduces the feature set using the Waikato Environment for Knowledge Analysis software, using the CfsSubsetEval method and the features were reduced to 11.

And the last stage is the classification it used the SVM algorithm to classify the nodules from 2 mm to 30 mm in diameter. The method gave 90.90% sensitivity which is better than many methods compared in the paper.

Singh, Singh, and Vijay (IOSR Journal of Computer Engineering (IOSR-JCE), 2016) [10] focused on feature extraction stage in image processing techniques, in order to define cancer cells and determine Malignancy. Preferring CT images in their study due to better clarity, low noise and distortion. They eliminated all unwanted parts of chest, such as bones, heart, in the preprocessing stage. After that, lung nodule obtained as Region of Interest, features are calculated for classification stage using Neural Network. The study focused on techniques that operate in Spatial Domain. Histogram Equalization was used for enhancement stage, the Median filters for image denoising, and Otsu's Method for thresholding the CT image. Morphological Operations were performed to fill the indentation caused by the pulmonary vessels in Segmentation stage and isolating the ROI by its shape, size, and texture of a given image.

Singh and Asuntha, (2016) [21] aimed in this paper on developing an efficient system able to detect lung cancer in early stages, and further

stages if cancer had spread to other organs through the Ultrasound images, based on an automatic diagnosis of lung regions in CT, MRI, and Ultrasound images.

The preprocessing stage was applied by a Gabor filter to smooth and remove noise from the image. Layer separation was performed to reduce complexity and proper conversion to gray-level images. Edge detection performed with Canny Adaptor due to its fast performance and suitability for real-time detection, the edges were dilated to locate edge regions and enlarge them in order to be removed by masking in later stages.

For the segmentation stage, a super-pixel segmentation algorithm was used because the super-pixel level reduced the computational cost and makes the detection of the cancer area faster. The basic idea of their super-pixel method was introducing an initial seed growth based on pixels' similarity, maintaining the simple linear clustering manner. To enforce connectivity between irregular cancer-prone regions, they assigned a new label to the disjoint segment, and thus the small cancerous region can show up as a single super-pixel, the reason for this step as mentioned is because most methods relabel pixels with the labels of the larger neighboring cluster which may be a normal region. Thus, it is hard to detect cancerous from this super-pixel since most pixels are noncancerous. Features extracted were the shape, area of interests, size of nodules, contrast enhancement, and calcification and they were used to increase the accuracy of detecting small lung cancer nodules. The authors noted for a future work

that they can add a classification stage using Pearsons and Spearman algorithm and SVM algorithm.

Cheela (International Journal of Engineering and Advanced Technology (IJEAT) Volume-6 Issue-3, February 2017) [9] reviewed the algorithms for Enhancement, Segmentation, and Feature Extraction to detect lung cancer cells in different sizes and compared between those algorithms using accuracy, sensitivity, and specificity. He used Mean, Median, Adaptive Median, and Weiner filters in the preprocessing stage. The Deep Neural Network classification algorithms gave more accuracy. Higher sensitivity was achieved with the Wavelet-Based Support Vector Machine (SVM), and specificity was in the LDA classifier method for large tumors and having three segmentation masks for small tumors. Recent papers (2012- 2016) for his comparison were taken into consideration.

Makajua, and Prasad (6th International Conference on Smart Computing and Communications, ICSCC 2017, Dec 2017, Procedia Computer Science 125 (2018) 107–114) [16] aimed in their research at evaluating the various computer-aided techniques, by analyzing the current best technique and finding out their limitations and drawbacks. Then they proposed a new model with improvements in the current best model. There were no 100% accuracy results, so their target was to achieve an increase towards 100%.

The proposed model used the Median Filter and Gaussian Filter instead of the Gabor filter in the preprocessing stage. The image

segmentation with Watershed to mark cancer nodules. Features extracted like area, perimeter, eccentricity, Centroid, diameter, and pixel mean intensity for the detected cancer nodules. For the classification stage, they used the Support Vector Machine algorithm on the extracted features and trained the model to predict benign and malignant.

They increased the accuracy up to 92% of lung cancer detection and classified the cancer type as benign or malignant with 86.6%. They also pointed to two limitations in their model, the first was about the accuracy still not 100%, and the second was about the classifier not being able to specify which stage the cancer nodules belonged to.

Perumal, Velmurugan (2018) [18] proposed a method for detecting lung cancer in early stages and with high accuracy. It consisted of three processes; the preprocessing is carried with image enhancement using the Adaptive Histogram Equalization and Gabor filter within the rules of Gaussian. The segmentation using Sobel. The extraction process using a 5-level HAAR wavelet transformation technique. The third process is the classification using Enhanced Artificial Bee Colony (EABC) optimization.

It detected the lung nodules around 3mm diameter at an earlier stage and differentiates between cancerous and non-cancerous nodules using the EABC approach with a detection accuracy of 92.4% and False Error Rate (FERR) of 7.6.

Bhalerao and colleagues (2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (P.577-

P.583)) [46] used image processing techniques and Convolutional Neural Network (CNN) to early detect lung cancer and classify cancerous images from non-cancerous. Their proposed system converted images from RGB to gray scale then to binary images, these images were used as input for CNN model to predict whether lung image was cancerous or not. They used Deep Learning as a new branch of Artificial Intelligence to enhance the performance of CNN based systems. Their system took also in consideration the processing power and time delay of the cancer detection process for efficiency.

The proposed system was implemented in MATLAB and the cumulative accuracy obtained was 94.34%.

Muthazhagan, Ravi, and Rajiniginath, (2020) [47] applied feature extraction and proper combination of Adaptive thresholding in image processing stage using CT images as inputs. For classification stage, they used SVM and Content-Based Image Retrieval technique (CBIR) was used to compare lung image features such as contrast, intensity, texture, and shape.

CBIR retrieves lung images from online repository for relevant patient records based on their proposed feature dataset that aimed on lesions found in the lung image, location spread of the lesion, and area of lesions. The proposed method produced 98% prediction accuracy and measured exact tumor areas in lung CT images. Also, they added a treatment recommendation based on the CBIR output with choices that

might include surgical procedures, chemotherapy, radioactivity therapy, and specific usage of drugs for certain time periods.

In general, most of the related works used three to four stages systems. The difference mainly was in the algorithms used. The proposed system consists of a four-stages system similar to the mentioned methods in the related work. The aim of this thesis is to enhance the design of some stages in a way to achieve higher accuracy and improve performance.

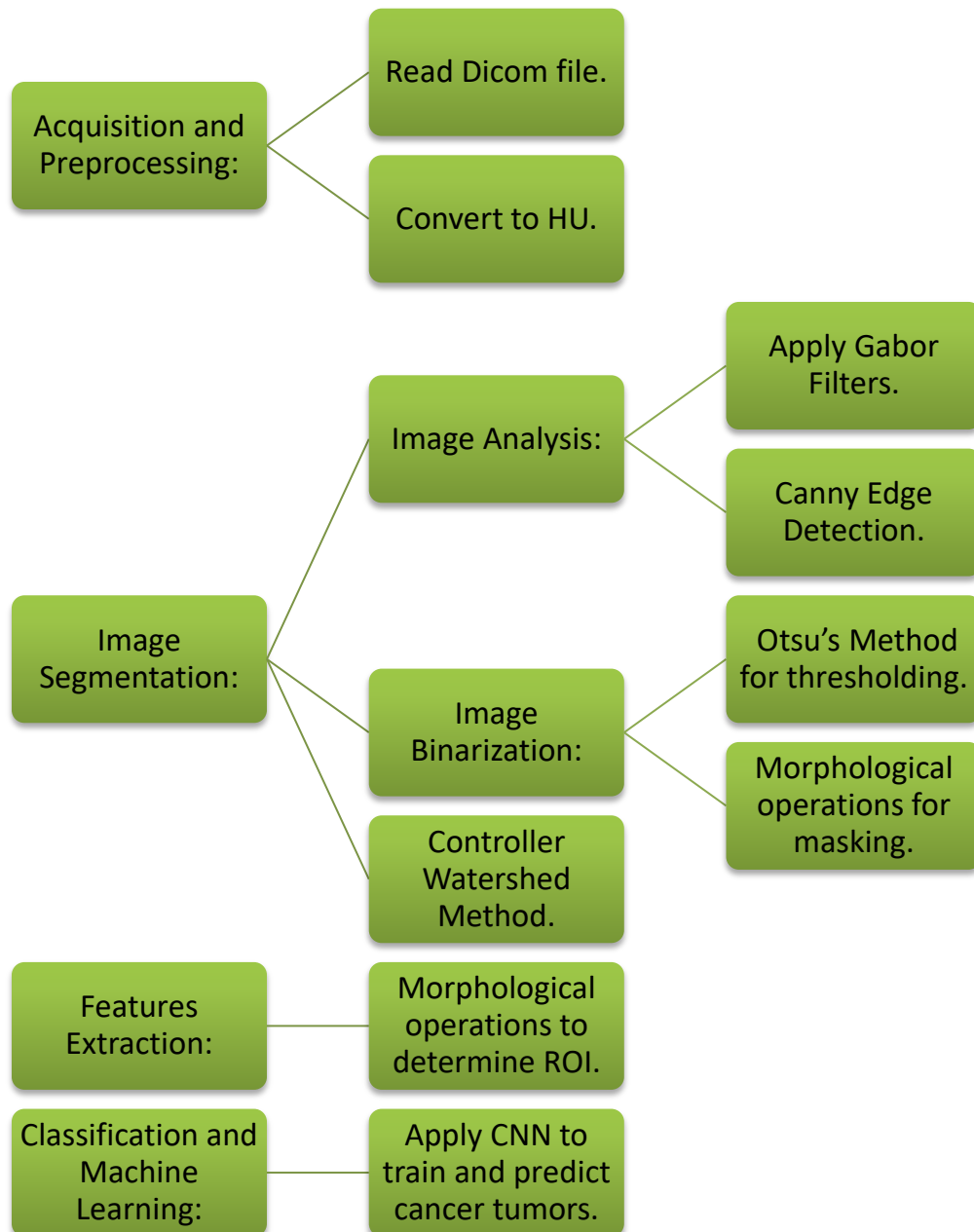
Chapter Three

Methodology

Chapter Three

Methodology

The proposed system consists of a four-stage system described in the following block diagram:



3.1. Preprocessing Stage:

3.1.1. Read DICOM files:

Each patient's data is processed in a single iteration so that all the properties and Region of Interests (ROI) for a patient are stored separately. At the beginning of each iteration, the number of images per patient is known in order to process all the images and extract ROIs in all slices. The size and spacing between slices are the same for all patients. The size of each image is 512X512 pixels, and the depth between slices equals 1. Using the command *dicominfo()*, the number of rows, the number of columns, and the samples per pixel for each image can be obtained. Then the pixel values of the image are read and stored in a matrix for processing and analyzing the image.

3.1.2. Read HU images:

The pixels of the original image have Hounsfield Unit (HU) values stored by the machine that was used to take the CT scan. They make up the grayscale in medical CT imaging.

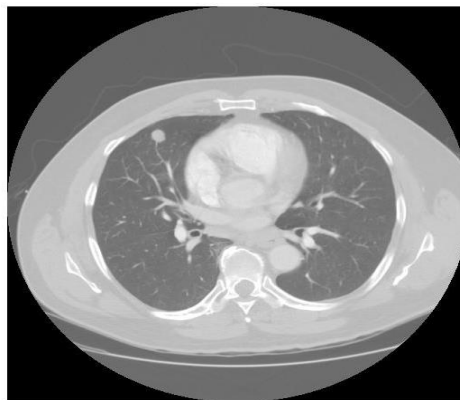


Fig.3.1. 16-bit DICOM image

It is a scale from black to white of 4096 values (bit depth of 12-bit) and ranges from -1024 HU to 3071 HU. In the HU scale, -1024 HU represents black color and corresponds to air (in the lungs), 0 HU corresponds to water (since human body consists mostly of water, there is a large peak at this value), 3071 HU represents white color and corresponds to the densest tissue in a human body like tooth enamel. All other tissues are within this scale, for example fat is around -100 HU, muscle around 100 HU and bone spans from 200 HU (trabecular/spongy bone) to about 2000 HU (cortical bone). Metal implants typically have very high Hounsfield units. Therefore, they have been attributed the maximum value in typical 12-bit CT scans (3071) [48].

The Hounsfield unit (HU) scale is a linear transformation of the original linear attenuation coefficient measurement into one in which the radiodensity of distilled water at standard pressure and temperature (STP) is defined as zero Hounsfield units (HU), while the radiodensity of air at STP is defined as -1000 HU. In a voxel with average linear attenuation coefficient μ , the corresponding HU value is therefore given by [54]:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu - \mu_{air}} \quad (3.1)$$

where μ_{water} and μ_{air} are respectively the linear attenuation coefficients of water and air. Thus, a change of one Hounsfield unit (HU) represents a change of 0.1% of the attenuation coefficient of water since the attenuation coefficient of air is nearly zero.

MATLAB by default converts the 12-bit range to 16-bit when reading the DICOM file (Figure (3.1)). To convert the image back to its original HU values, the following linear equation is used:

$$Image_{HU} = Slope \times Image_{MATLAB} + Intercept \quad (3.2)$$

where $Image_{HU}$ is the output in HU unit, $Image_{MATLAB}$ is the image with a bit depth of 16-bit, $Slope$ is the Rescale Slope read from the DICOM information file and equals 1, and $Intercept$ is the Rescale Intercept read from a DICOM information file and equals -1024.

The rescale slope and rescale intercept allow to transform the pixel values to HU or other units as specified in the tag (0028,1054) stored in DICOM header.

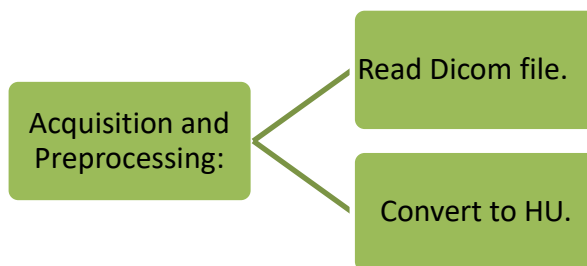
With CT images, the unit should be HU (Hounsfield) and the default value is HU if the tag (0028,1054) is not present. However, the tag may be present and specify a different unit (OD=optical density, US=unspecified). The rescale slope and intercept are determined by the manufacturer of the hardware. If the transformation from original pixel values to Hounsfield is not linear, then a LUT (Lookup Table) is applied. The output image (Figure (3.2)) has attributes in a matrix of the same size as the original image.



Fig.3.2. HU Image

Two matrices will be conducted with pixels' values of both bit depths, 12-bit and 16-bit. The first matrix of bit depth 12-bit is for HU values and will be referenced by HU matrix in this thesis. The second matrix is for pixels' values with bit depth of 16-bit and will be referenced by DICOM matrix.

The following diagram shows a summary of the steps in this section.



And the following pseudo code explains the steps to read the DICOM files from the images and convert the image from 16-bit width into HU intensity value of 12-bit depth:

START

Read DICOM folder and subfolders

Read DICOM files

read DICOM header information

dicominfo(fileName)

read image and store in matrix

dicomread(fileName)

Convert 16-bit image into 12-bit ...

(Hounsfield Unit intensity values of pixels)

3.2. Image Segmentation Stage

Segmentation stage is applied to analyze the acquired image, remove noise, detect edges and isolate regions of interests for features extraction stage.

3.2.1. Image Analysis:

3.2.1.1. Gabor Filters

Gabor filter is used to enhance the image quality along with a Gaussian filter to smooth the image and remove speckle noise (Figure (3.3) if found.

Speckle noise is the noise that arises due to the effect of environmental conditions on the imaging sensor during image acquisition. Speckle noise is mostly detected in the case of medical images, active Radar images, and Synthetic Aperture Radar (SAR) images [30].

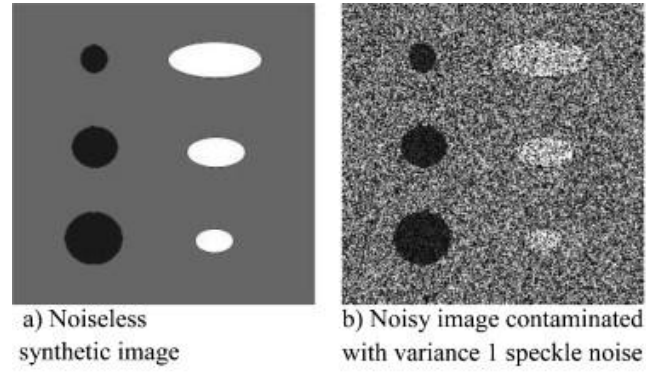


Fig.3.3. Speckle noise ^[32].

Source: Latifoğlu, Fatma: A novel approach to speckle noise filtering based on Artificial Bee Colony algorithm: An ultrasound image application

Gabor filter is bandpass linear filter used for feature extraction and texture analysis. It basically analyzes whether there is any specific frequency content in the image in specific directions in a localized region around the point or region of analysis (interest).

Here the Gabor filter is considered as two outputs of phase filters allocated in the real and the imaginary parts of the complex function. The real part holds the filter:

$$g_r(t) = \omega(t)\sin(2\pi f_o t + \theta) \quad (3.3)$$

And the imaginary part holds the filter:

$$g_i(t) = \omega(t)\cos(2\pi f_o t + \theta) \quad (3.4)$$

Where θ is the angle between the direction of the sinusoidal wave and the x-axis of the spatial domain, and f_o is the central frequency of the filter.

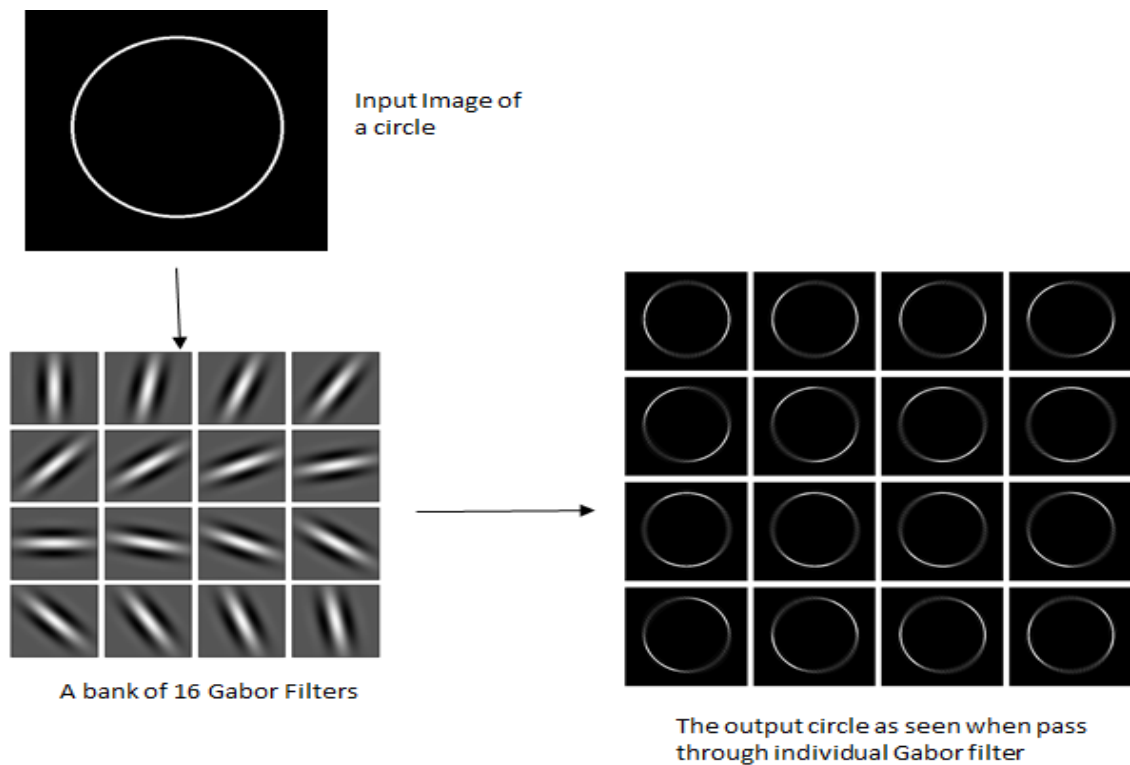


Fig.3.4. Set of 16 Gabor Filters ^[33].

