Classification of Lung Carcinoma using Texture Features of Ultrasound Images

Dissertation submitted in partial fulfillment of the requirements for the Degree of
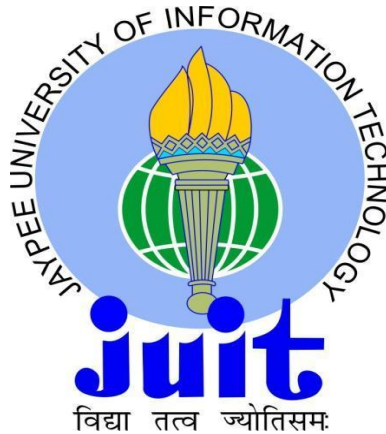
MASTERS OF TECHNOLOGY

By

AMAN DEEP
Enrollment No: 152012

UNDER THE GUIDANCE OF

Dr. Shruti Jain

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
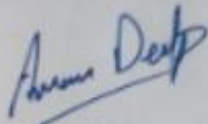JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
May-2017

TABLE OF CONTENTS

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the M-Tech thesis entitled "**Classification of Lung Carcinoma using Texture Features of Ultrasound Images**" submitted at **Jaypee University of Information Technology, Waknaghat India**, is an authentic record of my work carried out under the supervision of **Dr. Shruti Jain**. I have not submitted this work elsewhere for any other degree or diploma.

Aman Deep

Department of Electronics and Communication Engineering
Jaypee University of Information Technology, Waknaghat, India

JAYPEE UNIVERSITY OF INFORMATIO TECHNOLOGY
(Established by H.P. State Legislative vide Act No. 14 of 2002)
P.O. Waknaghat, Teh. Kandaghat, Distt. Solan - 173234 (H.P.) INDIA
Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: +91-01792-245362

CERTIFICATE

This is to certify that the work reported in the M.Tech project report entitled **"Classification of Lung Carcinoma using Texture Features of Ultrasound Images"** which is being submitted by **Aman Deep** in fulfillment for the award of Masters of Technology in Electronics and Communication Engineering by the Jaypee University of Information Technology, is the record of candidate's own work carried out by him under my supervision. This work is original and has not been submitted partially or fully anywhere else for any other degree or diploma.

01/05/17.

Dr. Shruti Jain
Associate Professor
Department of Electronics & Communication Engineering
Jaypee University of Information Technology, Waknaghat

ACKNOWLEDGEMENT

I am highly indebted to Dr. Shruti Jain for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her constant support as we ascend toward the completion of this project.

I would like to express my gratitude towards Mr. Sahil Bhusri for his kind co-operation in spite of being occupied with his work, he managed to guide me and help me with the problems which I faced during the completion of this project.

I would like to express my special gratitude to my family and my friend Shreya Sharma for their immense support, help and motivation throughout the work. My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

LIST OF ACRONYMS & ABBREVIATIONS

AC	Adenocarcinomas
ANN	Artificial Neural Network
BMP	Bitmap
CAD	Computer Aided Diagnosis
CADe	Computer Aided Detection System
CADx	Computer Aided Diagnosis System
CCL	Connected Component Labeling
CM	Confusion Matrix
CR	Compression Ratio
CT	Computed Tomography
DFT	Discrete Fourier Transform
DICOM	Digital Imaging and Communications in Medicine
FITS	Flexible Image Transport System
FN	False Negative
FOS	First Order Statistics
FP	False Positive
FPS	Fourier Power Spectrum
GA	Genetic Algorithm
GGO	Ground Glass Opacity
GIF	Graphic Interchange Format
GLCM	Gray Level Co-occurrence Matrix
GLDS	Gray Level Difference Statistics
GWT	Gabor Wavelet Transform
HIV	Human Immunodeficiency Virus
ICA	Individual Classification Accuracy
ICAB	Individual Classification Accuracy of Benign
ICAM	Individual Classification Accuracy of Malignant
IGMC	Indra Gandhi Medical College
JPEG	Joint Photographic Experts Group
K-NN	K- Nearest Neighbor

LC	Large cell Carcinomas
LDA	Linear Discriminant Analysis
LibSVM	Library Support Vector Machine
MATLAB	Matrix Laboratory
MRI	Magnetic Resonance Imagining
MTANN	Massive Training Artificial Neural Network
NGTDM	Neighborhood Gray Tone Difference Matrix
NN	Neural Network
NSCLC	Non Small cell lung cancer
OCA	Overall Classification Accuracy
PCA	Principal Component Analysis
PDF	Probability Distribution Function
PNG	Portable Network Graphics
PNN	Probabilistic Neural Network
PSNR	Peak Signal-to-noise Ratio
RBF	Radial Basis Function
ROC	Receiver Operating Characteristics
ROI	Region of Interest
RMSE	Root Mean Square Error
SC	Squamous Carcinomas
SCLC	Small cell lung cancer
SD	Standard Deviation
SFM	Statistic Feature Matrix
SSVM	Smooth Support Vector Machine
SVM	Support Vector Machine
TEM	Texture Energy Measurement
TFV	Texture Feature Vector
TI	Texture Image
TIFF	Tag Image File Format
TNM	Tumor Node Metastases
TN	True Negative
TP	True Positive
WPT	Wavelet Packet Transform