Source: M.Janani, K.Nandhini , K.Senthilvadivel ,S.Jothilakshmi: Digital Image Technique using Gabor Filter and SVM in Heterogeneous Face Recognition

By extending these functions to two dimensions, it is possible to create filters that are selective for orientation. Under certain conditions, the phase of the response of Gabor filter is approximately linear.

In order to establish a multi-resolution strategy, the image can be filtered with a set of N Gabor filters with different bandwidths and modulation frequencies (Figure (3.4)). In this research, angle orientation changes with 45° on the positive y-axis (between 0° - 180°) to observe any change in every quarter of the frequency plane. We could consider the entire plane from 0° - 360° , but this will give repeated frequencies with opposite directions. The set of frequencies and orientations is designed to localize different, roughly orthogonal, subsets of frequency and orientation

information in the input image. Regularly sample orientations between $[0,180]$ degrees in steps of 45 degrees. Sample wavelength in increasing powers of 2 starting from $(4/\sqrt{2})$ up to the hypotenuse length of the input image [31]. The total number of Gabor filters used are 28.

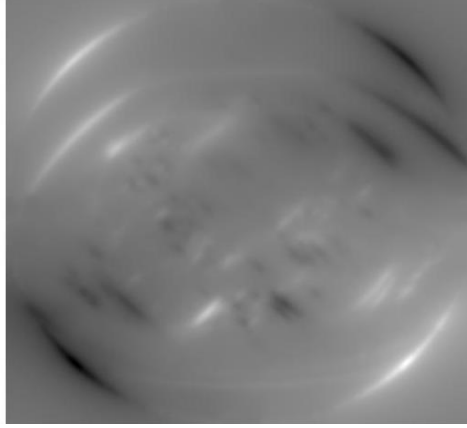


Fig.3.5. Image after applying Gabor Filters.

The resulting Gabor magnitude responses are used as features for defining ROI. Spatial information is added by applying a Low-Pass Gaussian filter to smooth the output. Then, the feature is reshaped to a form similar to input image using Principal Component Analysis (PCA). The output features are then normalized using the mean value of the pixels and their standard deviation.

PCA is used to transform the 30D representation (28 Gabor filters and 2 spatial features) of each pixel in the input image to a 1D intensity value for each pixel. PCA is a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components. It can be defined also as a dimension-reduction tool that used to reduce a large set of variables to a

small set that still contains most of the information in the large set. *pca* is used to eliminate extra features and extract a smaller number of features with more information to determine regions of interest.

Spatial location information is mapped by the image pixels' coordination's values X and Y for grouping pixels spatially close together. The image after applying Gabor filter is shown in figure(3.5).

The steps used to remove noise is shown in the following pseudo code:

```
% Apply Gabor filter

    Define wavelength

    Define Theta

    Define orientation range ( from 0° - 180°)

    Define the set of filters according to orientation range

        Filters = gabor(wavelength, orientation)

    For i from 1 to length of filters % here parallel processing is used

        Apply gabor filter

        Smooth the result with Gaussian filter

    Reshape the image into same size as input image
```

3.2.1.2. Edge detection:

Regions of interests should be extracted accurately as much as possible to help detect lung tumors in early stages. The output image after grouping features is processed with an edge detection algorithm. It's needed to have a method that shows a wide range of edges in the image in order to detect ROI and obtain an early detection of nodules.

Edge detection is an image processing technique for finding the boundaries of objects within an image. It works by detecting discontinuities of brightness in pixels' values. Many methods are used for edge detection, such as Prewitt, Sobel, Canny, and fuzzy logic methods. The most appropriate method that experimentally tested in the system gave the best results was the Canny algorithm.

Canny Method is an edge detection operator that uses a multi-stage algorithm to detect wide range of edges in an image with noise suppressed at the same time. The first step is smoothing the image with a Gaussian filter to reduce noise and unwanted details and textures.

$$g(m, n) = G_o(m, n) * f(m, n) \quad (3.5)$$

$$\text{Where } G_o = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{m^2+n^2}{2\sigma^2}\right)$$

The second step is computing the gradient of $g(m, n)$ using the Sobel gradient operator. It can be implemented by convolving the image with Sobel kernels K_x and K_y , respectively:

$$K_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad K_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

To get the following magnitude

$$M(m, n) = \sqrt{(g_m^2(m, n) + g_n^2(m, n))} \quad (3.6)$$

And slope

$$\theta(m, n) = [g_n(m, n)/g_m(m, n)] \quad (3.7)$$

Then compute the threshold

$$M_T(m, n) = \begin{cases} M(m, n) & \text{if } M(m, n) > T \\ 0 & \text{otherwise} \end{cases} \quad (3.8)$$

Where T is the threshold value chosen, so that all edge elements are kept while most of the noise is suppressed.

The final step is thinning edge ridges by checking each non-zero $M_T(m, n)$ with its two neighbors. If it is greater along the gradient direction keep the value unchanged, otherwise set it to zero. Then edge segments are linked together to form a continuous edge by thresholding the thinned edge by two threshold values τ_1 and τ_2 (where $\tau_1 < \tau_2$) to obtain two binary images T_1 and T_2 . Note that T_2 with greater threshold τ_2 has less noise and fewer false edges but greater gaps between edge segments when compared to T_1 with smaller threshold τ_1 . Then, edge segments in T_2 are linked together to form a continuous edge. To do so, each segment in T_2 was traced to its end and then its neighbors were searched in T_1 to find any edge segment in T_1 to bridge the gap until reaching another edge segment in T_2 .

There are different edge detection techniques that can be used in image processing [27]:

- Sobel operator is one of the pixel-based edge detection algorithm. It can detect edge by calculating partial derivatives in 3 x 3 neighborhoods. The reason for using Sobel operator is that it is insensitive to noise and it has relatively small mask in images.

- Prewitt Operator is similar to the Sobel operator and it is used for detecting vertical and horizontal edges in images. The Prewitt edge detector is an appropriate way to estimate the magnitude and orientation of an edge. It is estimated in the 3 x 3 neighborhood for eight directions. The entire eight masks are calculated and the one with the largest module is selected.
- Canny edge detector is one of the most accurate technique because it addresses good detection, good spatial localization, and good response rate; giving the optimal solution for edge detection issue.

In this research Canny method was used (Figure (3.6.a)), because accuracy of detecting edges between the lungs and other chest organs and within the lungs is very important to obtain optimal edge detection of the ROI, and Canny Method gave the most suitable results compared to other methods when applied on the images. The final output of applying Canny algorithm is shown in Figure (3.6.b).

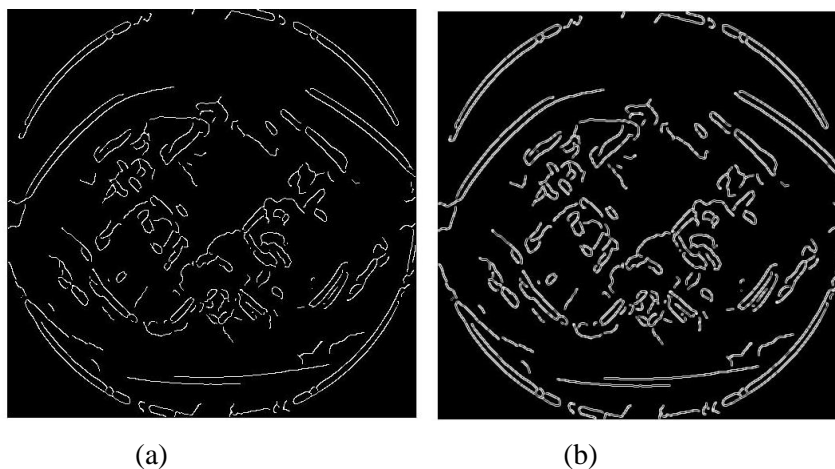


Fig.3.6. (a). Canny Edge. (b). Edge smoothing using Sobel.

The following pseudo code shows the steps to determine edges in order to eliminate unwanted parts and separate the candidate regions of interests:

Apply Canny method for edge detection

Edge(image, 'canny')

Sharpen and clear the edges with Sobel method

Apply gradient(edged_image, 'sobel')

3.2.2. Image Binarization:

3.2.2.1. Histogram and Binarization:

In order to extract the region of interests, we need to reduce the area studied specifying some constraints on the image. One of these methods is binarizing the image to separate background as the unwanted areas and foreground which contains the regions specified for more study. This binarization method is achieved by choosing a threshold value to separate pixels with smaller value to the background, and pixels with bigger values are considered foreground.

Many methods and algorithms, such as, Otsu's thresholding, Global thresholding, and Local (Adaptive) thresholding were used to apply this step on an image in this research. Otsu's method converts a grayscale image to monochrome in image processing. It involves iterating through all the possible threshold values and calculating a measure of spread for the pixel levels each side of the threshold, i.e., the pixels that either fall in

foreground or background. The aim is to find the threshold value T , where the sum of foreground and background spreads is at its minimum. The method uses the image intensity histogram to read the number of pixels at each grayscale level. The command is *otsuthresh(counts)*, where counts are the number of bins resulting from image histogram analysis (Figure (3.7)).

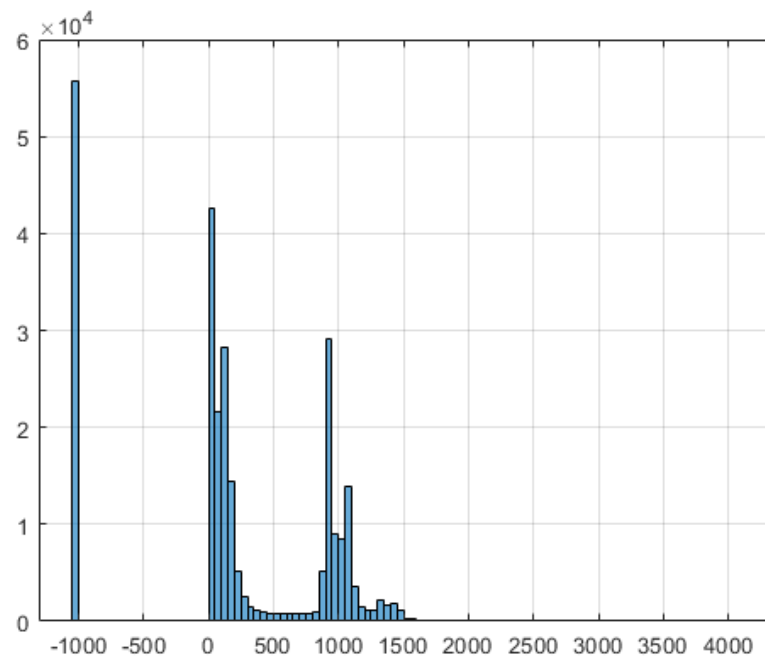


Fig.3.7. Image Histogram

Another method is the Global thresholding. It is based on the assumption that the image has a bimodal histogram and, therefore, the object can be extracted from the background by a simple operation that compares image values with a threshold value T . the command used is *imbinarize(image)*, where image is in gray level, it chooses the threshold value to be in the middle of intensity values of pixels as possible. THRESH_TOOL from MATLAB Central File Exchange (Figure (3.8)) launches a GUI (graphical user interface) for thresholding an intensity input image using the Global method.

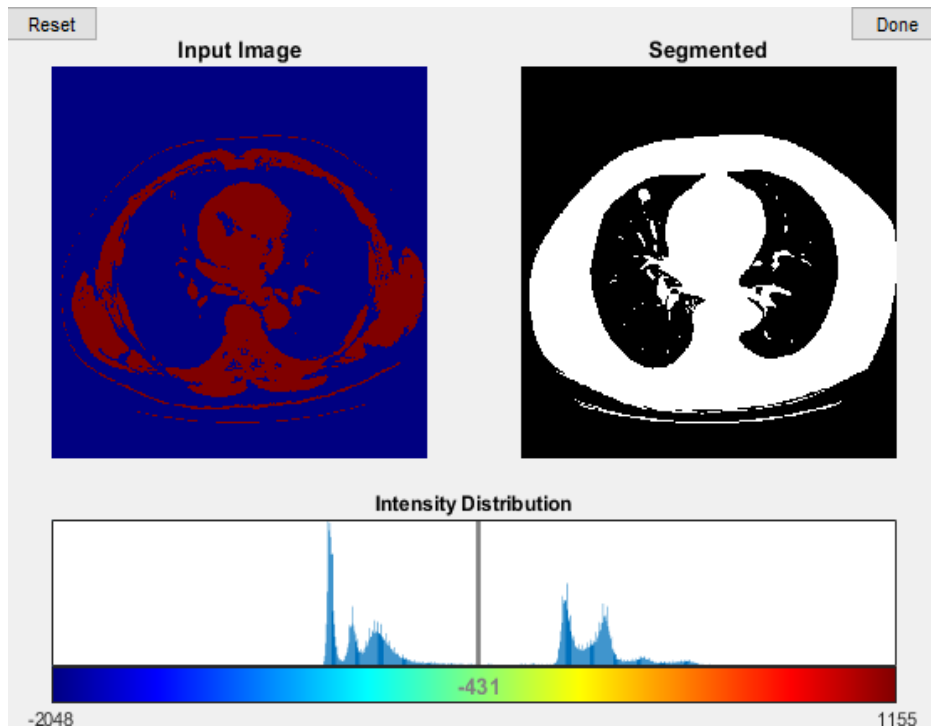


Fig.3.8. THRESH_TOOL GUI interface from MATLAB Central File Exchange

It was used to find the most suitable threshold value in this research since it allows the user to evaluate the threshold value that well separates the image pixels into background, and foreground. After many experiments, the best averaged threshold value obtained, which achieved optimal separation of foreground from background, was approximately '0.4947395'.

This function gave the best threshold value when processing a single image, but the result was not as good when processing many images at once. This is because for each image, a manual selection for the threshold is required making the performance time too long. For this reason, Otsu's method is the one used in this research.

The result is a binary image, with logical pixel values of 0 or 1 (Figure (3.9)), used to create a binary mask to separate lungs from all organs in the chest.

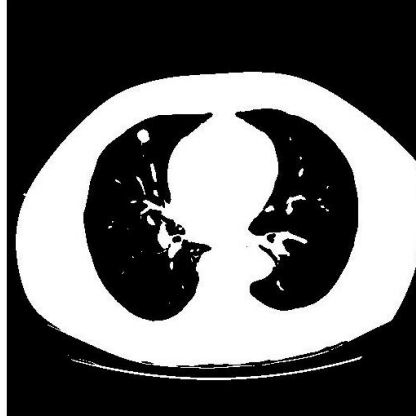


Fig.3.9. Binarization using Otsu's Method.

The binary output image is used to create a mask for lungs (Figure (3.10.a)) which will be used to multiply the original image by this mask to get the ROI (Figure (3.10.b)).

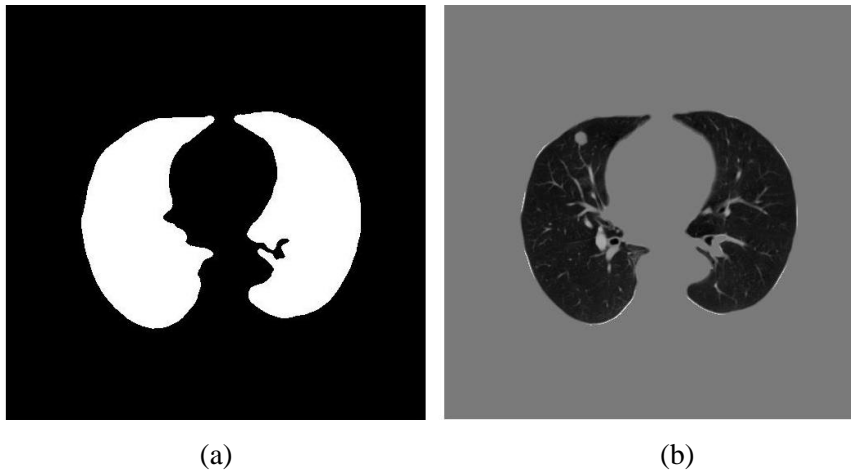


Fig.3.10. (a) Lung Mask. (b) Lungs from Image in (Fig.3.2)

3.2.2.2. Morphological Operations:

Morphological operations process images based on their shapes. Basically, these operations apply a structural element on the input image, and the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors (Figure (3.11.a)).

The main morphological operations are erosion, dilation, opening and closing. Dilation adds pixels to the boundaries of objects in an image if any of the neighboring pixels have the value 1 then the pixel is set to 1, while erosion removes pixels on object boundaries if any of the neighboring pixels have the value 0 then the pixel is set to 0. 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.

The opening operation erodes an image and then dilates the eroded image, using the same structuring element for both operations. It is useful for removing small objects from an image while preserving the shape and size of larger objects in that image. The closing operation dilates an image and then erodes the dilated image, using the same structuring element for both operations. It is useful for filling small holes from an image while preserving the shape and size of the objects in that image. The morphological outputs are multiplied by the original image as shown in Figure (3.11).

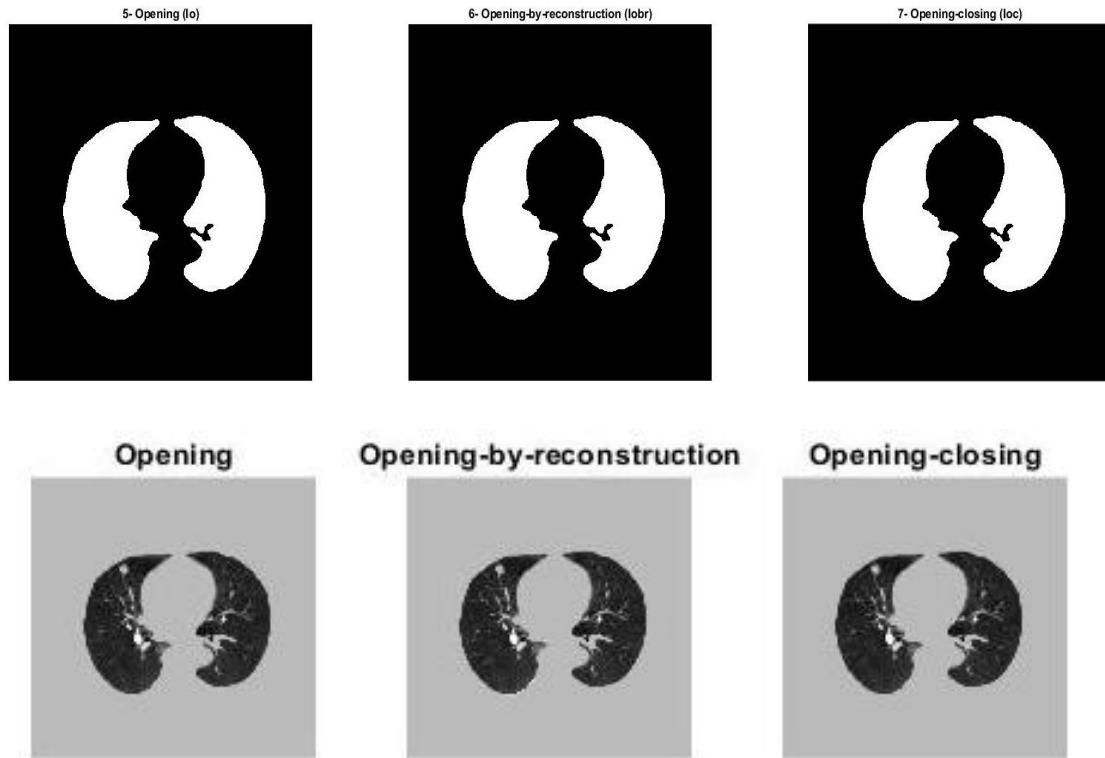
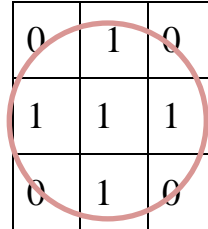


Fig.3.11. Masks and lungs after applying Morphological operations.

A structuring element is a matrix that identifies the pixel in the image being processed and defines the neighborhood used in the processing of each pixel (Figure (3.12)). Typically, a structuring element is chosen of the same size and shape as the objects processed in the input image. In this research, we look for disk-shaped areas to help classify tumors from other parts within lungs. Benign tumors have smoothed circular shape, and malignant tumors have circular shape with sharper edges [34].

The parameters needed to define a structural element are: 1) the shape, it is set to disk and 2) the neighboring radius, it defines the number of neighboring pixels to compare the central pixel's value with. For this step in extracting the ROI, two structural elements were used. One to be applied on dilation operations with radius equals to 4, and the other is

applied on erosion operations with a radius of value 1. The radii values were chosen based on experimental studies that gave appropriate results for this research.



0	1	0
1	1	1
0	1	0

Fig.3.12. Disk-shaped SE with Radius =1.

The input image is processed first to find its complement. Then dilation operation with a structural element of size 1, followed by an erosion operation with a structural element of size 4. The output of these operations is an image with thickened edges. The next step is clearing the edges to extract the lungs without other organs and bones of the chest. Then, a filter based on area is applied to extract all connected components from the binary image within the range 2. This range specifies the minimum and maximum sizes that an object needs to achieve to be extracted. After the extraction of the lungs, the open areas inside them must be filled to get a mask of full lungs.

Another erosion operation is applied to enhance the lungs' edges. The binary image is multiplied by the original input image to get the lungs' intensity mask. This mask is then multiplied by the original image to get the original values of pixels inside the lungs as shown in Figure (3.11).

The output image is then processed with another set of morphological operations. The opening operation followed by erosion operation to

remove pixels from edges with applied structuring element of radius 1. A reconstruction operation is applied to the resulting image (the marker) by the intensity mask (the mask) using 8-connected neighborhoods as shown in Figure (3.13). Morphological reconstruction can be thought of as repeated dilations of an image, called the marker image, until the contour of the marker image fits under a second image, called the mask image. In morphological reconstruction, the peaks in the marker image “spread out,” or dilate.



Fig.3.13. Opening and closing by Reconstruction

A closing operation is then performed to fill any small holes with the structuring element of radius 1, followed by another dilation operation. A reconstruction operation is applied with the complement image of the dilated output as the marker, and the complement of previously reconstructed image as the mask.

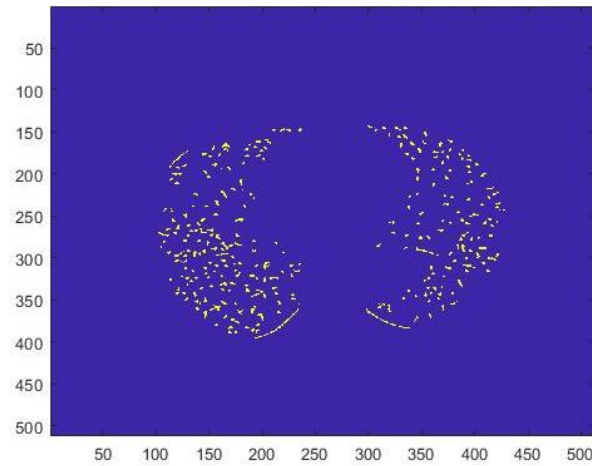


Fig.3.14. Regional Maxima after Reconstruction

The complement of the output is taken to get back the regions inside the lungs as the foreground areas of interest (Figure (3.14)). The maximum ROI is taken as the final output of this step and a closing operation followed by an erosion operation is applied to sharpen the ROI edges as shown in Figure (3.15) after multiplying the regional maxima with the original image.

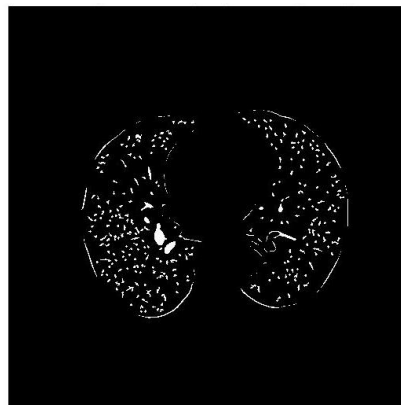


Fig.3.15. Regional Maxima superimposed on Original Image.

The next step eliminates all connected components with a total number of pixels less than the specified value 5 (Figure (3.16) and Figure (3.17)). This value was chosen to reduce any fault detection of small vessels and lymph nodes as tumors as much as possible.

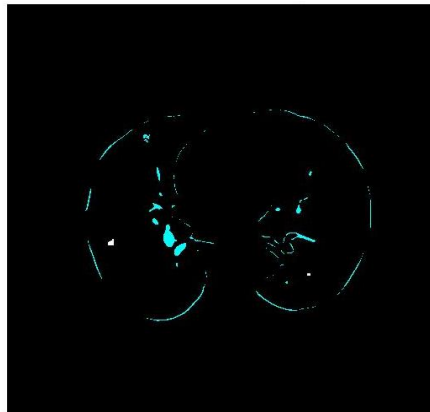


Fig.3.16. Clear Borders and Extract ROI.

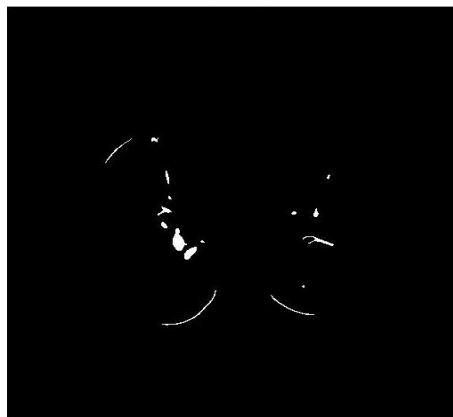


Fig.3.17. Binarized ROI after taking image complement of Fig.3.16.

The following pseudo code shows the steps to binarize the image and apply the morphological operations to extract ROI:

Binarize image

apply otsu_threshold(image)

Extract Regions of Intesets ROI

mask lungs

take complement of image

apply dilation, erosion with disk structuring element of radii=1,4

fill holes inside lungs

apply morhplogical operations

Opening

Dilation

Erosion

Closing

eliminate unwanted objects with pixel values more than 5

image reconstruction

take region maxima

3.2.3. Watershed segmentation:

This step segments the output image from the previous morphological step to isolate each connected component and consider it a separate ROI to extract features for the classification stage. The most used segment method is Watershed [31][26].

Watershed, in image processing, is a transformation in grayscale images. The aim of this technique is to segment the image. When two regions-of-interest are close to each other, that their edges touch. The Watershed Transform finds "catchment basins" and "watershed ridge lines" in an image by treating it as a surface where light pixels are high and dark pixels are low surfaces.

Segmentation using the watershed transform works better if you can identify, or "mark," foreground objects and background locations. To do so, first the Euclidean distance is computed between each pixel and the nearest nonzero pixel of the binary output image from the last step, one at a time. Then, watershed segmentation is applied on the image. Finally, the impose

minima is found which modifies the grayscale mask image obtained from the edge detection step, see Figure (3.18). Using morphological reconstruction, it only has regional minima wherever binary marker image - yielded from applying the watershed method- is nonzero.

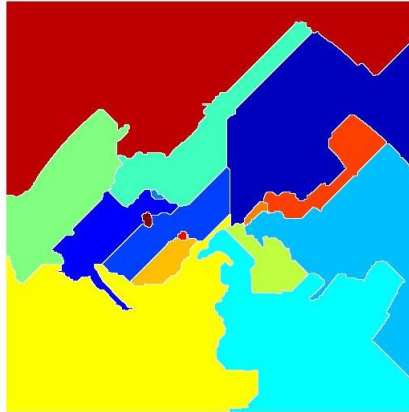


Fig.3.18. Colored Watershed Label Matrix.

Another watershed segmentation step is applied to segment the impose minima image in order to extract regions of interest that meet the desired conditions the most as shown in Figure (3.19). Erosion operation is performed to remove any holes within the objects, and this is the final output processing of the image before the output is taken to extract features from each object for the classification stage (Figure (3.20)).

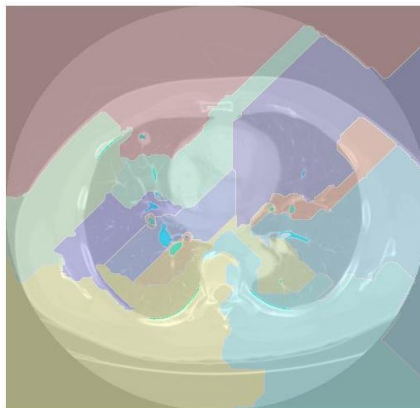


Fig.3.19. Watershed Labels imposed on Original Image.

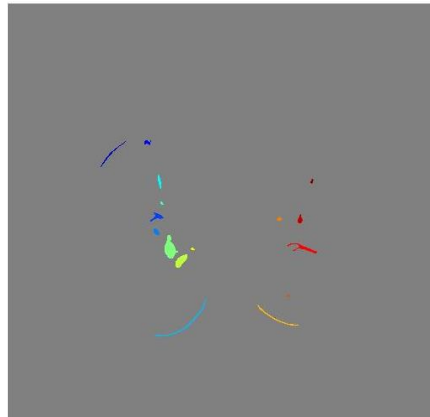
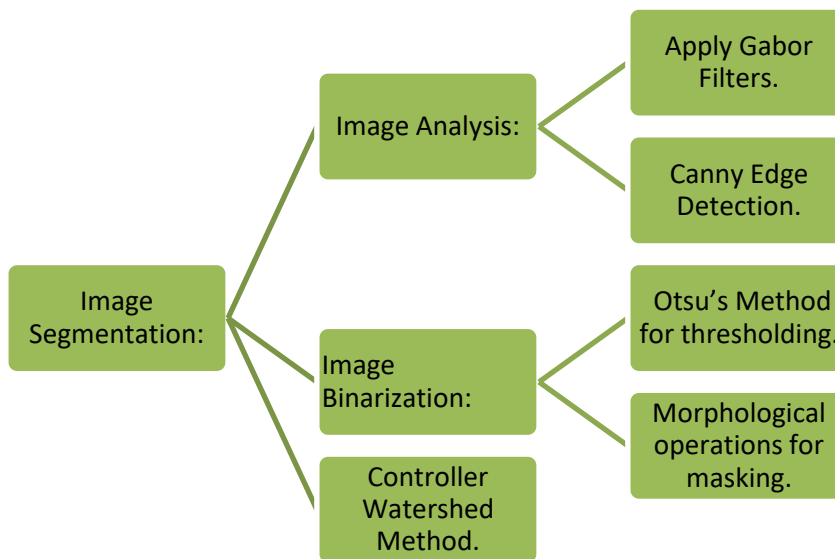


Fig.3.20. Extracted ROI.

The following diagram is a summary for the steps used in this section.



3.3. Features Extraction Stage:

The final output of image processing techniques is used to determine the candidate connected components and store each candidate as a separate item to be studied in the classification stage.

This study and analysis depend on the properties of ROI extracted, such as the 'Area', 'Centroid', 'Equivalent Diameter', 'Eccentricity', 'Euler Number',

'Filled Area', 'Major Axis Length', 'Minor Axis Length', 'Orientation', 'Perimeter', and 'Image'. They are computed by the MATLAB command *regionprops(image)* for each connected component in the binary image and stored in a table for later use. A brief definition of these features is the following:

Area: It is the actual number of pixels in a ROI

Centroid: It is the center of mass of the region, it has two values, the first element is the centroid in the x-coordinate, and the second element is the y-coordinate.

Equivalent diameter: It is the diameter of a circle with the same size as the ROI.

Eccentricity: This is for an ellipse that has the same points' distribution as the ROI. It is the ratio of the distance between the center of the ellipse and its major axis length. It has a value between 0 and 1. When the eccentricity is 0 it means that the ellipse is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.

Euler number: It is the number of objects in the region minus the number of holes in those objects. It is useful to know the connectivity between pixels within the object.

Filled area: It is the number of pixels in the region's image.

Major and Minor axis lengths: They are the lengths of the major axis and the minor axis respectively of the ellipse that has the same points' distribution as the region of interest.

Orientation: It is the angle between the major axis and the x-axis of the region in degrees. It ranges from -90° to 90° .

Perimeter: It is the distance around the boundary of the region. It is computed by calculating the distance between each adjoining pair of pixels around the border of the region.

Image: It is the region of interest with the same size as the bounding box created by the function, and it has binary pixel values 1 correspond to the region and all other pixels are 0.

The values computed are returned as a structure so we need to convert them to table in order to read the values and use the file in the classification stage.

The diameter of each region is needed to specify the length of the vectors used to store the pixels' values of each extracted image. The diameter is computed by taking the mean of the major and minor axes lengths. Then the radius is computed and stored in the properties table for each region, along with the folder number for each patient and the slice number being processed in the current loop.

The properties extracted from this step is stored in a csv file. A sample is shown in Table(3.1).

Table.3.1. Section of Properties' Output.

FolderNum	ImageNum	Area	Centroid_1	Centroid_2	EquivDiameter	Perimeter
1	42	53	123.0754717	180.1132075	8.214724333	84.474
1	42	30	166.1333333	166.2	6.180387232	23.603
1	42	64	177.546875	254.875	9.027033337	41.897
1	42	38	176.8421053	272.7894737	6.955796338	20.229
1	42	115	210.8608696	381.7217391	12.10051849	159.128
1	42	40	180.55	212.9	7.136496465	32.824

Another table is made with the values of folder number, image number and the region of interest pixels' identifier numbers from 1 to 62500. This table is used for machine learning step. It is the input for the Convolutional Neural Network to detect whether the region is cancer or not. We can use the binary table as obtained from region properties function, but for more accuracy the HU matrix is used for classification.

The HU table is obtained by multiplying the binary mask with HU valued image $Image_{HU}$. A sample of this table is shown in Table (3.2) where each row identify a single ROI extracted and the cloumns are the pixels of each ROI numbered from 1 to 62500 pixels.

The features extracted and stored in a table is explained in the following pseudo code:

```

define connected components with neighbors with radius 4
extract geometrical features from image

```

```

regionprops('table', binarized_image, {'features extracted', '
ROI_image'})

compute diameters and radii for each ROI

export to table 'properties_table'

for each ROI

multiply ROI with original HU_image

save as row vector

append in HU_matrix

for each ROI

multiply ROI with original DICOM_image

save as row vector

append in DICOM_matrix

```

Table3.2. Section of MAT file for ROI Intensity values (HU matrix).

Pixel#	4851	4852	4853	4854	4855	4856	4857
ROI 1	642	436	0	0	0	0	0
ROI 2	10	10	21	14	0	0	3
ROI 3	0	4	63	99	92	81	53
ROI 4	102	235	292	275	106	0	0
ROI 5	0	0	0	0	0	0	0
ROI 6	0	115	265	235	0	0	0
ROI 7	0	54	126	0	0	0	0

A boundary detection process is applied on the output of the multiplication above with no holes inside the object using the command *bwboundaries* (binary-mask). This function returns the minimum and maximum boundary values for both the x-coordinate and y-coordinate of

the region, and these values are used to extract each element if the slice has more than one ROI to isolate each region as a single object and store as a row vector for later analysis in ROI array. This array is stored as a MAT file

The DICOM table is obtained from multiplying the binary mask by the bit depth input image of 16-bit (Image_{MATLAB}) to get the intensity values of that image. This array is also stored as a MAT file.

Table.3.3. Section of MAT file for Original DICOM values in ROI (DICOM matrix).

Pixel#	4851	4852	4853	4854	4855	4856	4857
ROI 1	1666	1460	0	0	0	0	0
ROI 2	1034	1034	1045	1038	0	0	1027
ROI 3	0	1028	1087	1123	1116	1105	1077
ROI 4	1126	1259	1316	1299	1130	0	0
ROI 5	0	0	0	0	0	0	0
ROI 6	0	1139	1289	1259	0	0	0
ROI 7	0	1078	1150	0	0	0	0

Each region of interest has different dimensions. It is stored as a row vector in the array containing all ROIs for the classification. So, all regions have the same dimensions which is the biggest extracted region. It has the dimensions of 250X250 pixels for each region, and when flatten the region into a row vector it will be of the size 62500, plus a column determining a label whether the region is cancerous or not, and three other columns containing folder number, image number, and ROI number, with total size

of 62504. A sample of this table is shown in Table (3.3) where each row identify a single ROI extracted and the cloumns are the pixels of each ROI numbered from 1 to 62500 pixels. Regions of bigger size are neglected since the aim of this research is detecting tumors in early stages with small sizes of 3mm at least. This ensures that any candidate ROIs will be extracted.

The following diagram is a summary for the steps used in this section.



The following table is a summary of MATLAB functions used in image processing stages:

Table.3.4. Summary of MATLAB functions used in image processing stages.

#	Command	Description	In Page#
1	Dicominfo(filename)	reads the metadata from the compliant DICOM file specified in the string or character vector FILENAME. Returns a structure of attributes,such as number of rows and columns, slope and intercept.	p.21
2	Otsuthresh(counts)	computes a global threshold from histogram counts COUNTS that minimizes the intraclass variance for a bimodal histogram. Threshold returned is a normalized intensity value that lies in the range [0, 1]	p.32
3	Imbinarize(image)	Binarize grayscale 2D image or 3D volume by thresholding. It binarizes image with a	p.33

#	Command	Description	In Page#
		global threshold computed using Otsu's method. Returns a matrix.	
4	THRESH_TOOL(IM)	launches a GUI (graphical user interface) for thresholding an intensity input image. IM is displayed in the top left corner. A colorbar and IM's histogram are displayed on the bottom. A line on the histogram indicates the current threshold level. A binary image is displayed in the top right based on the selected level. To change the level, click and drag the line. The output image updates automatically and returned as 2D matrix. (from MATLAB Central File Exchange)	p.33
5	Regionprops(image)	measures a set of properties for each connected component (object) in the binary image, returns 2D matrix.	p.45
6	Bwboundaries()	traces the exterior boundary of objects, as well as boundaries of holes inside these objects. Returns 1D matrix	p.48

3.4. Classification and Machine Learning Stage:

The classification stage is divided into three steps: the first step is reading the annotations and extract the desired attributes (pathologic Features), the second step is the classification of these features using neural network algorithms to determine whether a nodule is cancerous or not, and the third step is using the output results of the model trained in the second step to test the output HU matrix that resulted from features' extraction stage in section 3.3 using CNN model.

3.4.1. Read Annotations Step:

The dataset used in this work is the LIDC-IDRI dataset [1], it consists of 1012 thoracic CT scans with nodule size reports and diagnosis

reports. Four radiologists reviewed each scan using two blinded phases. The results of each radiologist's unblinded review were compiled to form the final unblinded review. The LIDC radiologists' annotations include outlines of nodules ≥ 3 mm in diameter on each CT slice in which the nodules are visible, along with the subjective ratings scale of the following pathologic features: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy.