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Lung cancer is the main cause of cancer death in women and men across the planet. The lung carcinoma is divided into two categories: Small Cell Lung Cancer (SCLC) and Non Small Cell Lung Cancer (NSCLC). A large number of techniques are being used for detection and diagnosis of Lung Cancer. The Computer Aided Diagnosis (CAD) is the most common and accurate technique for early detection of abnormal cells which can cause cancer to healthy lung tissues. CAD system works on the basis of analysis of condition of ultrasound images. CAD system follows different steps: Data collection (ultrasound image), Data Preprocessing (ROI Selection), Feature Extraction, Data Partitioning (hold- out method), Feature Classification and Result Calculation. To classify input ultrasound images into benign and malignant, different classifiers were used. The system work is based on the calculation of parameters such as individual accuracy, overall accuracy and sensitivity. These benchmarks are obtained by calculating the matrix of Support Vector Machine (SVM). The results were obtained by using various features using Statistical Methods. The best results achieved were having accuracy of 91.4% by using Gray Level Difference Statistics (GLDS). The results will be used in CAD system for detection of Lung Cancer in initial stage to enhance the capability of survival of patient.

CHAPTER 1

INTRODUCTION

Cancer is leading cause of death in human across the world. It is basically unwanted and uncontrolled growth of cells in particular organ of body. Lung Cancer is of various types according to its occurrence in various body parts. Through which Lung Cancer is most common occurring Cancer in men and women worldwide. It contributes 13% of total number of diagnosed cases of all type of cancers. For men it contributes approximately 17% of all types of cancers. 8.5% of Lung Cancer out of all the other types of cancers occurs in women.

Lung Cancer is uncontrolled cell growth in lung tissue. It originates in (trachea) the main airway (bronchus) or the lung tissue. It occurs due to the deformity in genetic material of lung cells. The main cause of lung cell deformity is tobacco which is the main reason of 88% deaths of Lung Cancer. Only 12-13% lung cancer occurs in non-smokers. On basis of size cancer can be categorized in two types: Benign and Malignant. Benign type of lung cancer is the starting of abnormality in lung and it is present in only one lung. This is the initial stage of lung cancer and it usually undetected[1-3]. But when benign is out of control then it grows into malignant. Malignant type is that stage of cancer when it spread in both lungs and also start spreading in other body parts. Lung Cancer is divided in two types on the basis of benign and malignant:

- Small Cell Lung Cancer (SCLC)
- Non Small Cell Lung Cancer (NSCLC)

SCLC:

The SCLC is initial stage of Lung Cancer, when the abnormality starts growing in only one lung and smaller in size. Most of the time, it is undetectable because of its smaller size. There are only 20% cases of SCLC occurring in Lung cancer[4]. It is rapidly growing and very aggressive type of lung cancer. Only 1% of this Lung Cancer present in non-smokers. When benign type abnormality occurs in lung tissues and growing as tumor then it is of SCLC type cancer. The Fig 1.1 shows how the cell looks under microscope:
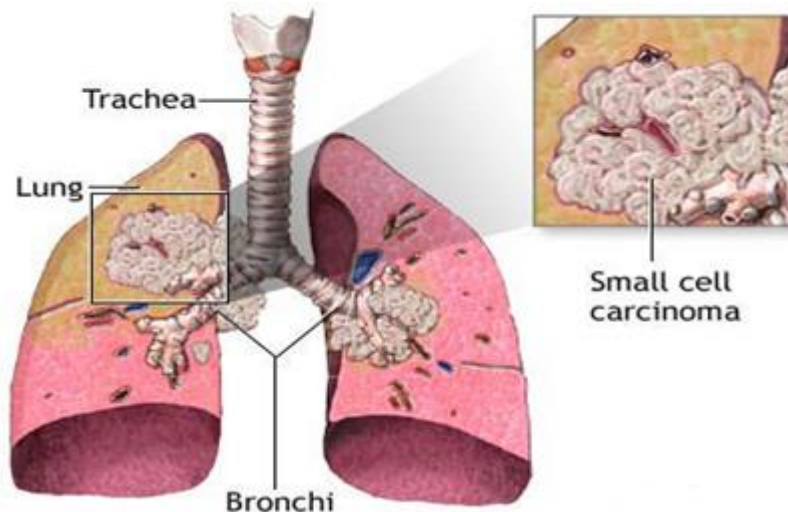
Fig 1.1: Structure of SCLC [5]

Smoking is the major risk factor for SCLC. The increase in the quantity or time of smoking will increase in the risk of occurring small cell lung cancer. Then causes of SCLC are:

- Passive smoking
- Being exposed to radioactive gases like Radon.
- Being exposed to superfine fibers like Asbestos fibers.
- Air pollution
- Hereditary problem
- Infected with Human Immunodeficiency Virus (HIV)

As human age is increases the risk factor of cancer is also increases as compare to young age. But when smoking is combined with old age the risk factor reaches to extreme dangerous level in Lung Cancer.