1. **Subtlety:** is a measure between 1 to 5 to determine the difficulty of detection. 1 stands for Extremely Subtle and 5 stands for Obvious.
2. **Internal Structure:** is the internal composition of the nodule, and it has a value between 1 and 4, where
 1. 'Soft Tissue'
 2. 'Fluid'
 3. 'Fat'
 4. 'Air'
3. **Calcification:** defines the pattern of how Calcium builds up in body tissue and around nodules. If calcification is present around a nodule, it takes a value of the following:
 1. 'Popcorn'
 2. 'Laminated'
 3. 'Solid'

4. 'Non-central'

5. 'Central'

6. 'Absent'

4. Sphericity: is the three-dimensional shape of the nodule in terms of its roundness, and it has a value in the range 1 to 5 where 1 is 'Linear' and 5 is 'Round'.

5. Margin: is the description of how well-defined the nodule's margin is. It has a value between 1 and 5 where 1 is 'Poorly defined' and 5 is 'Sharp defined'.

6. Lobulation: Lobulation is a category of CT imaging signs, which is dependent on the ingrowth of connective tissue septa containing fibroblasts derived from perithymic mesenchyme [50]. It is normally related with malignant lesion, though it also occurs in up to 25% of benign nodules [50]. Visually, a lobulation shows the indentation which appears at the edge of round or oval lesion, as shown in Figure (3.21). it ranges between 1 (No Lobulation) and 5 (Marked Lobulation).

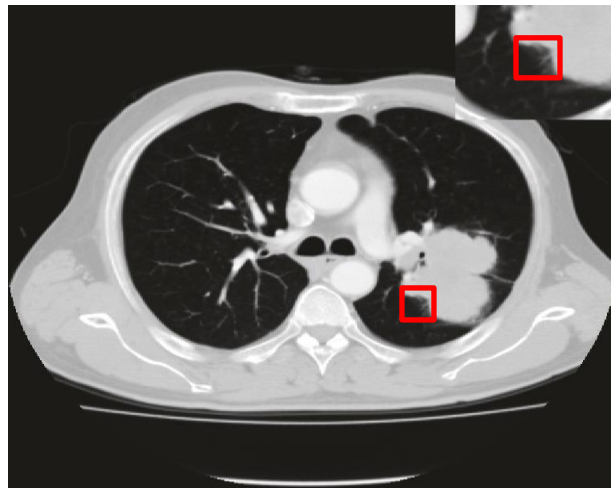


Fig.3.21. An example of annotated lobulation CT image, where the red rectangle indicates the lobulation region annotated by the radiologist. [50]

Source: Guanghui Han, Xiabi Liu, Nouman Q. Soomro, Jia Sun, Yanfeng Zhao, Xinming Zhao, and Chunwu Zhou: Empirical Driven Automatic Detection of Lobulation Imaging Signs in Lung CT,

- 7. Spiculation:** is a measure between 1 and 5 of stellate distortion caused by the intrusion of cancer into surrounding tissue. Its existence is a clue to characterizing malignant tumors, where 1 stands for ‘No Spiculation’ and 5 for ‘Marked Spiculation’.
- 8. Texture:** it determines the radiographic solidity of the nodule’s internal structure whether it is Solid, Ground Glass, or Mixed. It ranges between 1 ‘Non-solid/GGO’ and 5 ‘Solid’.
- 9. Malignancy:** is a subjective assessment of the likelihood of malignancy, this feature is used as the output class with a range between 1 and 5 where:
 1. ‘Highly Unlikely’
 2. ‘Moderately Unlikely’
 3. ‘Indeterminate’

4. ‘Moderately Suspicious’

5. ‘Highly Suspicious’

The annotations also include an approximate centroid of nodules $\leq 3\text{mm}$ in diameter as well as non-nodules $\geq 3\text{ mm}$ [1][49].

The maximum number of annotations obtained is up to 4 which is the same as the number of radiologists who reviewed the scans.

The classification stage is performed using Python 3.7.7 and its libraries. The main libraries used are pylidc [51] for analyzing and querying only annotation data and pydicom to access DICOM image data. The extraction of attributes that will be used for classification is conducted on the pathologic features extracted from the annotations marked by the radiologists in the LIDC-IDRI dataset.

Pylidc library: is an Object-relational mapping (using SQLAlchemy) for the data provided in the LIDC dataset. This means that the data can be queried in SQL-like fashion, and that the data are also objects that add additional functionality via functions that act on instances of data obtained by querying for particular attributes.

Pydicom library: is a pure Python package for working with DICOM files. It lets you read, modify and write DICOM data in an easy "pythonic" way.

For this research, the features used for classification as inputs are :subtlety, internal structure, calcification, sphericity, and margin, while malignancy is used as the output class and since the main purpose of this

research is to determine whether a nodule is cancer or not, the categories of malignancy feature is reduced to binary classes where categories 1 and 2 from the definition is converted to 0 output class (not-cancer), whereas 4 and 5 are converted to 1 output class (cancer), category 3 is eliminated in this research because it is defined as Indeterminate.

The features are returned in two methods:

1. Each feature can be returned separately from the other features according to the purpose of the research for each annotation as shown in Table (3.5).

Table.3.5. Features extracted for classification step.

Pathologic Features	Annotation1	Annotation2	Annotation3	Annotation4
Subtlety	5	5	5	5
Internal Structure	1	1	1	1
Calcification	6	6	6	6
Sphericity	3	4	3	5
Margin	3	4	2	4
Malignancy	3	5	5	4

2. All features can be called at once as a row vector arranged by default as explained in order in (Page.47) for each annotation as shown in Table (3.6).

Table.3.6. All Features extracted at once.

Features Annotations	#1	#2	#3	#4	#5	#6	#7	#8	#9
Annotation1	5	1	6	3	3	3	4	5	5
Annotation2	5	1	6	4	4	5	5	5	5
Annotation3	5	1	6	3	2	3	3	5	5
Annotation4	5	1	6	5	4	1	5	4	4

The root folder containing the LIDC-IDRI dataset is defined and all subfolders are called using for loop to read each scan annotations and extract the necessary features for the research and store them in a CSV file for later use in the classification stage.

Two main classes are used from the *pylidy* library, the Scan Class which refers to the top-level XML file from the LIDC-IDRI data, and the Annotation Class which belongs to the Scan Class and includes querying feature values, determine the contour-derived data and helps visualize the annotations.

A query is defined for each scan based on the patient ID read from the root folder, and the first scan is returned in the query results

```
scans = pl.query(pl.Scan).filter(pl.Scan.patient_id == pid).first()
```

The number of annotations each case study has is determined, knowing that the maximum number of annotations any case study should have equals 4 corresponding to the number of radiologists who reviewed the dataset, and each annotation has its pathologic features based on the radiologist who reviewed and marked the scan. Estimated annotations referring to the same physical nodule in the CT scan are clustered by the command and shown in Figure (3.22).

```
nodules = scan.cluster_annotations()
```

This function uses a distance function to create an adjacency graph to determine which annotations refer to the same nodule in a scan.

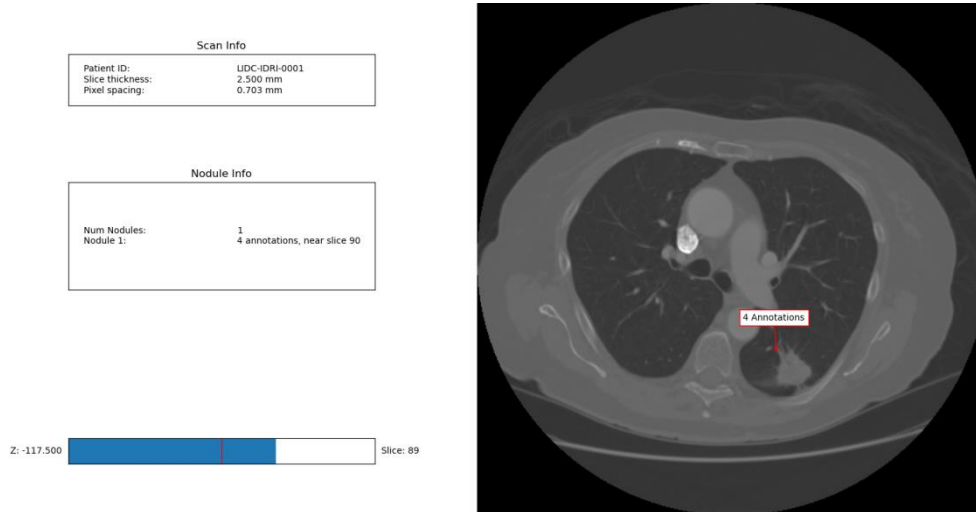


Fig.3.22. Scan Information and its total number of annotations.

To get all annotations corresponding to the same nodule, a for loop is used to iterate through the annotations and return the pathologic features associated with it as a new appended row in the CSV file which is used in the neural network classification step.

Define annotations associated with a scan, based on its patient ID or any other feature needed to be used in the research, to show nodule's contours in metric values (Figure (3.23)), along with physical information about the annotations like the diameter, centroid (Figure (3.24)), and the shape.

```
anns = pl.query(pl.Annotation).join(pl.Scan).filter(pl.Scan.patient_id == pid)
```

```
contours = ann.contours
```

```
dia, surf_area, vol = ann.diameter, ann.surface_area, ann.volume
```

The code above draws a red line around the nodule and computes its diameter in milli-meters (mm), surface area in mm², and the nodule’s volume in mm³.

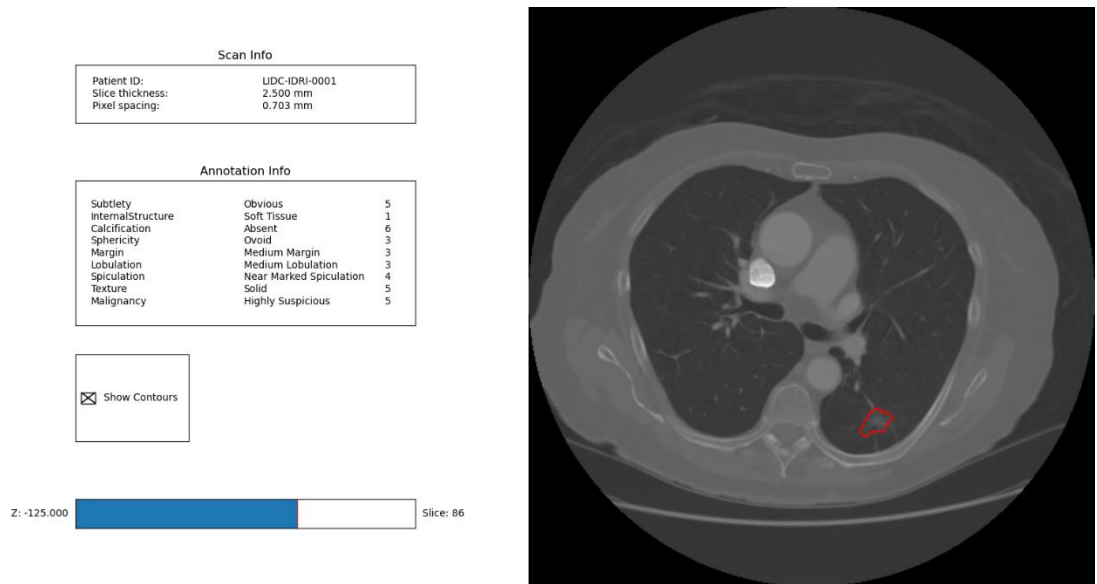


Fig.3.23. Nodule’s Contour.

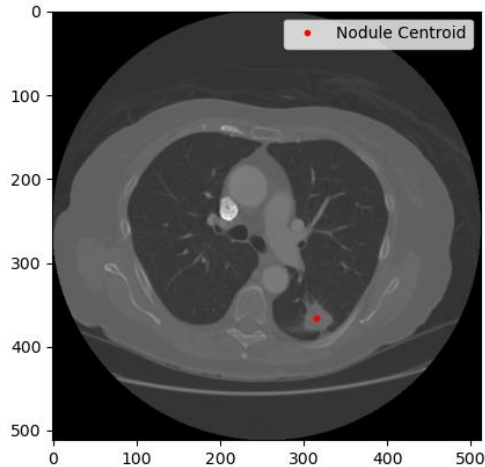


Fig.3.24. Nodule’s Centroid.

A Boolean mask for the nodule is created by the command

```
mask = ann.boolean_mask()
```

This mask doesn’t occupy the entire extent of the CT image volume (which is 512pixelsX512 pixelsX number of slices). It sits within the

computed bounding box of the nodule [52], as shown in Figure (3.25). The Bounding Box is the computed extent of the contour indices of the annotation. It returns a tuple of slices for the edge boundaries with the largest diameter for every coordination axis (X-axis, Y-axis, and Z-axis, respectively) surrounding the nodule, as shown in the following code:

```
bbox = ann.bbox() // compute the bounding box of annotation

print(bbox) // the location of a nodule inside an image

# => (slice (340, 390, None), slice (297, 338, None), slice (86, 94, None))

//nodule dimensions: {X=340-390},{Y= 297-338}, {Z=86-94}.
```

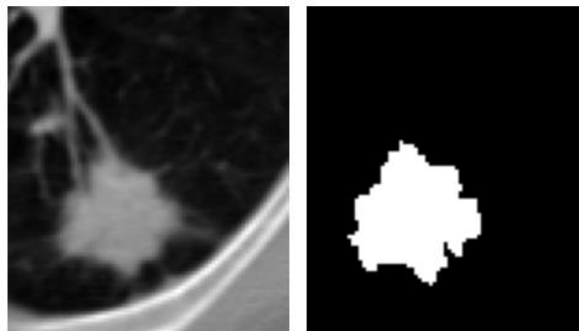


Fig.3.25. Boolean mask of the nodule.

The final output CSV file is of size $4000 \text{ annotations} \times 6 \text{ features}$ and it is used as the input for the classification step to detect whether a nodule is cancerous or not.

These ROI are also soterd as row vectors in a matrix of size 352×62501 to use it as the training dataset for CNN model.

3.4.2. Classification Step using Neural Networks:

In this section, the features extracted are used as the input to the neural network model to determine if a nodule is cancerous or not.

The input features are subtlety, internal structure, calcification, sphericity, margin, and the output class malignancy. The total input size of the data used to classify nodules is $4220 \text{ annotations} \times 6 \text{ features}$, and it is divided with 80% training data set of actual size of $3376 \text{ annotations} \times 6 \text{ features}$, and the remaining 20% of actual size equals to $844 \text{ annotations} \times 6 \text{ features}$ is defined as the testing set, where the first 5 columns are the input features and the 6th column in the dataset is the output class.

Multiple classification methods are used to determine whether a nodule has cancer tumor or not. These are Logistic Regression, Linear Discriminant Analysis, Classification and Regression Trees, Naïve Bayes, K-Nearest Neighbors classifier, and Support Vector Machine. Also, a comparison between these algorithms is made to evaluate the most suitable method for the proposed model.

Logistic Regression (LR) is a supervised learning classification linear algorithm based on the statistical method logistic function (sigmoid function). It is best used with binary classification problems. Input values(X) are combined in the dataset linearly to predict the output value (y) [35].

Linear Discriminant Analysis (LDA) [36] is a dimensionality reduction technique commonly used for supervised classification problems to separate two or more classes. It reduces dimensions of 2D graph into 1D graph to maximize the separability between the two classes.

Classification and Regression Trees (CART) is a combination of two types of decision trees: classification trees and regression trees. In classification trees, the output variable is categorical and the tree is used to identify the class to which the output value falls into. In regression trees, the output variable is continuous and the tree is used to predict its value. It is structured in a way to split the data at each tree node based on the variable value, decide whether a branch is terminal or not by making stopping rules, and predict the output variable in each terminal node [37].

Gaussian Naïve Bayes (NB) is classification technique based on the Bayes' Theorem with the assumption that the features (predictors) are independent from each other. Gaussian distribution function is used with NB in case the input data was continuous to avoid output zero for all input data [38].

K-Nearest Neighbors classifier (KNN) [10] uses data and classifies new data points based on similarity measures between variables, and the classification process of an input variable occurs by majority of similar features of that input with its neighbors. The greater the number of nearest neighbors, the accuracy value increases.

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm that outputs a hyperplane that divides the dataset into two classes (in this model) or more classes based on the categorical output variable of the dataset.

These algorithms are used with the features extracted CSV file and trained with each classifier to give the accuracy of cancer prediction:

```
models.append(('LR', LogisticRegression()))

models.append(('LDA', LinearDiscriminantAnalysis()))

models.append(('KNN', KNeighborsClassifier()))

models.append(('CART', DecisionTreeClassifier()))

models.append(('NB', GaussianNB()))

models.append(('SVM', SVC()))
```

Accuracy is printed out for all algorithms, then confusion matrix, and classification report were printed out for the classifier that returned the highest accuracy value which was tested on 20% of the data with the size *844 annotations × 6 features*.

```
print ('Accuracy = ', accuracy_score(Y_test, predictions) *100)

print(confusion_matrix(Y_test, predictions))

print(classification_report(Y_test, predictions))
```

Another classification method used is the Neural Network classification (NN). Neural networks are set of algorithms designed to

recognize patterns and help cluster classify inputs to the given categorical output classes.

The model used the libraries of *TensorFlow*(an end-to-end free-open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that helps building and deploying machine learning programs easily) and *Keras*(a deep learning API written in Python, running on top of the machine learning platform TensorFlow defined as '*tf.keras*').

The model is sequential; meaning its layers are organized as a linear stack of layers, which is the simplest type of models in neural networks.

The model consists of an input layer, five hidden layers, and an output layer. The first hidden layer contains 128 neurons, the second has 64 neurons, the third has 32 neurons, the fourth layer has 16 neurons, and the fifth layer has 8 neurons. The final output layer has one neuron to classify the nodule in binary output, either 1 for cancer, or 0 for not-cancer nodule with a Sigmoid activation function.

The activation function used in the hidden layers is Rectified Linear Unit function (ReLU), and with the final output layer a Sigmoid activation function is used. The Batch size is set to 10 and number of epochs is 350. The prediction set is taken randomly from the same input dataset.

Activation functions are mathematical equations that determine the output of a neural network. An activation function is attached to each neuron in the network and determines whether this neuron should be activated or not

(outputs a 0-value) based on that neuron's input. Also, it performs a normalization operation on each neuron's output.

Multiple types of activation functions are used in this model with different classification algorithms, such as Rectified Linear Unit function, Sigmoid function, and SoftMax function.

Rectified Linear Unit function (ReLU) is a type of activation functions. It's linear for all positive values and zero for all negative values.

$$y = \max(0, X) \quad (3.9)$$

It's used to eliminate unwanted features from dataset keeping the features mostly can help classify the input variables to one of the output classes.

Sigmoid function (also called logistic function) is another type of activation function (Fig(3.26)), used to predict the probability of an output in neural networks [41]. It maps the whole real domain of t into $[0, 1]$.

$$\sigma(t) = \frac{1}{1+e^{-t}} \quad (3.10)$$

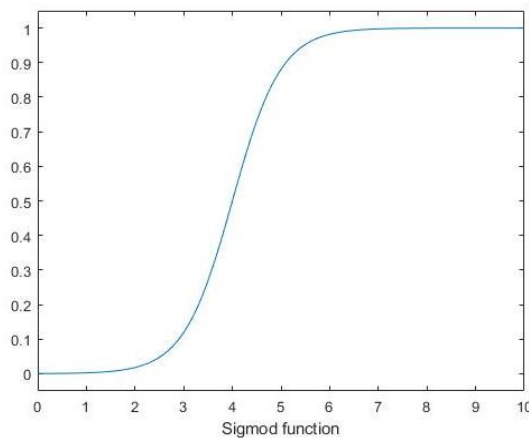


Fig.3.26. Boolean mask of the nodule.

SoftMax function is an output activation function that estimates class probabilities in multi-layer class. It is used in the last layer of neural network model

$$S(y(i)) = \frac{e^{y(i)}}{\sum_j e^{y(j)}} \quad (3.11)$$

It represents the categorical probability distributions of potential outcomes.