NSCLC:

NSCLC is basically the developed size of SCLC, when this small size increases and spread into lungs and other body organs then the carcinoma is NSCLC. The NSCLC is divided into three types:

- Adenocarcinomas,
- Squamous Carcinoma

- Large Cell Carcinoma

Adenocarcinomas:

The adenocarcinomas are of most common occurring lung cancer which occurs approximately 50% in all Lung Cancer patients. These cancers start in secret substance of cell, known as mucus that is early version of the cells[5]. This type of Cancer occurs mainly in smokers and former smokers, but it is also most common type of Lung Cancer in non-smokers. It usually occurs in young people as compared to older. More women are suffering from this cancer as compare to the men. So, young women are most likely to suffering from this type of Lung Cancer. Adenocarcinoma is usually growing on the outer layer of lungs. The growth rate of this Lung Cancer is very slow as compare to other types of cancer. Because of its slow growth, it is more likely to be detected before it starts spreading in other body organs. The Fig 1.2 shows the structure of adenocarcinoma:
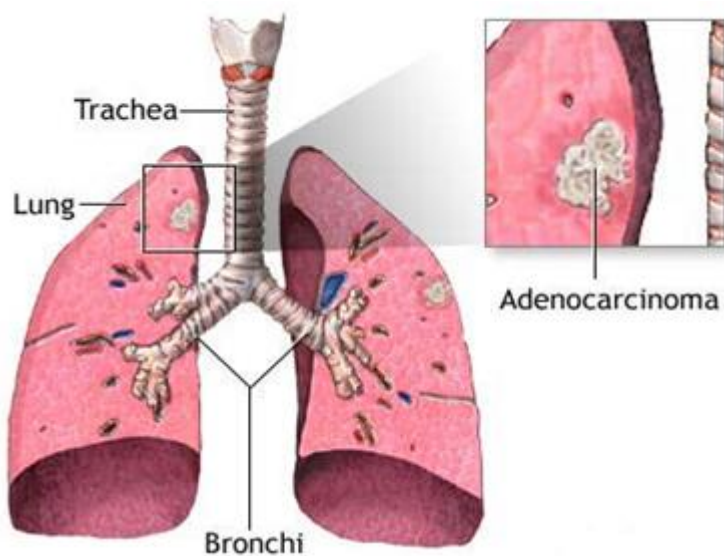
Fig 1.2: Structure of Adenocarcinoma [5]

Squamous Carcinoma:

The Squamous type of Lung Cancer is present in 30% cases of lung cancer. In past, It is more common occurring Lung Cancer than adenocarcinoma. This cancer is grows in the flat cells of lungs that are present inside of the airways of the lungs, also known as squamous cells. These cells present in the central part of the lungs, near the main airways. The Fig 1.3 shows structure of squamous carcinoma:
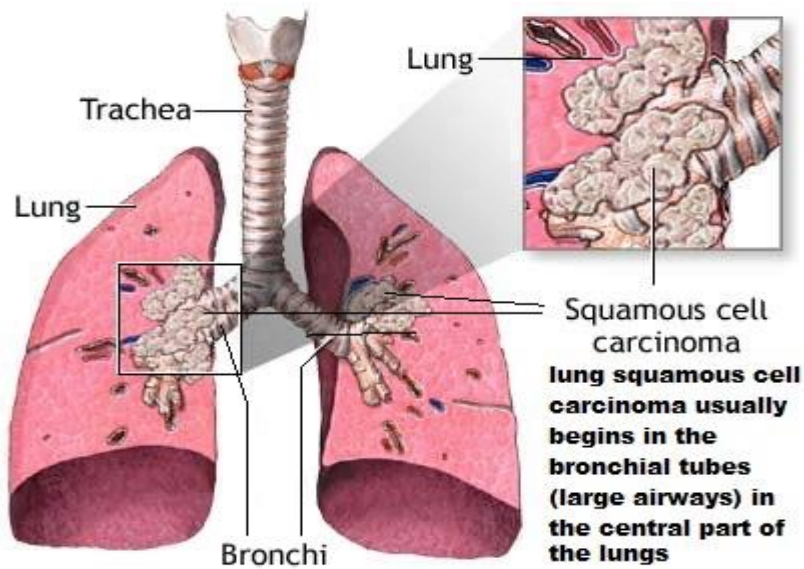
Fig 1.3: Structure of Squamous Carcinoma [5]

Large Cell Carcinoma:

The third type is Large Cell Carcinomas, which is least common occurring Lung Cancer. It is referred to as undifferentiated carcinomas. This type of NSCLC can present in any part of lungs[48]. They are very rapid growing Lung cancers and starts spreading in body very quickly. It is very similar to SCLC. Non SCLC is often treated with surgery whereas for SCLC the chemotherapy and radiotherapy are responds better. The structure of large cell carcinoma is shown in Fig 1.4:
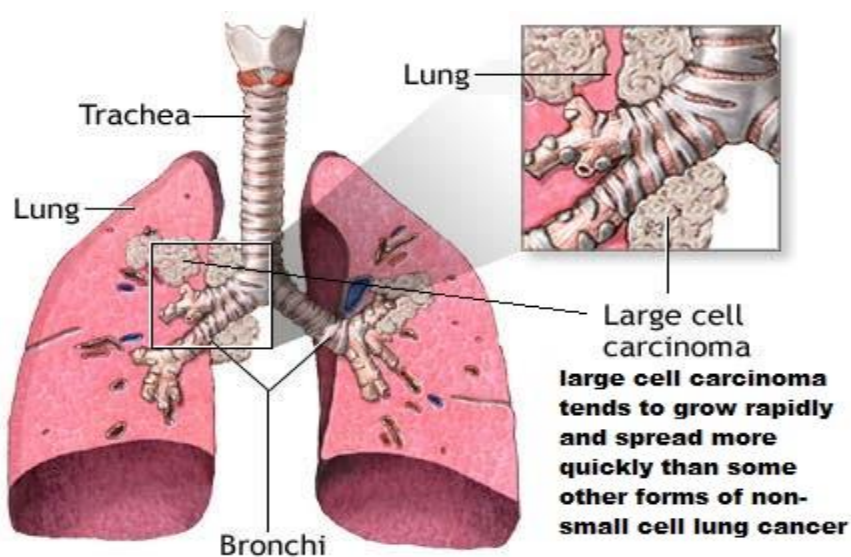
Fig 1.4: Structure of Large Cell Carcinoma [48]

Lung cancer is measured on the bases of how big it is and how abnormal its cell looks under the microscope. The abnormal cell growth is explained using grades:

Grade 1: In the initial stage of lung cancer, the cells look like normal cells and less likely to spread.

Grade 2: Cells look abnormal and start spreading. This grade is also known as moderate grade.

Grade 3 and 4: The cells are completely deformed and growth is very quick. At this grade the cancer start spreading in other body parts. This grade is known as poorly differentiated or high grade. The grading is for NSCLC.

If Lung Cancer is detected in early stage of cancer, then it is curable. The primary cause of Lung Cancer is smoking tobacco. About 88% cases of Lung Cancer are due to smoking. Smoking can be categorized into two types: Active smoking and Passive smoking. The active smokers are those who smoke tobacco and passive smoker are those who are not smoking directly but are inhaling the smoke produced by smoker. The pie chart shows in Fig 1.5:
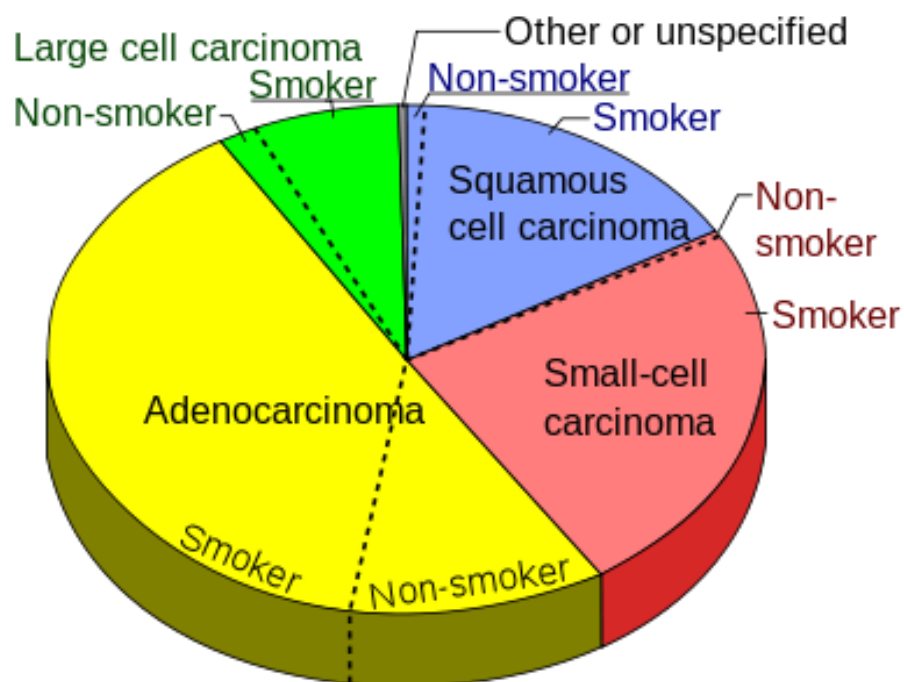
Fig 1.5: Pie chart of Lung Cancer for NSCLC and SCLC with smokers versus non-smokers. [36]

Only 10-12% Lung cancer cases are due to secondary causes. The secondary causes of lung cancer are passive smoking, asbestos fiber, radon gas, lung diseases etc. The graph of risk of death in smokers is presented in Fig 1.6:
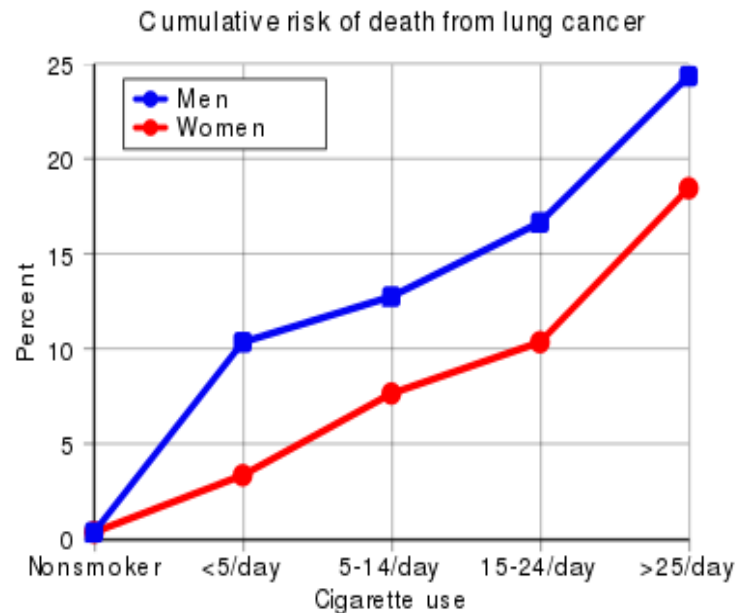
Fig 1.6: Risk of death in smokers [49]

In NSCLC, the TNM staging is used for classification purpose. This staging is based on the size of the primary Tumor, lymph Node involvement and Metastasis. This TNM staging gives information about tumor.

- T- The size and position of tumor.
- N- Presence of spread into lymph node.
- M- Presence of Metastases.

In treatment of lung cancer, the time of detection of abnormality and tumor size plays an important role in patient's life. If the disease is detected in early stage then it doesn't only help in proper treatment and diagnosis of tumor but also minimizes the risk of death. For early detection and diagnosis purpose, there are number of techniques available these days[6-9]. The X-Ray, Sputum cytology, Ultrasonography, Magnetic Resonance Imagining (MRI), Computed Tomography (CT) and biopsy are the main techniques of present time. When a sample of lesion is taken out from the tumor for diagnosis and

analysis that results in unbearable pain to patient, this is the procedure of biopsy. For the comfort of patient, to reduce unnecessary biopsies CT scan and Ultrasonography is most frequent method that used. The ultrasound imagining is used for providing the radio neutral, non-invasive, low cost and better penetration ability as compared to others like X-ray. The SCLC and NSCLC can be distinguishing with the help of modern ultrasound imagining.

The CT scan is more accurate as compare to X-ray. It can reveal a very small abnormality in infected lungs that might not be revealed by X-ray. The CT scan allows early diagnosis of the abnormal area. The Lung Cancer detection rate is approximately 2.6 to 10 times greater than by using radiography. But the use of CT scan gives direct impact on the workload of radiologist who needs to diagnose and analyzes an increasing number of screen test in very short time. So the errors in detection or misinterpretation take place because of workload. So to avoid these type of situations, the computational systems are needed to assist radiologist in the properly diagnose of tumor. For interpretation of images the nodule detection and determination of their characteristics are needed.

So, for more accurate and better results, the quality of image is improved by using CAD system so that the ultrasonic images of carcinoma can be easily diagnosed by radiologist. To be a good system, CAD system must have the following requirements: providing high sensitivity in diagnosis to improve the performance of radiologists, lesser number of False Positive, providing higher level of automation, achieving high processing speed, cheaper (in implementation, training, support and maintenance), ability to detect different size, shape and types of lesion and assurance of software security.

In CAD system, to help the radiologists in diagnosis of cancer, there are two main types of computational systems: Computer Aided Detection system (CADe) and Computer Aided Diagnosis System (CADx). The purpose of CADe system is to detection of lesion through medical images while CADx system is used to measure the characteristics of the lesion. The main characteristics are to determining the malignancy and staging of the cancer. CADe system must achieve the following goals:

1. Enhance accuracy in diagnosis.

2. Abetment in early detection of tumor.
3. Reduce the complexity and time of the radiologist in exam evaluation.

CADe system is very useful and important tool of medical technology; however, many systems don't yet have all the fundamental requirements to be considered helpful by most radiologists. The fundamentals that are required by radiologist are:

- By providing high sensitivity in diagnosis the performance of radiologist is improved.
- By providing high accuracy in diagnosis, the performance of radiologist is improved.
- The lower number of False Positives (FP). The FP represents that when system gives the significance of the diseases but sample showed no diseases. FP increases the reading time and can result in errors in detection and diagnosis[10].
- High speed of processing. The time taken to respond the request of detection for the system.
- The presence of high level automation to avoiding the occurrence of manual operations. The system should receive DICOM files automatically, then start a processing and store results in standardized report.
- The low cost of maintenance, training, support and implementation.
- Detection of different types, shapes and size of nodules.
- Detect solitary nodules, small nodules which are less than 3 mm in size, ground glass opacity nodules and cavity nodules.
- Assurance of software security for avoiding potential harms that can be result because of loss, inaccuracy, alteration, incorrect use of data, unavailability etc.

The average detection rate of nodules is increases to 84% from 77% at initial evaluation with the aid of the CADe system. The observation shows that the detection of nodules with the use of these tools improves radiologist's performance. Even through the CADe systems are proven to enhance the efficiency and reduce the time of radiologists in the diagnosis and detection of nodules, although they are not commonly used in clinical

practice. That's why the CAD system become one of the most interesting and important areas of research in medical processing.