Each of these activation functions were used and tested in our model to see which distributes output prediction probabilities better.

The loss in data is computed by Binary Cross-Entropy (Log Loss) function. The loss function is used to return probability values for predictions. If predictions are good, then it should return low values. If the predictions are bad, the probability values returned are high.

To compute the learning rate for all weights on each neuron in the network; Adam optimizer was used. It updates the network weights iteratively based on training set.

The input dataset, as mentioned earlier, consists of 4220 input rows. This size of data is relatively small when trained using neural networks and this issue can cause the training data to overfit. So, to avoid overfitting a Dropout layer is added after the input layer and every hidden layer.

Dropout is a regularization method that approximates training a large number of neural networks with different architectures in parallel.

Dropout is implemented per-layer in neural network models. It can be implemented on the input layer and on any or all the hidden layers in a network, but not used on the output layer.

Dropout layer takes one parameter as an input. This parameter specifies the probability at which outputs of the layer are ignored (dropped out). The value used is 0.5 to drop out 50% of the outputs from a layer and keep 50% of these outputs to be used for the next layer in the model.

3.4.3. Classification Step using Convolutional Neural Networks on HU matrix:

The third step of classification stage is to test the output HU matrix, obtained from applying image processing techniques, using a CNN model trained with the dataset used to train the Neural Network model explained in section 3.4.2.

The testing dataset, the HU matrix, is stored as a matrix of size 33542X62501. The 2-D ROI image extracted has size of 250X250 pixels converted to row vector of size 62500, plus the output labeled class.

This matrix is used as input dataset to a Convolutional Neural Network (CNN) for predicting output classes based on the results from the training process.

Convolutional Neural Network (CNN) is emerged from the study of the brain's visual cortex. They have been used in image recognition. CNN is a sequence of layers where each layer passes one volume of activations

to another layer through differentiable functions. It is made of neurons that have learnable weights.

Convolutional Network consists of three types of layers:

- **Convolutional layer:** is the most important building block of CNN. Neurons in the first convolutional layer are not connected to every single pixel in the input image, but only to pixels in their receptive fields. Each neuron in the second layer is connected only to neurons located within a small rectangle in the first layer concentrating on low-level features in the first hidden layer, then assemble them into higher-level features in the next hidden layer, see Figure (3.27). Each neuron receives inputs, perform dot product and optionally follows it with a non-linearity function.
- **Pooling layer:** reduces the dimensions of the data input by combining the outputs of neurons at one layer into a single neuron in the next layer. There are many types of pooling: Local pooling combines small clusters of neurons, such as 2X2, Global pooling performs reduction on all the neurons of the convolutional layer, Max pooling which uses the maximum value from each group of neurons at the previous layer, and the Average pooling uses the average value of each neurons' group at the previous layer.
- **Fully Connected layer:** connects every neuron in one layer to every neuron in another layer, the matrix goes through this layer to classify the images.

In CNN the layers have neurons arranged in 3D: width, height, and depth (number of channels).

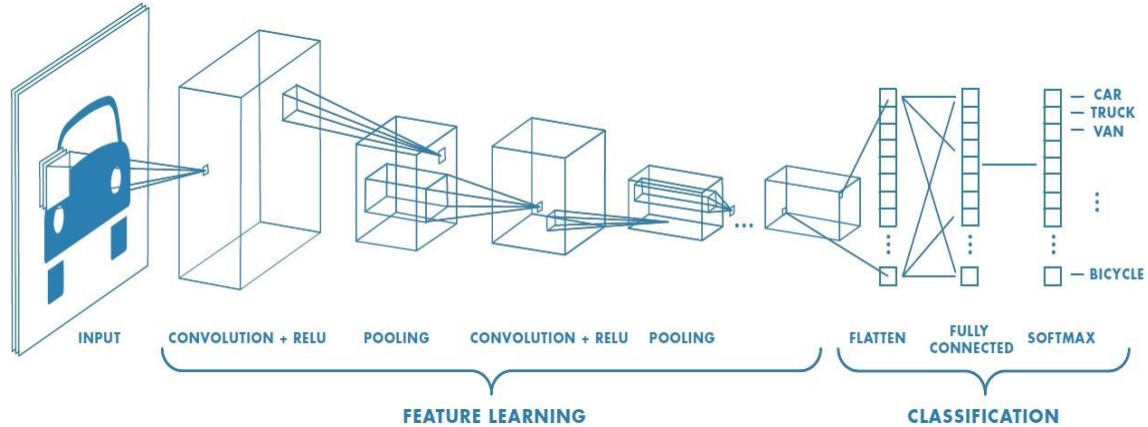


Fig.3.27. CNN architecture made of several 3D layers. [40]

Before inserting the images to CNN, each row vector is reshaped to an image of size $250 * 250$ pixels.

These images are divided into 70% training set, 20% testing set, and 10% validation set.

The training data is of size 1228 samples, the testing data is 352, and the validation data is 176.

The CNN model is sequential consisting of:

- Four convolutional layers with Rectified Linear Unit functions for activation
- Four Max pooling layers reducing the size of each input image by 2, and reduce the input size by 3 in the last Max pooling layer.
- A flatten layer to reshape each 2D image into 1D vector.
- A fully connected layer to classify the image with Sigmoid activation function

The CNN model summary is shown below:

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 248, 248, 32)	320
max_pooling2d_1 (MaxPooling2)	(None, 124, 124, 32)	0
conv2d_2 (Conv2D)	(None, 122, 122, 64)	18496
max_pooling2d_2 (MaxPooling2)	(None, 61, 61, 64)	0
conv2d_3 (Conv2D)	(None, 59, 59, 128)	73856
max_pooling2d_3 (MaxPooling2)	(None, 29, 29, 128)	0
conv2d_4 (Conv2D)	(None, 27, 27, 256)	295168
max_pooling2d_4 (MaxPooling2)	(None, 9, 9, 256)	0
flatten_1 (Flatten)	(None, 20736)	0
dense_1 (Dense)	(None, 256)	5308672
dense_2 (Dense)	(None, 1)	257
Total params: 5,696,769		
Trainable params: 5,696,769		
Non-trainable params: 0		

The loss function used is the binary-cross entropy. It computes the loss in data-input recognition. The optimizer used is Adam's. It returns prediction's accuracy value of the model.

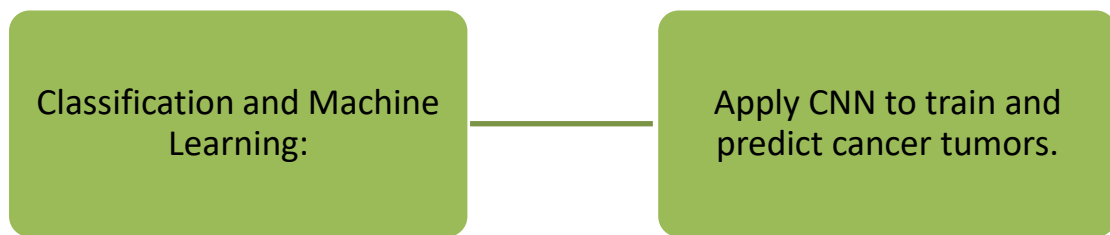
The testing dataset used to test the accuracy of prediction on new input data. The learnable weights are saved in an h5 file along with the model itself.

The confusion matrix and classification report are printed out viewing the statistical results for the proposed model.

After the model is trained, the dataset HU matrix, obtained from applying Image processing techniques is tested to classify whether the nodules extracted are cancerous or not. The tested dataset is of size 33254X62501.

The dataset set is tested and the accuracy and loss are compared to the results obtained from training the model.

The following diagram shows a summary of the steps in this section.



The following pseudo code shows the classification steps applied to determine if a ROI is classified as cancer or not-cancer:

Find annotations from the labeled dataset

Build ANN model to predict tumors

ANN model

Input layer

Five hidden layers

Dense layer

Output layer

Add label to each annotations

```
if label == 1
    then it is cancer
else
    then it is not-cancer
```

Use the trained dataset from ANN model to extract ROI as images

Save the ROIs in csv file 'ann_ROI.csv'

Build CNN model

```
train the model with 'ann_ROI.csv' dataset
open extracted HU_matrix from image processing stage
evaluate CNN model and predict output class using HU_matrix
if predicted_label == label
    then model succeeds
else
    model fails
```


Chapter Four

Results

Chapter Four

Results

The method introduced in this thesis consists of four-stages system:

- Preprocessing stage.
- Image Segmentation stage.
- Features Extraction stage.
- Classification stage.

The methodology is applied to multiple samples' patients. Those samples are obtained from The Lung Image Database Consortium image collection (LIDC-IDRI) [42]. The dataset consists of 48 folders with a total of 8395 images and 33542 ROIs in these images.

This dataset consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. It is a web-accessible international resource for development, training, and evaluation of computer-assisted diagnostic (CAD) methods for lung cancer detection and diagnosis. Initiated by the National Cancer Institute (NCI), further advanced by the Foundation for the National Institutes of Health (FNIH), and accompanied by the Food and Drug Administration (FDA) through active participation, this public-private partnership demonstrates the success of a consortium founded on a consensus-based process [1].

Applying the algorithms explained in the methodology; the 64-bit MATLAB R2018a application was used as a high-level language and

interactive environment that supports image processing techniques with a special toolbox contains many built-in functions that returned best results for this research.

The application installed on a 64-bit Windows10 Pro, with 3.30 GHz Intel Core i5 CPU Processor, 16GB RAM, and NVIDIA GeForce GTX 1050Ti Display adapter.

Parallel processing was used for better performance especially the loops. MATLAB parallel pool activated to reduce code run-time on some parts of the code, and other parts were run on the NVIDIA display adapter which supports parallel processing.

For time performance and handling the large dataset, An-Najah Computer Science Lab's Cluster was used for image analysis and segmentation stages.

4.1. Preprocessing stage:

The first stage in the program is preprocessing. DICOM image was read by MATLAB functions, the image intensity values are stored in 16-bit depth by default, so we need to convert it back into 12-bit in order to read accurate HU intensity values as the original acquired images from CT medical scanning devices. A linear transformation is applied by equation (3.2). The input image before and after transforming the intensity values are shown in Figure (4.1):

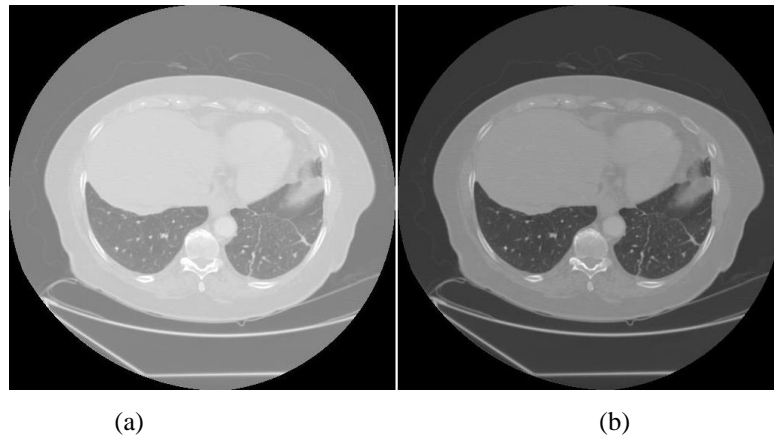


Fig.4.1. (a) DICOM image. (b) HU Image.

4.2. Image Segmentation stage:

The output of the first stage is the input for the segmentation stage. In this stage, Gabor filter was used to remove any noises from the images during acquisition. This filter returns the magnitude and phase of the input image, and to achieve high accuracy and eliminate noise and unwanted features; a set of 28 Gabor filters were applied and the resulting image was reshaped by using PCA on pixels' values mean and standard deviation, as shown in Figure (4.2) :



Fig.4.2. Result after applying 26-Gabor filters.

The edge detection step is applied to the processed image to help define the ROI boundaries. Many techniques are famous for edge detection

such as Prewitt, Roberts, Sobel, and Canny. The difference between each method is the accuracy of determining ROI boundaries and by experimenting with all these methods; Canny gave the desired output for this research. The outputs for applying each of these methods are shown in Figure (4.3):

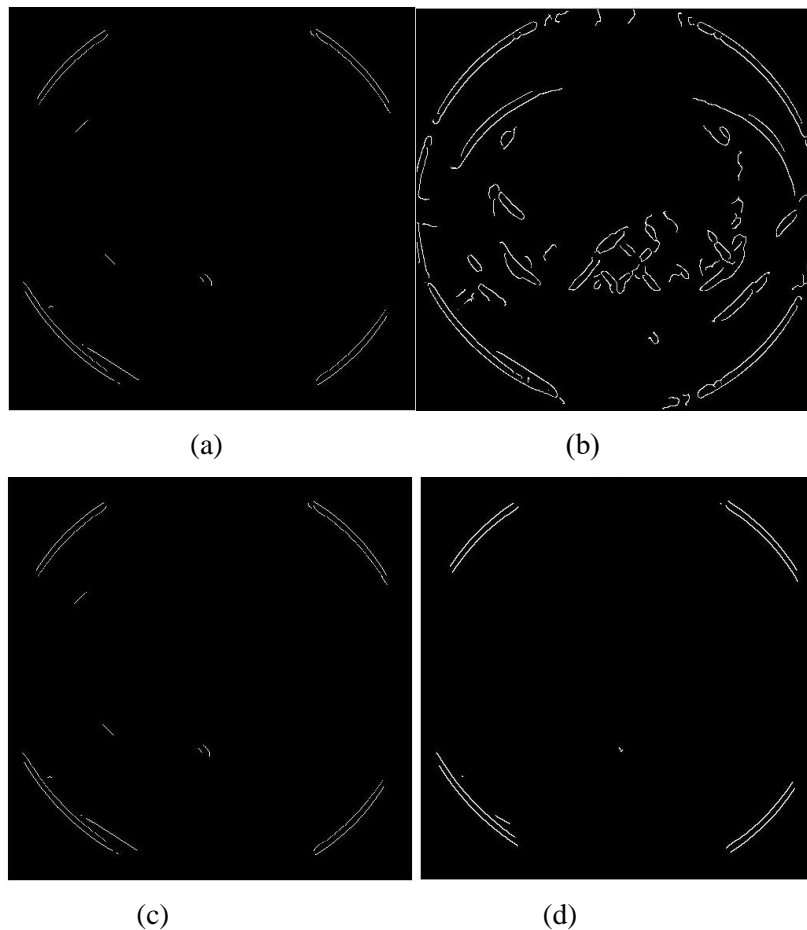


Fig.4.3. Edge Detection Methods: (a) Canny, (b) Prewitt, (c) Roberts, and (d) Sobel.

After detecting the edges, pixels' values are binarized to choose a suitable threshold for background and foreground separation.

When computing the Histogram for the original input image as shown in Figure (4.4), it is noticed that there is an obvious peak between

pixels' intensities near 0. This peak indicates that there are some pixels in the image with abnormal intensity values. This abnormality will help defining a suitable threshold value to separate image regions in order to allocate ROI.

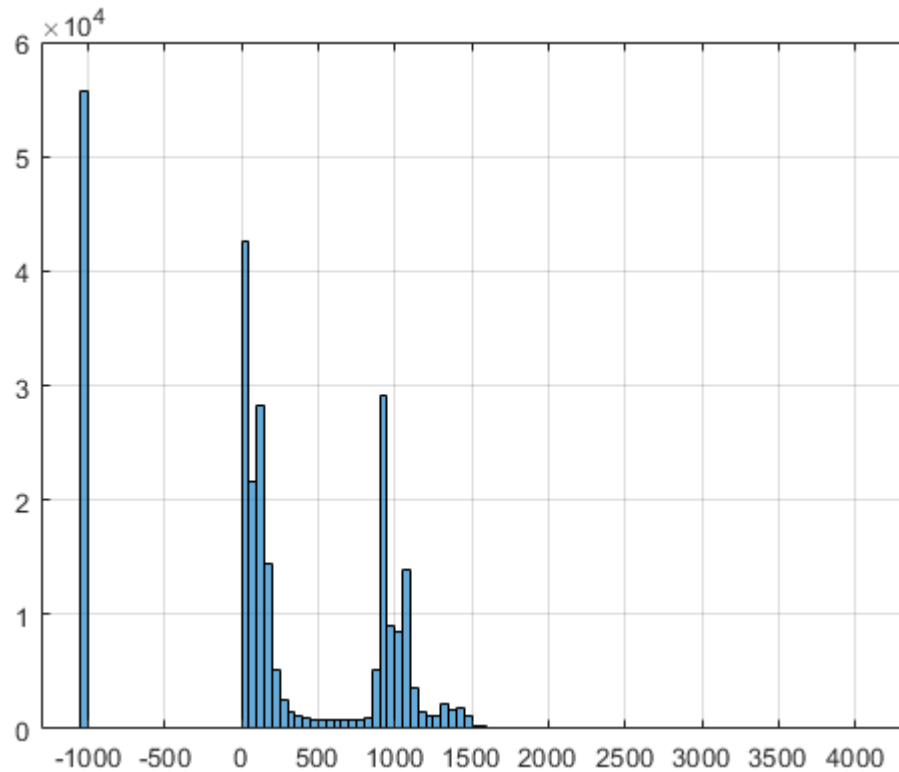


Fig.4.4. Image Histogram.

This step is accomplished by Binarization Method. The most common methods used in image processing to evaluate a good threshold for image details separation into background and foreground include Otsu's Method, Global Thresholding, and Local (or Adaptive) Thresholding. The foreground is the ROI in this research.

By experimenting and comparing the outputs from each method, a GUI MATLAB from Central File Exchange threshold tool gave the best output as shown in Figure (3.8). However, this method requires the user to

choose the proper threshold manually which causes a problem because thousands of images to have to be processed. Therefore, an automatic method must be used to select the proper threshold. After several trials with different automatic methods, Otsu's Method gave the best acceptable threshold values compared to other methods that compute the threshold automatically.

The outputs of the mentioned binarization methods are shown in Figure (4.5), and the optimal averaged threshold value was approximately '0.4947395'.

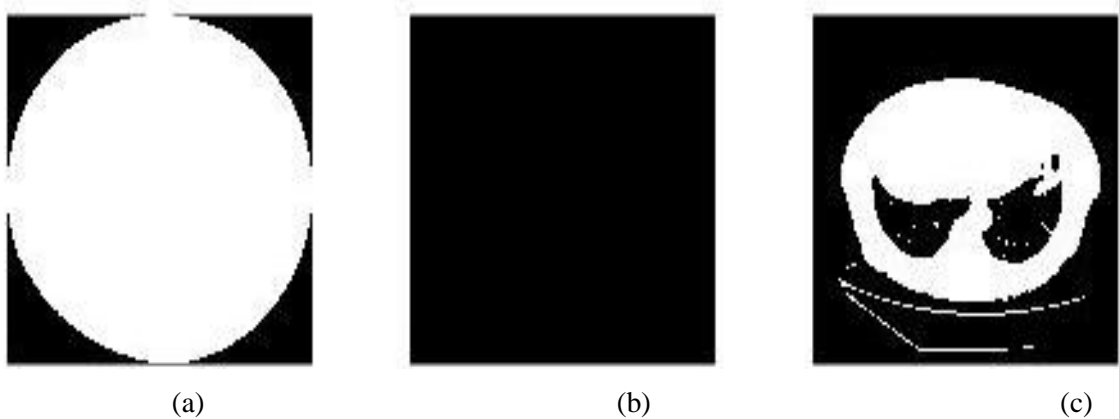


Fig.4.5. Binarization using: (a) Global, (b) Adaptive, and (c) Otsu's Thresholding.