As we proceed further we will see in the upcoming chapters about the work that has been done. In chapter 2 : Literature Review has been done of various authors using different methods for detection of Lung Cancer. In chapter 3 : all the methods, classifiers were discussed regarding the work done for this project. Chapter 4 : The CAD system for non small cell Lung cancer using laws' mask analysis has been discussed along with its work done in detail showing the results. In this chapter the Laws' mask of dimension 3 gives best results. For chapter 5: CAD for two class classification of Lung Cancer using statistical features has been discussed showing all the obtained results. Out of all the various features, GLDS gives desired results, followed by conclusion and references.

CHAPTER 2

LITERATURE

REVIEW

Name of Paper	Year	Method	Classification	Result
Lung cancer detection on CT scan images: A review on the analysis techniques.	Jan-16 [38]	Computer Aided Diagnosis (CAD) and Neural Network	KNN, SVM, LDA	Overviewed the current detection techniques.
Computerized system for lung module detection in CT scan image by using MATLAB.	2016 [44]	Connected Component Labeling Algorithm (CCL)	SVM. ANN	Discriminate benign modules from malignant modules.
Detection and classification of lung cancer using artificial neural network.	2016 [30]	CAD	ANN	Color feature and texture features are used to compare in ANN.
Techniques for lung cancer module detection: A survey	Dec-15 [4]	1. Pre-processing 2.Segmentation 3. Feature extraction and identification	Minimum Distance classifier, Cascade Classifier, RBF SVM ANN	SVM= 92.67% ANN= 85.27%
CAD system for early lung cancer detection	Nov-15 [31]	CAD, receiver operating characteristics (ROC)	Neural network, SVM K-NN	Achieve good results using classifiers.

Name Of Paper	Year	Method	Classification	Result
A review of techniques for lung cancer detection	May-15 [1]	Gabor filters within Gaussian rules.	ANN	ANN= 85.14%
Lung cancer detection from CT image using image processing techniques.	May-15 [6]	1. capture of input image 2. Pre-processing 3. Feature extraction and identification	SVM	Find better segmentation algorithm which produce good result.
Lung cancer detection using circular hough transforms techniques.	Mar-15 [41]	Circular Hough transform Techniques, CAD, sobel method	KNN,SVM	Efficient in terms of cost and time consumption.
Identifying lung cancer in its early stage using neural network & GA algorithm	Feb-15 [39]	1-Pre-processing 2-feature extraction 3-Cell identification	Genetic Algorithm Neural network	NN and GA give accurate results.
Comparative study review of lung cancer detection using neural network and clustering algorithm.	Feb-15 [20]		Neural network & Fuzzy C mean clustering algorithm	Neural network is better as compare to Fuzzy C mean clustering algorithm.
Review on lung cancer detection using image processing techniques.	Feb-15 [35]	CAD & DFT	Shape, size or appearance based classification	Give overview of different detection techniques.

Name Of Paper	Year	Method	Classification	Result
Image enhancement to detect the lung cancer at early stage using median filter.	2015 [42]	1. Image enhancement 2.Image Segmentation 3. Feature Extraction	ANN,SVM	Median filter is an effective way of detecting lung cancer.
Detection of lung cancer stages on CT scan images by using various image processing techniques.	Oct-14 [16]	1-Pre-processing 2-feature extraction 3-Cell identification	SVM	The result shows good potential for lung cancer detection in early stages
Feature Extraction and principal component analysis for lung cancer detection in CT scan image.	Mar-13 [21]	Principal Component Analysis (PCA) Histogram equalization	ANN SSVM	It gives very promising results compare with other used techniques.
Early detection of lung cancer using neural network technique.	Aug-15 [27]	Remove white Gaussian noise from CT scan image by using non local mean filter.	1. ANN 2. k-NN	ANN = 92.68% k-NN = 85.37%
Detection and classification of cancer from Microscopic biopsy image using clinically significant and biologically interpology feature.	Jul-15 [2]	1.Contain limited adoptive histogram equalization approach 2.K-mean segmentation 3.k-nearest neighborhood method	1-KNN 2-Fuzzy KNN 3-SVM	KNN is best. SVM is Average. Fuzzy KNN is comparatively a less good classifier.
Pre-processing methods for module detection in lung CT.	2011 [34]	CAD in low dose lung multi slice CT.	SVM	Studied pre processing methods for Lung CT.

Name Of Paper	Year	Method	Classification	Result
Multistage approach for biomedical image processing: a review	Mar-13 [8]	Multistage CAD, Ground-Glass opacity (GGO)	SVM, ANN, MTANN	SVM= 79.20%
A segmented Morphological approach to detect tumor in lung images.	Sep-11 [33]	FFT, Histogram modeling, Gabor transform	SVM	The quality of image is increased.
A novel approach for the detection of small cell lung cancer based on entropy and PSNR value.	2015 [43]	1. Image enhancement 2.Image Segmentation 3. Feature Extraction	SSVM	Gabor Filter gives better enhancement but Marker watershed algorithm is the best.