After that, the lungs were extracted from the chest cavity by creating a mask and taking the image complement, using Morphological Operations, and clearing borders. Morphological operations began with dilation using a disk-shaped structural element (SE) of radius equals 4, followed by erosion with a one radius SE. Clear the borders and fill the area inside the lungs. Another erosion step is applied and the final mask is completed (Figure (4.6)). This mask is multiplied with the original input image to obtain the

lungs' intensity values and extract ROI. The output is shown in Figure (4.7).

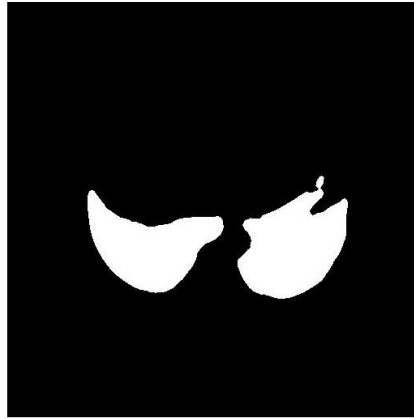


Fig.4.6. Lung Mask.



Fig.4.7. Lungs from multiplying Lung mask with the original input image.

Another set of Morphological operations is applied to eliminate all unwanted parts inside the lungs except for ROI to be tested if being a tumor or not as shown in Figure (4.8).



Fig.4.8. Applying Morphological operations to create a mask and multiply it with original input image.

The operations applied are opening operation followed by erosion and reconstruction steps, then a close operation is applied followed by dilation operation and another reconstruct operation (Figure (4.8)), respectively. The final output after applying the morphological operations is shown in Figure (4.9). The complement is computed (Figure (4.10)) and the regions with maximum pixel values are extracted as shown in Figure (4.11).



Fig.4.9. Final output image after Morphological Operations.

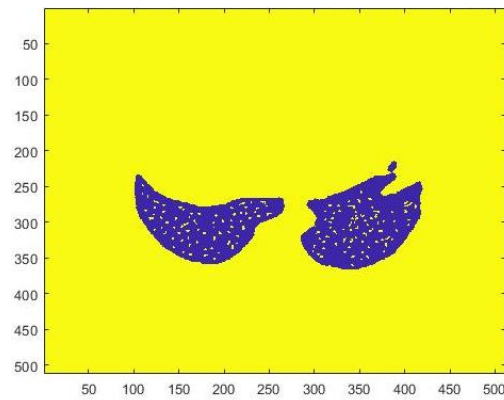


Fig.4.10. Regional maxima extraction

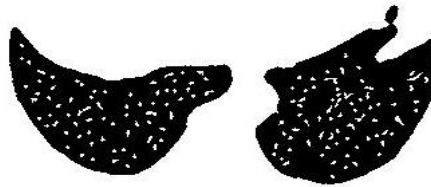


Fig.4.11. Regional Maxima superimposed on the original image.

With those steps applied, any object with a size greater than 5 pixels will be eliminated from the image to extract the candidate ROI.

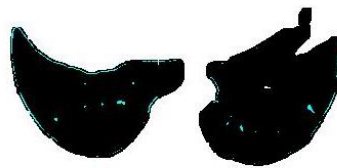


Fig.4.12. Clear Borders and eliminate unwanted ROIs.

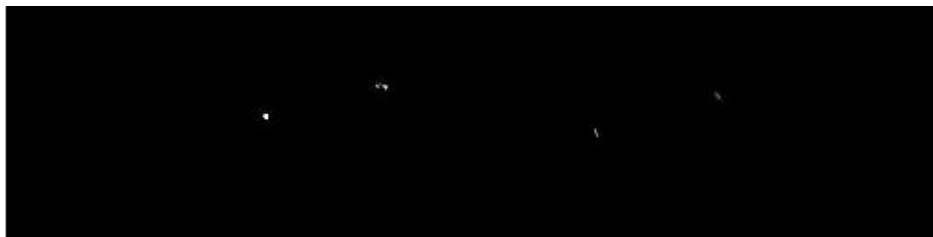


Fig.4.13. final ROI in the binary image.

The final ROI is extracted from the binary image by taking the image complement to remove the borders from Figure (4.12) and keep regions of interest inside the lungs as shown in Figure (4.13).

Watershed-Controller Technique helped to separate each region extracted from the previous step and treat it as a single candidate for examination whether it indicates a tumor or not. So, computing the watershed ridgelines and filling each region with different color (Figure (4.14.a)), then imposing the colored labels on the original input image as shown in Figure (4.14.b) returns each ROI labeled with different color as shown Figure (4.15).

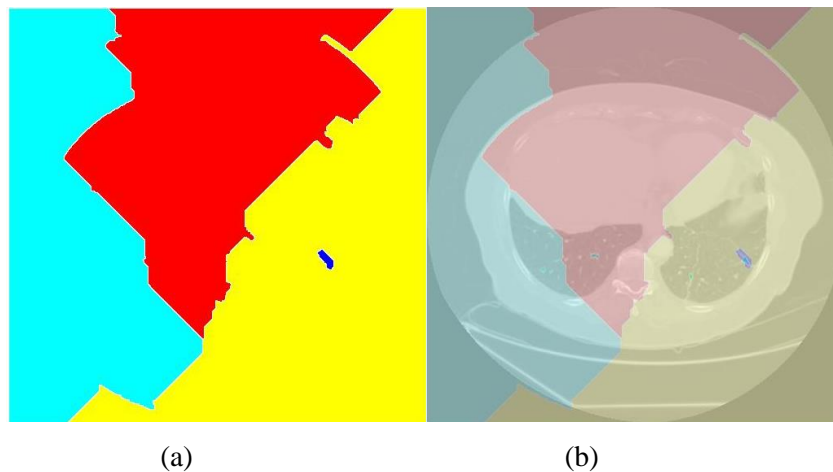


Fig.4.14. (a) Watershed labeled regions, (b) Labels imposed on the original input image.



Fig.4.15. Final ROI labeled.

4.3. Features Extraction stage:

The third stage is feature extraction. In this stage several features of each ROI are extracted. In this thesis, we considered three different approaches for features extraction and classification.

- 1) In the first approach we extracted geometrical properties of each region and used these properties as features to train the model. These properties include the area, centroid, diameter, and others as mentioned in Chapter Three.

A built-in MATLAB function returns the properties and pixels' values as a MAT file. For ease of comparison and to use these features as the input for the machine learning stage, this MAT file is converted to a table.

- 2) In this approach we replaced the values of the pixels in each ROI with its corresponding value from HU matrix. Then an image of each ROI is extracted.

The images are computed by multiplying the binary image that contains the ROIs with the HU matrix. Then these images are flattened and combined to form a matrix of row vectors.

- 3) This approach is similar to the second approach, but we used the values of the pixels in the ROI from those in the original DICOM image.

The images are obtained by multiplying the extracted ROI with the DICOM matrix returning the intensity values of image pixels.

To determine the suitable size of the ROI images in the second and the third approaches, several sizes were examined based on the biggest ROI detected. The final size was 250X250 pixels. The size was reduced from 512X512 pixels for performance issues on the machine used to run the program. Each ROI was stored in the matrix as a row vector of the same size.

The accuracy of detecting ROI with cancer compared to the labeled dataset [1] is 100% for the nodules inside the lungs. For a random sample of folders taken from the trained dataset with nodules identified inside and near the boundaries, the accuracy of detecting suspicious area of tumor near lungs' boundaries was 72.92%.

4.4. Classification and machine learning stage:

The final stage was the classification. In this stage nodules are classified whether they are cancer or not.

The first step of this stage is reading the annotations and extract the desired attributes (pathologic Features), and the second step is the classification of these features using six different classification algorithms in an approach and neural network algorithm in a second approach to determine whether a nodule is cancerous or not. The third step is using the ROIs extracted in the image processing stages as input to predict if the nodules are cancer or not using the Convolutional Neural Network model trained in the second step.

4.4.1. Reading annotations and extract features:

The dataset used in this step was the DICOM header files from the LIDC-IDRI dataset and the annotations were read from XML file associated with the dataset [1].

The size of the input dataset is 1012 patients' scans with annotations' information on the nodules reviewed independently by four radiologists.

The LIDC radiologists' annotations include outlines of nodules ≥ 3 mm in diameter on each CT slice in which the nodules are visible, along with the subjective ratings scale of the pathologic features mentioned in Chapter 3.

The features extracted to be used for classification as inputs were: subtlety, internal structure, calcification, sphericity, and margin, while malignancy was used as the output class. Its categories were reduced from 5 outputs to binary output with 1 for cancer and 0 for not-cancer nodules.

The root folder is defined and all the scans within it were processed to extract the corresponding annotations and the associated pathologic features with each annotation.

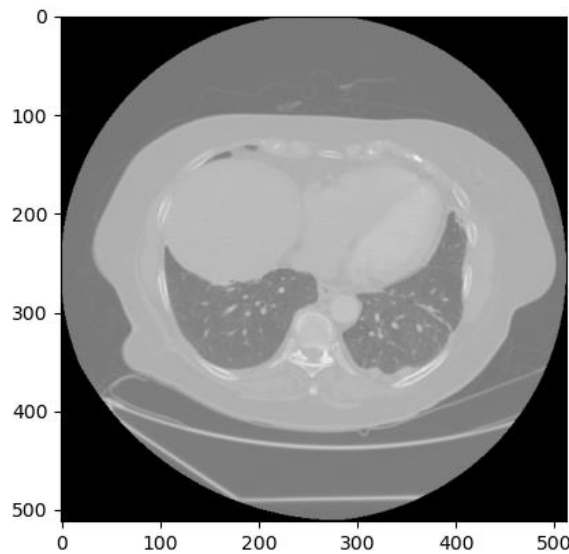


Fig.4.16. Original DICOM image.

The original DICOM image from scan LIDC-IDRI-0004 is shown in Figure (4.16). The scan information such as patient ID, size, slice thickness, nodule centroid as shown in Figure (4.17), nodule's diameter, and other information can be read and printed out from the DICOM header files and XML file.

Patient ID: LIDC-IDRI-0004

Study Instance UID:

1.3.6.1.4.1.14519.5.2.1.6279.6001.191425307197546732281885591780

Slice Thickness: 1.25

Shape: (512, 512, 241)

Diameter: 8.46574661350882

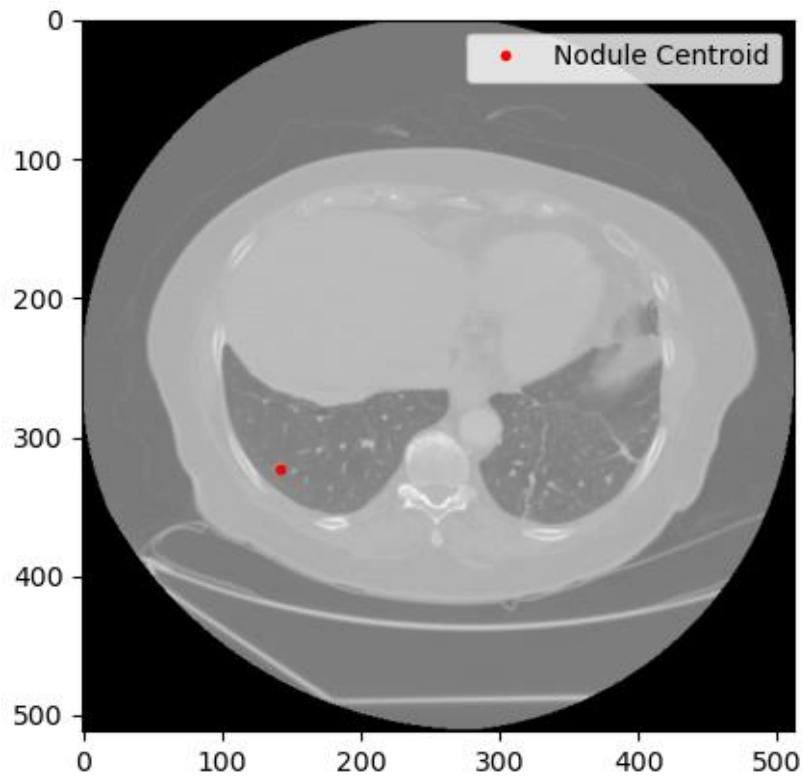


Fig.4.17. Nodule Centroid marked in red.

The number of nodules in the scan and total number of annotations for each nodule can be determined and printed out.

Scan(id=15, patient_id=LIDC-IDRI-0004) has 1 nodules.

Maximum number of annotations: 4

The nodule is defined by contours and the number of slices in which the nodule appears are determined starting with the mean slice for nodule appearance in the scan as Figure (4.18) demonstrates.

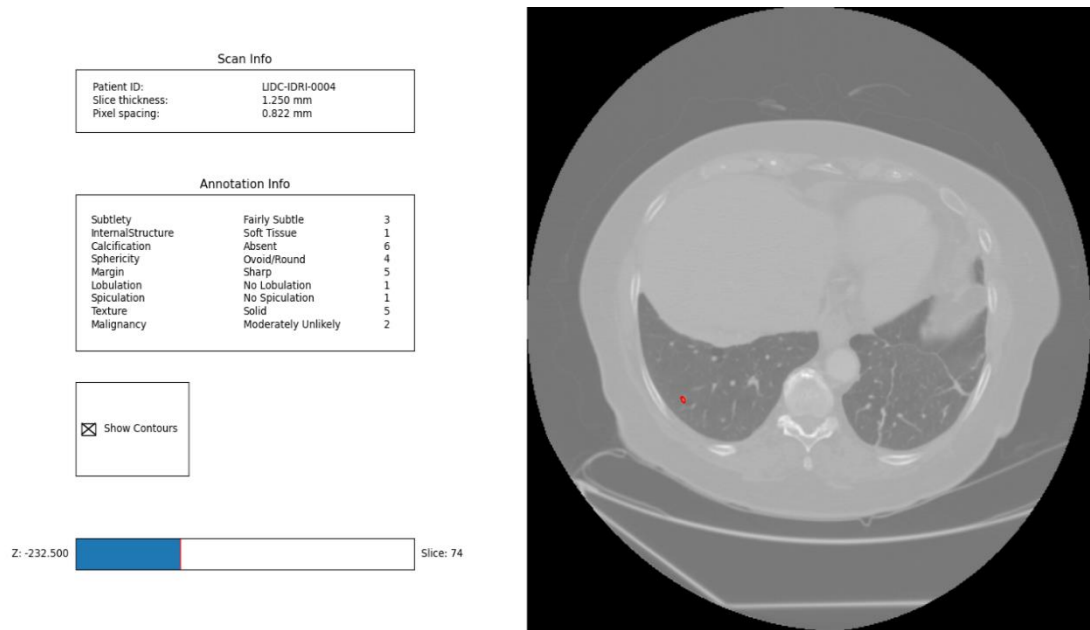


Fig.4.18. Scan information and nodule contours shown in red.

Figure (4.18) showed the DICOM image with red contour around the detected nodule defined by the radiologists. Also, this figure gave information about the scan, the annotation features, and the blue dash returned the number of slices where the nodule began to appear in the scan, in this case the slice number where the nodule in scan LIDC-IDRI-0004 was detected equaled to 74.

The returned number of annotations for a nodule is used to loop through these annotations to extract the needed features for the research. The features can be extracted separately by calling each feature by its name as shown in Table (4.1)

Table.4.1. The 6 features extracted for scan LIDC-IDRI-0004.

Pathologic Features	Annotation1	Annotation2	Annotation3	Annotation4
Subtlety	3	2	2	5
Internal Structure	1	1	1	1
Calcification	6	3	3	3
Sphericity	4	3	2	5
Margin	5	5	5	5
Malignancy	2	1	1	1

Or all the nine features mentioned in Chapter 3 can be extracted at once as a row vector for each annotation as shown in Table (4.2)

Table.4.2. All the features extracted at once as row vector.

Features Annotations	#1	#2	#3	#4	#5	#6	#7	#8	#9
Annotation1	3	1	6	4	5	1	1	5	2
Annotation2	2	1	3	3	5	1	1	5	1
Annotation3	2	1	3	2	5	1	1	5	1
Annotation4	5	1	3	5	5	1	1	5	1

Geometrical functions from the annotation class defined in *pylidy* library were applied to specify each annotations diameter, volume, and surface area measured in millimeters.

Diameter: 6.26 mm, surface-area: 76.41 mm², volume: 67.19 mm³

After extracting the annotations and defining the nodule of interest, a Boolean mask was taken for the nodule with image size less than the original image size (512×512 pixels) to reduce memory usage in further analysis of these images. This mask was multiplied with the original image to return the nodule in its original shape as shown in Figure (4.19). with the Boolean mask, a bounding box information is used to determine the coordinates of the nodule in 3D dimensions. These coordinates are defined as tuples of boundaries on each axis, X-axis, Y-axis, and Z-axis, respectively.

(slice (319, 328, None), slice (139, 146, None), slice (74, 78, None))

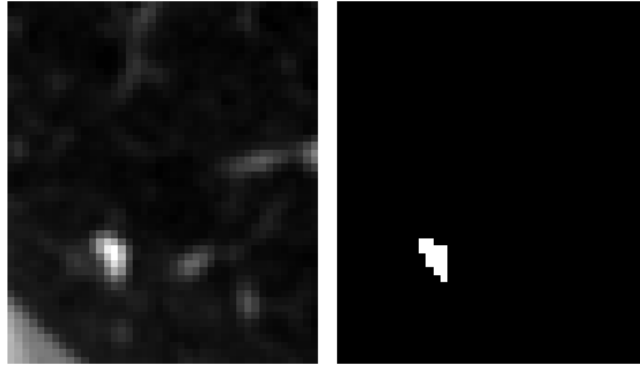


Fig.4.19. Boolean mask on right, original nodule on left.

The output of this step is a CSV file containing the extracted features from each annotation in all scans found in the root folder. The size of this file is $6770 \text{ annotations} \times 6 \text{ features}$, Where the rows are the total number of annotations defined by the radiologists, and the columns are the pathologic features extracted to be used in the classification step and they are: subtlety (#1), internal structure (#2), calcification (#3), sphericity (#4), margin (#5), and the output class malignancy (#6) as named in Table (4.3).

This size is reduced after eliminating the annotations categorized with malignancy equals to 3, because it stands for indeterminate whether the nodule is cancer or not. The size of the CSV file became 4220×6 . The malignancy feature is converted from 4 categories into binary class to classify if a nodule is cancerous (takes value 1) or non-cancerous (takes value 0). To achieve this conversion, the malignancy categories' values of 1 and 2 are defined as unlikely being cancer so these values were converted to the binary value Zero, while malignancy categories 4 and 5 are defined to be suspiciously being cancer so they were converted to the binary value One. Table (4.3) shows a part of the csv output file.

Table.4.3. Sample of the output CSV file.

#1	#2	#3	#4	#5	#6
4	1	6	3	4	1
4	1	6	4	4	1
3	1	6	5	5	0
4	1	6	4	5	1
5	1	6	5	5	1
3	1	6	5	5	0
3	1	6	4	5	0

4.4.2. Classification step:

The classification step took the features CSV file as the input to determine whether a nodule is cancerous or not. The input dataset consisted of five input parameters and a single binary output class.

Many classification algorithms were used to train and test the input dataset. The training process was applied on 80% of the dataset and the

testing process was applied on the remaining 20% of the set. A sample is shown in Table (4.4).

Table.4.4. A sample of the input dataset.

#	Subtlety	InternalStructure	Calcification	Sphericity	Margin	Malignancy
2	5	1	6	3	2	1
3	5	1	6	5	4	1
4	2	1	6	5	1	1
5	1	1	6	3	2	1
6	1	1	6	5	2	0

The box and the whisker plots were printed to view the distribution of the variables of the dataset as shown in Figure (4.20). the histograms for the same variables shown in Figure (4.21).