In our literature review, after studying and analyzing many publish papers and journals, we found that for classification of lung cancer there are various classifiers used. Some of them are ANN, k-NN, PNN, SVM, SSVM etc. Every classifier gives different result by using various filters. In July 2015, a journal "Detection and classification of cancer from Microscopic biopsy image using clinically significant and biologically interpology feature" showed that k-NN is best classifier and SVM is average one. But in Dec 2015 a journal "Techniques for lung cancer module detection: A survey" gives 92.67% of accurate result by using SVM classifier.

After the Analysis of many papers and journals, we got an idea that SVM is the best classifier with highest results accuracy of 92.67%. After the thorough reading of papers we found out that SVM uses number of filters for different type of cancer to obtain best results.

CHAPTER 3

METHODOLOGY

The system follows the following steps for the processing of data in order to extract results from its input data. The flowchart is developed to give the basic methodology of classification of Lung Cancer. The Fig 2.1 shows flowchart:

```
┌─────────────────────────┐
│     Data Collection     │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│     Data Selection      │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│    Feature Extraction   │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│    Data Partitioning    │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│   Classification Module │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│     Decision of CAD     │
└─────────────────────────┘
```
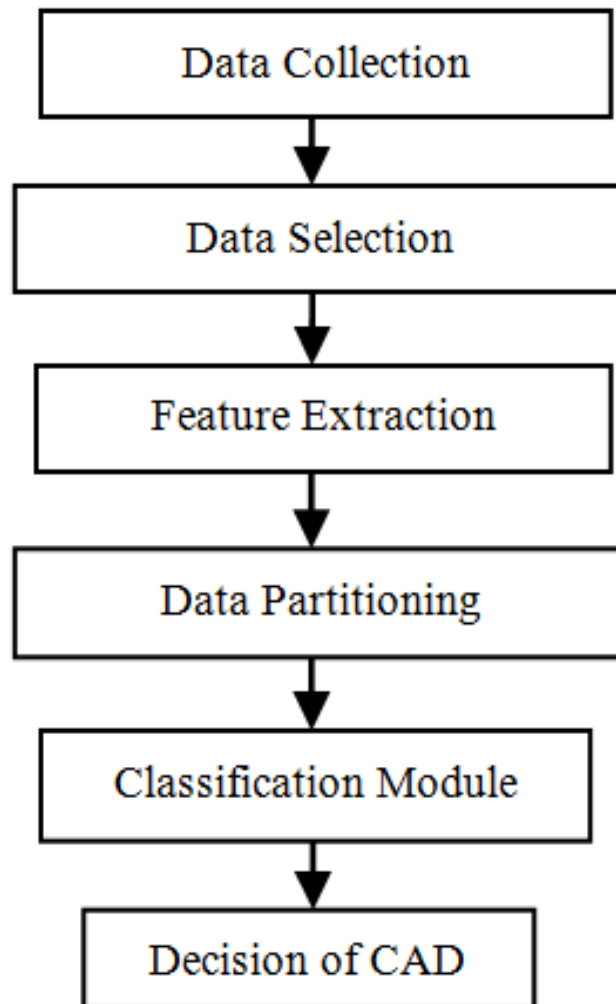
Fig 2.1: Flowchart of methodology

The CAD system follows the steps shown in flowchart. These steps are:

- Data Collection
- Data Selection
- Feature Extraction
- Data Partitioning
- Classification Module

17

- CAD system decision

1. Data Collection:

For CAD system the data collection is basically to collect the input for the CAD system to implement different operations. It can support different types of format as input. So the selection of input format is known as data collection. The images can be used as input for this system: X-ray images, CT scan images, Ultrasound images, Mammography images, MRI images etc. The selection of the input images is based on the type of cancer which we want to detect.

2. Data Selection:

The suitable data type for CAD system is selected from the collected data. Data Selection represents the selection of informative data from the collected data[11-12]. The radiologist helps in selecting the useful and informative data. The input images are studied and marked with the help of expert radiologist. Then the marked region is selected and extracted using various software; MATLAB, ImageJ etc.

MATLAB: It is a multi-paradigm numerical computing environment. It is also fourth generation language for programming. MALAB works on matrix manipulations, data and function plotting, algorithm implementation, user interface creation and interfacing with other language programs (C, C++, C#, Java, FORTRAN and Python). The MATLAB scripting language is the base of MATLAB application. Common window as an interactive mathematical shell or executing text files containing MATLAB code are common usage of the MATLAB application.

ImageJ: It is a public domain, java based image processing program. This software can display, edit, analyze, process, save and print 8-bit color and grayscale, 16 bit integer and 32 bit floating point images. It can read various file formats (TIFF, PNG, GIF, JPEG, BMP, DICOM and FITS) as well as raw formats. It also supports image stacks, a series of images that share a single window and it is multithreaded, so time consuming operation can be performed in parallel on multi CPU hardware. This software can calculate area and pixel value statistics of user defined selection and

intensity threshold objects. ImageJ can gives measurements of distance and angles. By using this, the density histograms and line profile plots. The logical and arithmetical operation between images, contrast manipulation, convolution, fourier analysis, sharpening, smoothing, edge detection and median filtering can be also supported by ImageJ software. It can support any number of images simultaneously.