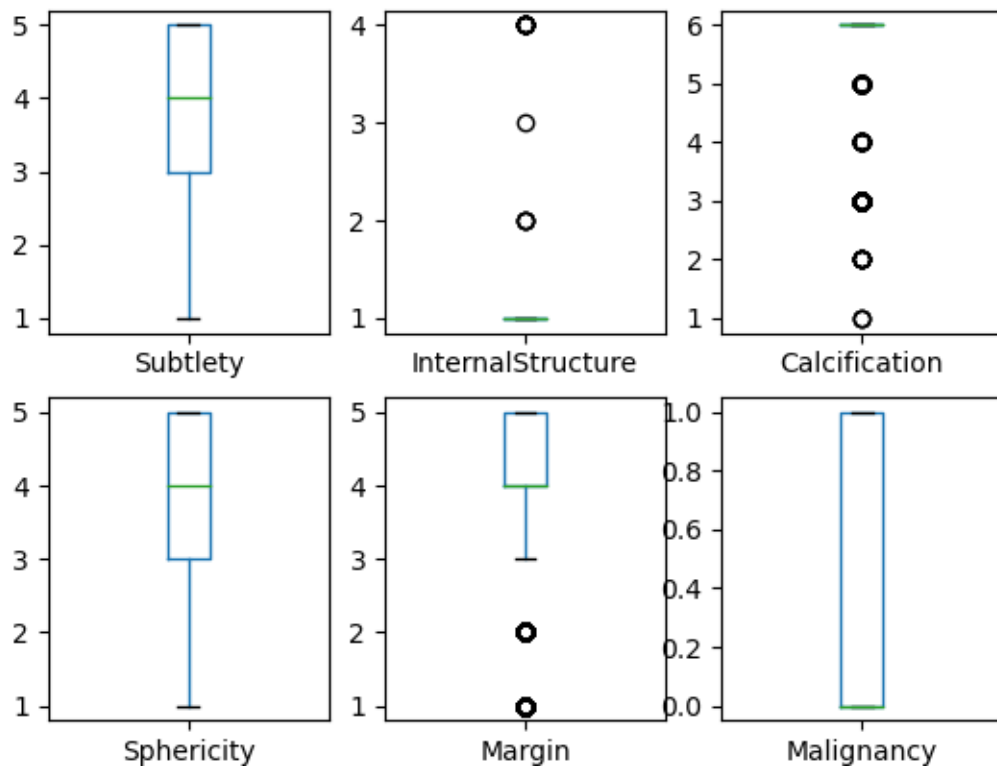


Fig.4.20. The box and whisker plots of the dataset's variables.

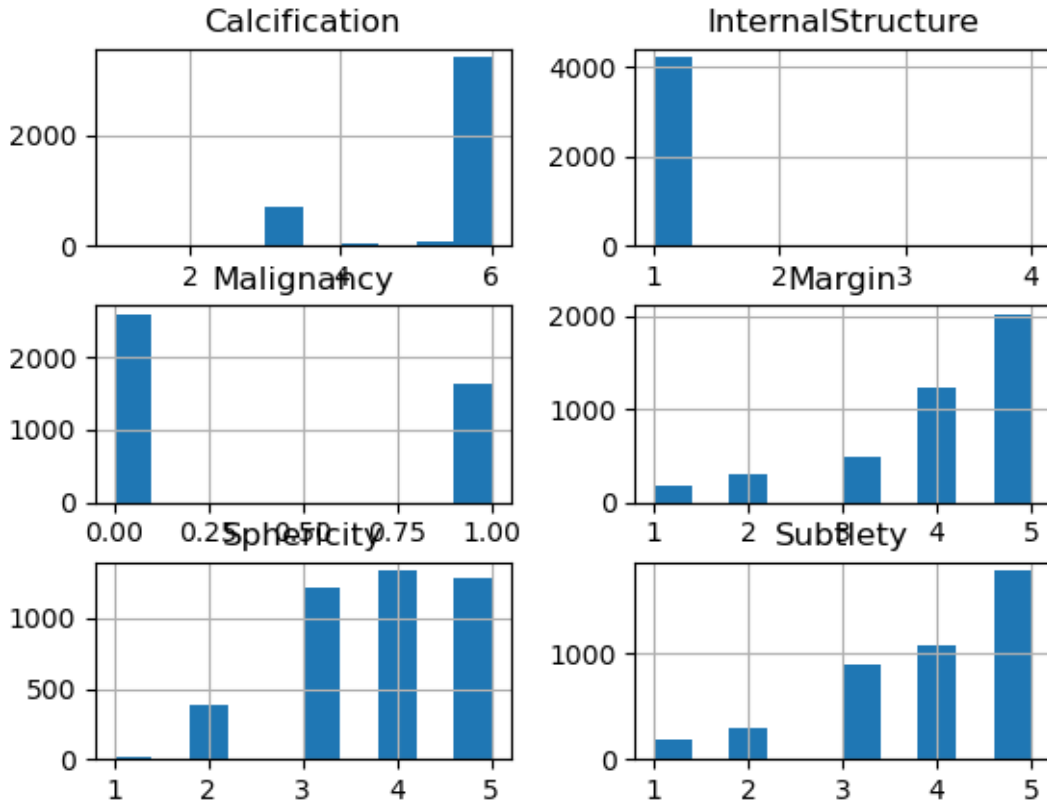


Fig.4.21. The Histograms of the variables in the dataset.

The evaluation results from this stage are the model accuracy, the confusion matrix, and the classification report [3].

The classification report shows the following classifier metrics:

Precision, is the accuracy of the positive predictions [43]

$$precision = \frac{TP}{TP+FP} \quad (4.1)$$

Where TP is the number of true positives, and FP is the number of false positives.

Recall (or Sensitivity), it is the ratio of positive instances that are correctly detected by the classifier [43]

$$recall = \frac{TP}{TP+FN} \quad (4.2)$$

FN is the number of false negatives.

F1-score: is the harmonic mean of precision and recall metrics. The harmonic mean gives much more weight to low values [43].

$$F_1 = \frac{1}{\frac{1}{precision} + \frac{1}{recall}} = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{FN + FP}{2}} \quad (4.3)$$

FP is the number of false positives.

As mentioned in Chapter 3, Six classification methods were used to train the dataset. The mean value, standard deviation, and prediction accuracy for each method are listed in Table (4.5), and the corresponding comparison Box plot is shown in Figure (4.22).

Table.4.5. Comparison Table of the used classification algorithms with the dataset.

#	Algorithm	Mean value	Standard Deviation	Accuracy%
1	Logistic Regression (LR)	0.829071	0.028098	83.06%
2	Linear Discriminant Analysis (LDA)	0.825809	0.028049	82.70%
3	K-Nearest Neighbors classifier (KNN)	0.842097	0.021421	82.82%
4	Classification and Regression Trees (CART)	0.828183	0.024408	80.57%
5	Naïve Bayes (NB)	0.8828786	0.020270	82.11%
6	Support Vector Machine (SVM)	0.852763	0.022368	85.43%

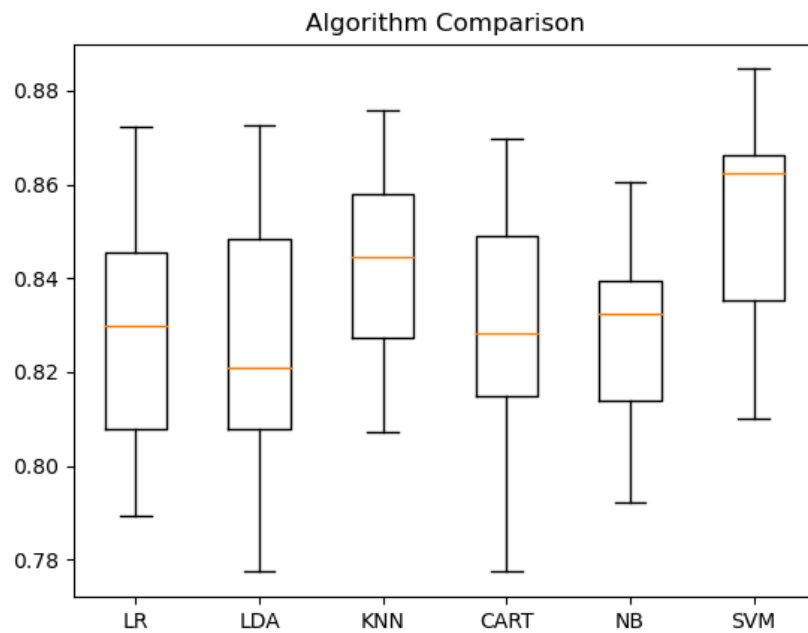


Fig.4.22. Classification Algorithms Comparison Box Plot.

As seen in Table (4.5), the SVM returned the highest prediction accuracy 85.43%.

The SVM confusion matrix for the 844 tested set (20% of the dataset) is shown in Table (4.6):

Table.4.6. SVM Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	421	83
1 Actual	40	300

The classification report for SVM classifier as shown in Table (4.7):

Table.4.7. SVM Classification Report.

	Precision	Recall	F1-score	Support
0	0.91	0.84	0.87	504
1	0.78	0.88	0.83	340

The CART confusion matrix is shown in Table (4.8):

Table.4.8. CART Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	430	74
1 Actual	88	252

The classification report for CART classifier as shown in Table (4.9):

Table.4.9. CART Classification Report.

	Precision	Recall	F1-score	Support
0	0.833	0.85	0.84	504
1	0.77	0.74	0.76	340

The confusion matrix for KNN classifier is shown in Table (4.10):

Table.4.10. KNN Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	437	67
1 Actual	78	262

The classification report for KNN classifier as shown in Table (4.11):

Table.4.11. KNN Classification Report.

	Precision	Recall	F1-score	Support
0	0.85	0.87	0.86	504
1	0.80	0.77	0.78	340

The confusion matrix for LR classifier, which returned the second highest prediction accuracy of 83.06%, is shown in Table (4.12):

Table.4.12. LR Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	432	72
1 Actual	71	269

And the classification report for LR as shown in Table (4.13):

Table.4.13. LR Classification Report.

	Precision	Recall	F1-score	Support
0	0.86	0.86	0.86	504
1	0.79	0.79	0.79	340

The confusion matrix for LDA classifier is shown in Table (4.14):

Table.4.14. LDA Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	433	71
1 Actual	75	265

The classification report for LDA as shown in Table (4.15):

Table.4.15. LDA Classification Report.

	Precision	Recall	F1-score	Support
0	0.85	0.86	0.86	504
1	0.79	0.78	0.78	340

The confusion matrix for NB classifier is shown in Table (4.16):

Table.4.16. NB Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	421	83
1 Actual	68	272

The classification report for NB classifier as shown in Table (4.17):

Table.4.17. NB Classification Report.

	Precision	Recall	F1-score	Support
0	0.86	0.84	0.85	504
1	0.77	0.80	0.78	340

The same dataset was trained and tested using a Neural Network (NN) algorithm, the confusion matrix is shown in Table (4.18):

Table.4.18. NN Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	2269	317
1 Actual	140	1494

And the classification report for NN as shown in Table (4.19)

Table.4.19. NN Classification Report.

	Precision	Recall	F1-score	Support
0	0.94	0.88	0.91	2586
1	0.82	0.91	0.87	1634

The prediction accuracy obtained from applying Neural Network method was 88.20%. The training accuracy is 88% and the testing accuracy is 86.5%, approximately, as shown in Figure (4.23), which is a plot for model accuracy between training set and testing set for the first 350 epochs.

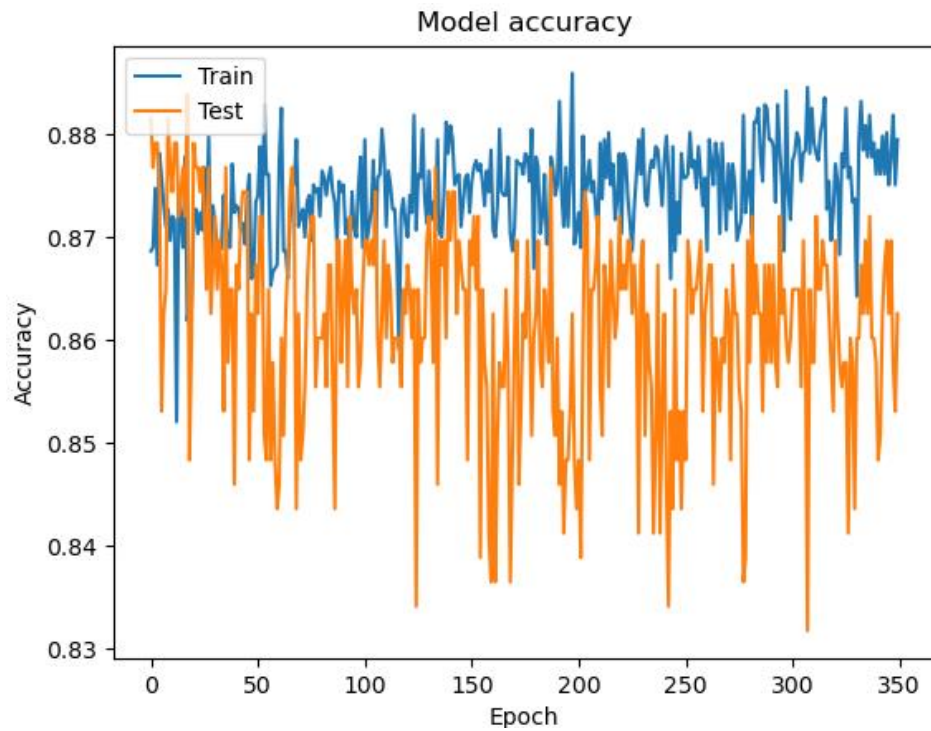


Fig.4.23 Model Accuracy.

The model loss is 0.376 for training set and 0.370 for the testing set as shown in Figure (4.24) for the 350 epochs defined.

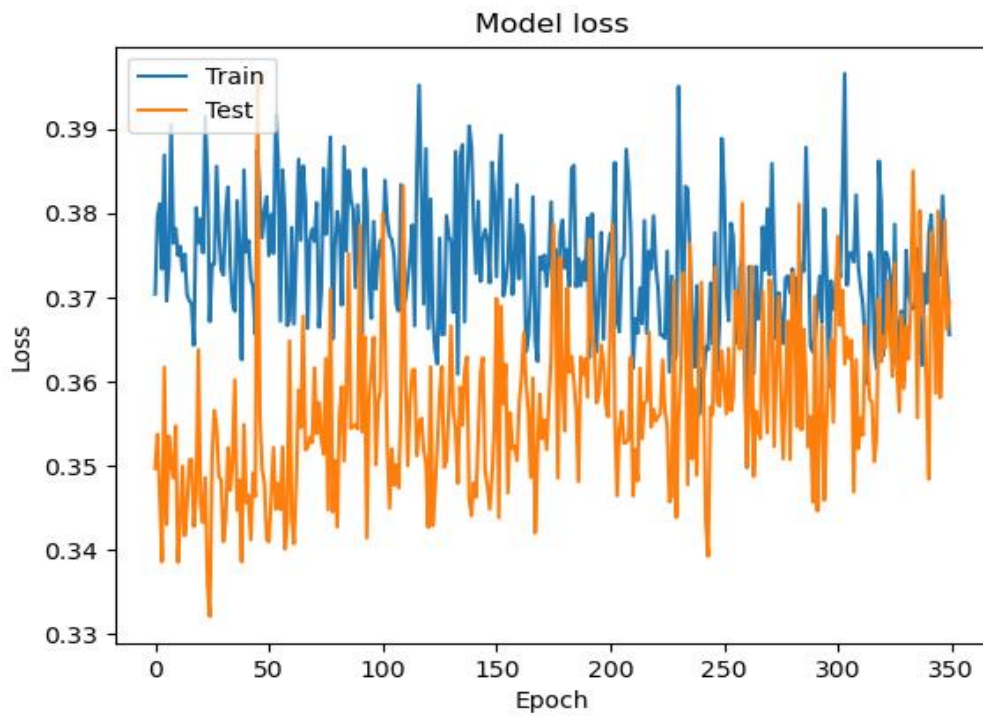


Fig.4.24. Model Loss.

The results obtained showed an improvement compared to the models proposed by the authors of [49] and [53]. The proposed model gave 88.20% prediction accuracy, while the model proposed in [49] gave 86.4%. Error rate in [53] gave 37.5% while the proposed model returned 37.0%

Neural Network Model summary:

Model: "sequential_1"

<i>Layer (type)</i>	<i>Output Shape</i>	<i>Param #</i>
=====		
<i>dense_1 (Dense)</i>	<i>(None, 256)</i>	<i>2304</i>
<i>dropout_1 (Dropout)</i>	<i>(None, 256)</i>	<i>0</i>
<i>dense_2 (Dense)</i>	<i>(None, 128)</i>	<i>32896</i>
<i>dropout_2 (Dropout)</i>	<i>(None, 128)</i>	<i>0</i>
<i>dense_3 (Dense)</i>	<i>(None, 64)</i>	<i>8256</i>
<i>dropout_3 (Dropout)</i>	<i>(None, 64)</i>	<i>0</i>
<i>dense_4 (Dense)</i>	<i>(None, 32)</i>	<i>2080</i>
<i>dropout_4 (Dropout)</i>	<i>(None, 32)</i>	<i>0</i>
<i>dense_5 (Dense)</i>	<i>(None, 32)</i>	<i>1056</i>
<i>dropout_5 (Dropout)</i>	<i>(None, 32)</i>	<i>0</i>
<i>dense_6 (Dense)</i>	<i>(None, 16)</i>	<i>528</i>
<i>dropout_6 (Dropout)</i>	<i>(None, 16)</i>	<i>0</i>
<i>dense_7 (Dense)</i>	<i>(None, 1)</i>	<i>17</i>
=====		
<i>Total params: 47,137</i>		
<i>Trainable params: 47,137</i>		
<i>Non-trainable params: 0</i>		

One epoch sample from training dataset. Train on 3376 samples, validate on 844 samples:

Epoch 316/350:

10/4220 [.....] - ETA: 0s - loss: 0.1402 - accuracy: 1.0000
 350/4220 [=>.....] - ETA: 0s - loss: 0.2688 - accuracy: 0.8914
 610/4220 [===>.....] - ETA: 0s - loss: 0.2835 - accuracy: 0.8803
 860/4220 [=====>.....] - ETA: 0s - loss: 0.2868 - accuracy: 0.8837
 1910/4220 [======>.....] - ETA: 0s - loss: 0.2806 - accuracy: 0.8864
 2020/4220 [======>.....] - ETA: 0s - loss: 0.2853 - accuracy: 0.8817
 2120/4220 [======>.....] - ETA: 0s - loss: 0.2914 - accuracy: 0.8788
 2390/4220 [======>.....] - ETA: 0s - loss: 0.2904 - accuracy: 0.8774
 2400/4220 [======>.....] - ETA: 0s - loss: 0.2909 - accuracy: 0.8775
 2430/4220 [======>.....] - ETA: 0s - loss: 0.2914 - accuracy: 0.8770
 2440/4220 [======>.....] - ETA: 0s - loss: 0.2913 - accuracy: 0.8770
 2630/4220 [======>.....] - ETA: 1s - loss: 0.2947 - accuracy: 0.8749
 2640/4220 [======>.....] - ETA: 1s - loss: 0.2945 - accuracy: 0.8750
 2650/4220 [======>.....] - ETA: 1s - loss: 0.2948 - accuracy: 0.8747
 4110/4220 [======>.] - ETA: 0s - loss: 0.2981 - accuracy: 0.8723
 4190/4220 [======>.] - ETA: 0s - loss: 0.2968 - accuracy: 0.8721
 4220/4220 [======>] - ETA: 4s - loss: 0.2967 - accuracy: 0.8718

A sample output of the prediction set:

[4, 1, 6, 3, 5, 5] => 0 (expected 1)

[5, 1, 6, 3, 4, 5] => 1 (expected 1)

$[5, 1, 6, 4, 5, 5] \Rightarrow 1$ (*expected 0*)

$[5, 1, 6, 4, 5, 5] \Rightarrow 1$ (*expected 1*)

$[3, 1, 6, 5, 5, 5] \Rightarrow 0$ (*expected 0*)

4.4.3. Testing a sample of HU Matrix with the CNN model:

The CNN model described in section 3.4.2. is used to test the ROI extracted and stored in HU matrix(Table3.2).

The dataset resulted from the second approach in features extraction is similar in properties and values of the dataset used in training the model in previous section. The datasets from the first approach with ROI's geometrical properties approach and the third approach with DICOM matrix will be used if a proper pre-defined dataset is available for training.

The model used had 10 epochs, with batch size equals to 50.

The dataset used to train and test the model was relatively small 352×62501 samples, compared to the whole dataset, due to machine and memory limitations for processing large image dataset with 16 GB RAM installed on a personal desktop.

The model was trained with labeled nodules obtained from the LIDC-IDRI dataset as mentioned in Section 4.4.2.

The accuracy achieved is 93.75%.

The confusion matrix for CNN model is shown in Table (4.20):

Table.4.20. CNN Confusion Matrix.

	0 Predicted	1 Predicted
0 Actual	156	0
1 Actual	22	174

The classification report for CNN model as shown in Table (4.21):

Table.4.21. CNN Classification Report.

	Precision	Recall	F1-score	Support
0	0.88	1.00	0.93	156
1	1.00	0.89	0.94	196

In order to test the model with the dataset obtained from image processing stages, the dataset was taken from 48 patients with total ROI extracted of 33542, HU matrix was used as the prediction set for the model. It has 33542 ROIs. The model predicted 179 as cancer and the remaining ROIs are classified as non-cancer.

This is considered satisfying prediction as the dataset is greatly imbalanced, considering the ROIs extracted are all the nodes the system can define as candidates within the lungs.

Chapter Five

Discussion and Conclusion

Chapter Five

Discussion and Conclusion

In this thesis, we proposed a system to detect lung cancer from CT images. The proposed system consists of four main stages, preprocessing stage, segmentation stage, features extraction stage, and machine learning and classification stage. In this system, CT images are passed through the preprocessing stage for image enhancement and noise removal. In the second stage, image is segmented to define boundaries between different tissues. In the third stage, important features are extracted. In the fourth stage, a machine learning procedure is applied to the extracted features to identify cancer regions. Finally, the system's detected cancerous regions are evaluated against similar work in literature.

The related steps conducted for the work are thoroughly discussed as follows:

Firstly, image was read by MATLAB which converts original HU values from bit depth of 12-bit into 16-bit, so a linear equation is applied to convert these pixels' values back to HU to make sure no data is lost.

Secondly, image was segmented with sequential steps; analyzing the image using Gabor filter of 28-Gabor filters and 2 spatial features for each pixel processed in the image. Gabor helped removing noise when found in some images and increased the quality of extracting ROI for later stages. By comparing Gabor with Median and Weiner filters; Gabor identified all

regions inside lungs from other organs more accurately. Detecting edges around candidate ROIs was performed using Canny Method. Canny method's returned high detection accuracy compared with Sobel and Prewitt, increasing the chance of isolating ROIs. Otsu's Algorithm was applied for binarization step to evaluate proper threshold value for background and foreground separation. The threshold value returned made this separation much convenient than Global Thresholding or Local Thresholding.

The Watershed algorithm isolates ROIs for features extraction. It is considered the best method for determining pixels assigned to a single ROI [2].

Thirdly, Features Extraction stage was performed on selected regions of interests producing a table of shape properties, and another two matrices of pixels' intensity values of both bit depths 12-bit and 16-bit.

Finally, annotations' pathologic features extracted from DICOM images were inserted into different machine learning algorithms to classify cancer.

Six classification algorithms were compared while training the extracted features from the annotations; as mentioned in Chapter 3. According to the results, SVM gave the highest classification accuracy among them with 85.43% percentage when applied on system's dataset.

Neural Network was used as the second approach for classification stage. The model consists of one input layer, five hidden layers, and one

output layer. The hidden layers are followed by a dropout layer to help reduce data overfitting. Also, they compared Sigmoid and SoftMax activation functions. Sigmoid returned 88.2% while SoftMax was 38.72%. This showed that Sigmoid was the most suitable method with the proposed model. Adam Optimizer in the fully connected output layer was used to return the final prediction accuracy value.

These steps are followed by a stage of testing the data extracted from image processing techniques with the Convolutional Neural Network model trained in this thesis.

The presented work has fulfilled the objective of designing and implementing a system that can detect the presence of lung cancer in CT images. By using the image processing techniques in the designed system, the methods implemented that were used in the system's stages were Gabor Filter for image enhancement, according to [2] and [3], it was the most suitable method between the algorithms they compared in their researches. In Segmentation stage; Canny algorithm was tested in the system and compared to other edge detection methods and it returned high detection rate, Binarization step was conducted by Otsu's method which was compared in the system with other methods and achieved second suitable threshold value after the GUI thresh-tool function which was eliminated due to the need of manual setting of the threshold , and Watershed algorithm was used for separating ROI boundaries since it is approved to achieve accurate separation level as mentioned in [2,3,4,16,26]. Features

extraction stage was implemented with several Morphological techniques to determine ROI and store the features and the pixels' intensity values corresponding to ROI extracted. The suspicious area of tumor detected by the system with a success percental of 100% for nodules inside lungs and detection percentage of 72.92%, for the 50 random cases, with nodules identified as cancer located inside and near boundaries of the lungs.

Design and implement a machine learning system to classify whether there exists lung cancer or not. Six algorithms were tested and compared to define the suitable algorithm with system's dataset as explained in Chapter 4. SVM algorithm returned high detection accuracy compared to the other algorithms with 85.43%.

Feed the annotations' features obtained from the dataset into neural network system to train the system to classify the output into two categories. The accuracy obtained was 88.20%.

Inserting the ROIs to the CNN model trained in this thesis returned a detection accuracy of 93.75% and detected 179 cancerous nodules from the dataset used with this thesis.

Limitations:

- Failed to detect tumors located at lungs' boundaries since the threshold value obtained recognized them as background along with chest wall.
- Otsu's method used for binarization stage didn't give a suitable threshold value for each image to separate background from foreground.
- Hardware limitation affected the sample size that the software can analyze.

Future Work:

- 1- Larger balanced dataset for both categories to enhance classification stage.
- 2- Use of local data set when available by local hospitals and specialists.
- 3- Use similar algorithm for detection of other respiratory symptoms such as COVID-19 symptoms and Pulmonary edema [44,45].
- 4- Implement more advanced machine learning, i.e., cluster, cloud, detection on the cloud, work on an app.
- 5- Implement advanced steps to detect cancer based on its stage.
- 6- To obtain samples that can be tested using the other features extraction approaches presented in the thesis.

References

- [1] Cancer Image Archive, <http://www.cancerimagingarchive.net/>.
- [2] Bhagyashri G. Patil, Prof. Sanjeev N. Jain, **Cancer Cells Detection Using Digital Image Processing Methods**, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 3 Issue 4 March 2014, ISSN: 2278-621X.
- [3] Mokhled S. AL-TARAWNEH, **Lung Cancer Detection Using Image Processing Techniques**, Leonardo Electronic Journal of Practices and Technologies, ISSN 1583-1078, Issue 20, January-June 2012.
- [4] Mr. Vijay, A. Gajdhane 1, Prof. Deshpande L.M., **Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques**, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. III (Sep – Oct. 2014), PP 28-35.
- [5] Amjed S. Al-Fahoum, Eslam B. Jaber, Mohammed A. Al-Jarrah, **Automated detection of lung cancer using statistical and morphological image processing techniques**, Journal of Biomedical Graphics and Computing, 2014, Vol. 4, No.2.
- [6] Arvind Kumar Tiwari, **PREDICTION OF LUNG CANCER USING IMAGEPROCESSING TECHNIQUES: A REVIEW**, Advanced Computational Intelligence: An International Journal (ACII), Vol.3, No.1, January 2016.

- [7] Disha Sharma, Gagandeep Jindal, **Identifying Lung Cancer Using Image Processing Techniques**, International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2011).
- [8] Md. Badrul Alam Miah, Mohammad Abu Yousuf, **Detection of Lung Cancer from CT Image Using Image Processing and Neural Network**. DOI:10.1109/ICEEICT.2015.7307530. Conference: International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2015, At JU, Savar, Dhaka, Bangladesh., Volume: IEEE.
- [9] Raja Rao.Chella, **A Qualitative Review on Image Processing Algorithms to Detect Early-Stage Lung Cancer**, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-6 Issue-3, February 2017.
- [10] Santosh Singh, Yogesh Singh, Ritu Vijay, **An Evaluation of Features Extraction from Lung CT Images for the Classification Stage of Malignancy**, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Special Issue - AETM'16 (2016).
- [11] Imaging Technology News Website, **Lung Screening: X-Ray or CT: Which One Better Detects Lung Cancer?**, <https://www.itnonline.com/article/lung-screening-x-ray-or-ct-which-one-better-detects-lung-cancer>, last accessed 14-4-2020.

- [12] American Cancer Society, **Tests for Non-Small Cell Lung Cancer**, <https://www.cancer.org/cancer/non-small-cell-lung-cancer/detection-diagnosis-staging/how-diagnosed.html>, last accessed 14-4-2020.
- [13] The World Health Organization, <http://www.who.int> .
- [14] Gonzalez, E. Woods, L. Eddins, **Digital Image Processing Using MATLAB**, Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©2003.
- [15] O. Marques Wiley, **Practical Image and Video Processing Using MATLAB**, _IEEE2011BBS.
- [16] Suren Makajua, P.W.C. Prasad, Abeer Alsadoona, A. K. Singhb, and A. Elchouemic, **Lung Cancer Detection using CT Scan Images**, 6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8, December 2017, Kurukshetra, India, Procedia Computer Science 125 (2018) 107–114.
- [17] Niveen M. E. Abu-Rmeileh, Emilio Antonio Luca Gianicolo, Antonella Bruni, Suzan Mitwali, Maurizio Portaluri, Jawad Bitar, Mutaem Hamad, Rita Giacaman, and Maria Angela Vigotti, **Cancer mortality in the West Bank, Occupied Palestinian Territory**, BMC Public Health. 2016; 16: 76.
- [18] Perumal, Velmurugan, **Lung cancer detection and classification on CT scan images using enhanced artificial bee colony optimization**, (2018).

- [19] Nunzio, Tommasi, Agrusti, Cataldo, De Mitri, Favetta, Maglio, Massafra, Quarta, Torsello, Zecca, Bellotti, Tangaro, Calvini, Camarlinghi, Falaschi, Cerello, and Oliva, **Automatic Lung Segmentation in CT Images with Accurate Handling of the Hilar Region**, (2011).
- [20] Magdy, Zayed, and Fakhr, **Automatic Classification of Normal and Cancer Lung CT Images Using Multiscale AM-FM Features**, (2015).
- [21] Singh and Asuntha, **Image Processing used for Lung Cancer Detection in Medical Images**, (2016).
- [22] Kuruvilla, Gunavathi, **Lung cancer classification using neural networks for CT images**, (2013).
- [23] Kumar and Kumar, **Lung Segmentation Using Region Growing Algorithm**, (2014).
- [24] Ada, and Kaur, **Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images**, (2013).
- [25] Orozco, Villegas, Sánchez, Domínguez, and Alfaro, **Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine**, (2015).
- [26] Sowjanya, Bharath, and Yadav, **Visualization and Segmentation of Lung CT Scan Images for Cancer Detection**, (2013).

- [27] Ireya. E. Igbiosa, **Comparison of Edge Detection Technique in Image Processing Techniques**, ITEE Journal Information Technology & Electrical Engineering, ISSN: - 2306-708X, Volume 2, Issue 1 February 2013.
- [28] Galbán, Craig J et al., **Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression**. Nature medicine vol. 18,11 (2012): 1711-5. doi:10.1038/nm.2971.
- [29] Subha B. Basu, **Is Lung Cancer better detected using an X-Ray or CT Scan?**, June 27, 2006, <https://www.itnonline.com/article/lung-cancer-better-detected-using-x-ray-or-ct-scan>, last accessed 20-3-2018.
- [30] Alenrex Maity, Anshuman Pattanaik, Santwana Sagnika, Santosh Pani, **A Comparative Study on Approaches to Speckle Noise Reduction in Images**, 2375-5822/15 \$31.00 © 2015 IEEE, DOI 10.1109/CINE.2015.36.
- [31] A. K. Jain and F. Farrokhnia, **Unsupervised Texture Segmentation Using Gabor Filters**,1991.
- [32] Fatma Latifoğlu, **A novel approach to speckle noise filtering based on Artificial Bee Colony algorithm: An ultrasound image application**, Computer Methods and Programs in Biomedicine, Volume 111, Issue 3, September 2013, Pages 561-569.

- [33] M.Janani, K.Nandhini , K.Senthilvadivel ,S.Jothilakshmi, **Digital Image Technique using Gabor Filter and SVM in Heterogeneous Face Recognition**, Research Inventy: International Journal of Engineering And Science Vol.4, Issue 4 (April 2014), PP 45-52 Issn (e): 2278-4721, Issn (p):2319-6483.
- [34] Ayman El-Baz, Matthew Nitzken, Ahmed Elnakib, Fahmi Khalifa, Georgy Gimel'farb, Robert Falk, and Mohamed Abou El-Ghar, **3D Shape Analysis for Early Diagnosis of Malignant Lung Nodules**, G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 175–182, 2011. c Springer-Verlag Berlin Heidelberg 2011.
- [35] GangadharShobha, ShantaRangaswamy, **Handbook of Statistics**, Volume 38, 2018, Pages 197-228, Chapter 8 - Machine Learning.
- [36] Carlo Ricciardi, Antonio Saverio Valente, Kyle Edmund, **Linear discriminant analysis and principal component analysis to predict coronary artery disease**. First Published January 23, 2020.
- [37] M.M Ghiasi, S Zendehboudi, A.A Mohsenipour, **Decision tree-based diagnosis of coronary artery disease: CART model**. Comput Methods Programs Biomed. 2020;192:105400. doi:10.1016/j.cmpb.2020.105400.

- [38] Nafizatus Salmi and Zuherman Rustam, **Naïve Bayes Classifier Models for Predicting the Colon Cancer**, 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546.
- [39] Sarah M. Ayyad , Ahmed I. Saleh, Labib M. Labib, **Gene expression cancer classification using modified K-Nearest Neighbors technique**, Biosystems Volume 176, February 2019, Pages 41-51.
- [40] Prabhu, Towards Data Science:
 “<https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148>” . Last accessed 6-5-2020.
- [41] Aurélien Géron, **Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems**, (2017, O’Reilly Media)
- [42] Armato III, Samuel G., McLennan, Geoffrey, Bidaut, Luc, McNitt-Gray, Michael F., Meyer, Charles R., Reeves, Anthony P., ... Clarke, Laurence P. (2015). **Data From LIDC-IDRI**. The Cancer Imaging Archive, (referenced in 26-9-2019), <http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX>.
- [43] Aurélien Géron, **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems**, 2nd Edition.
- [44] Soldati G, Demi M. **The use of lung ultrasound images for the differential diagnosis of pulmonary and cardiac interstitial**

- pathology.** J Ultrasound. 2017;20(2):91-96. Published 2017 Apr 7.
doi:10.1007/s40477-017-0244-7
- [45] RAMALHO; REBOUCAS FILHO; MEDEIROS, and CORTEZ. **Lung disease detection using feature extraction and extreme learning machine.** Rev. Bras. Eng. Bioméd. [online]. 2014, vol.30, n.3 [cited 2020-06-08], pp.207-214.
- [46] Bhalerao, R. Y., Jani, H. P., Gaitonde, R. K., & Raut, V. (2019). **A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks.** 2019 5th International Conference on Advanced Computing & Communication Systems, (ICACCS). doi:10.1109/icaccs.2019. 8728348.
- [47] Muthazhagan, B., Ravi, T. & Rajiniginath, D. **An enhanced computer-assisted lung cancer detection method using content-based image retrieval and data mining techniques.** J Ambient Intell Human Compute (2020). <https://doi.org/10.1007/s12652-020-02123-7>.
- [48] TD DenOtter, J. Schubert, **Hounsfield Unit.** In: StatPearls. StatPearls Publishing, Treasure Island (FL); 2020.
- [49] Wei Li, Peng Cao, Dazhe Zhao, and Junbo Wang, **Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images,** Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine, Volume 2016, Article ID 6215085, 7 pages
<http://dx.doi.org/10.1155/2016/6215085>.

- [50] Guanghui Han, Xiabi Liu, Nouman Q. Soomro, Jia Sun, Yanfeng Zhao, Xinming Zhao, and Chunwu Zhou, **Empirical Driven Automatic Detection of Lobulation Imaging Signs in Lung CT**, Volume 2017, ArticleID 3842659, 15 pages, <https://doi.org/10.1155/2017/3842659>.
- [51] Matthew C. Hancock, Jerry F. Magnan, **Lung nodule malignancy classification using only radiologist quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods**. *SPIE Journal of Medical Imaging*. Dec. 2016. <http://dx.doi.org/10.1117/1.JMI.3.4.044504>.
- [52] PyIidc Library Documentation: <https://pyIidc.github.io/index.html>, last accessed 3-8-2020.
- [53] Krizhevsky, A., Sutskever, I., & Hinton, G. E., **ImageNet classification with deep convolutional neural networks**. Communications of the ACM, 60(6), 84–90. doi:10.1145/3065386 (2017).
- [54] The specialist portal for the chemical sector and its suppliers in Europe and the USA, https://www.chemeurope.com/en/encyclopedia/Hounsfield_scale.html, last accessed 20-9-2020.

جامعة النجاح الوطنية

كلية الدراسات العليا

تصميم وتطبيق نظام محوسب للكشف المبكر لسرطان الرئة باستخدام الشبكة العصبونية ومعالجة الصور

إعداد

ديمة سهراب صوالحة

إشراف

د. عدنان سلمان

قدمت هذه الأطروحة استكمالاً لمتطلبات الحصول على درجة الماجستير في الحوسبة المتقدمة
من كلية الدراسات العليا في جامعة النجاح الوطنية في نابلس- فلسطين.

ب

تصميم وتطبيق نظام محوسب للكشف المبكر لسرطان الرئة باستخدام الشبكة العصبونية

ومعالجة الصور

إعداد

ديمة سهراب صوالحة

إشراف

د. عدنان سلمان

الملخص

يعتبر سرطان الرئة أكثر أنواع السرطانات شيوعاً بين الذكور في جميع أنحاء العالم. وهو يمثل 1 من كل 5 حالات وفاة بسبب السرطان ويحدث غالباً بين سن 55 و65. وهذا هو الحال أيضاً في فلسطين، حيث يمثل سرطان الرئة نسبة 22.8% من الإصابات بين الذكور.

يعد الكشف المبكر عن سرطان الرئة في المراحل الأولية خطوة حاسمة في عملية العلاج، حيث يمكن أن يزيد معدل البقاء على قيد الحياة بشكل ملحوظ.

في هذه الرسالة، تم تطبيق تقنيات معالجة الصور على صور التصوير المقطعي المحوسب (CT) لسرطان الرئة للعديد من المرضى لتحديد مناطق السرطان وحجمه. كما تم إجراء دراسة مقارنة بين خوارزميات معالجة الصور المختلفة على عدة صور لتحديد الخوارزميات الأكثر دقة التي سيتم استخدامها في عملية فحص سرطان الرئة. لقد تم استخدام خوارزميات التعلم الآلي والشبكات العصبونية وعمل مقارنة دراسية بينها لتحديد إذا كان الورم سرطان ام لا.