3. Feature Extraction:

The feature extraction techniques are used to extract basic feature (spot detection, ripple detection, edge detection, wave detection etc.) form the normal as well as gray level images. The feature extraction is the process used to transform the visually extractable and non-extractable features into mathematical descriptors. These descriptors are shape-based (morphological features) and the intensity distribution based (textural features)[13-16].The feature extraction is of two types:

- Morphological Features
- Texture Features

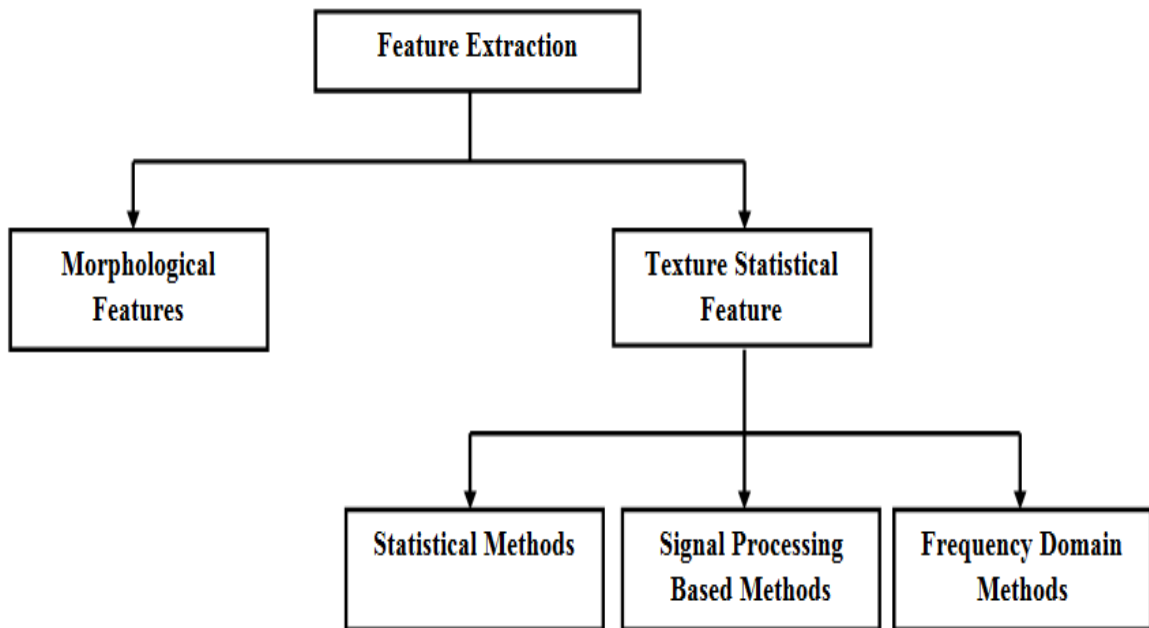These morphological and texture features can also further divided into sub groups as shown in Fig 2.2:

Fig 2.2: Feature extraction flowchart

The flowchart show that Texture features can be categorized in three parts:

1. **Statistical features**
2. **Signal processing based features**
3. **Transform domain features**

Statistical Features:

 The methods used for extraction of texture features based on pixel's gray level intensities of an ultrasound image, is known as statistical methods. The statistical methods can be classified on the bases of the number of pixels used to calculate the texture features [17]. The statistical features are divided into three types: First Order Statistics (FOS), Second Order Statistics and Higher Order Statistics.

- First Order Statistics (FOS)
- Gray Level Co-occurrence Matrix (GLCM or 2nd Order Statistics)
- Gray Level Run Length Matrix (GLRLM or Higher Order Statistics)
- Edge Feature
- Neighborhood Gray Tone Difference Matrix (NGTDM)
- Statistical Feature Matrix (SFM)
- Gray Level Differences Statistics (GLDS)

1. *FOS Feature:* The FOS is derived from the gray level intensities histograms of the image. The features like standard deviation, kurtosis, average gray level, smoothness, and entropy were calculated for every ROI.

2. *Second Order Statistics Feature:* In second order statistics, there are calculations with the Gray Level Co-occurrence Matrix (GLCM). In GLCM the different combinations of pixel gray level occurs in an image which are used for second order statistics calculation. The relationship between groups of 2 pixels is considered in GLCM which are called reference and neighbor pixel in different directions (0°, 45°, 90°, 135°). These pixel pair having different gray level occurring in image having separation of different dimensions in different directions. The GLCM includes

features of calculating contrast, entropy, sum entropy, difference entropy, correlation, inverse difference moment, sum average, variance, sum variance, difference variance angular moment, information measures of correlation-1 and 2, angular second moment.

3. *Higher Order Statistics-GLRLM Features:* Higher order statistics are computed with the use of GLRLM (Gray Level Run Length Matrix) [18].Texture features are computed using the different combinations of intensities at relative position of each other. Gray level run is made by the set of consecutive pixels of gray levels that are collinear to each other and run length denotes the no of times a run occurs. The 11 GLRLM features that are computed in this work are long run emphasis, short run emphasis, low gray level run emphasis, high gray level run emphasis, run length non-uniformity and run percentage, short run low gray level emphasis, long run low gray level emphasis, gray level non uniformity, short run high gray level emphasis, long run high gray level emphasis.

4. *Other Statistical Features:* Other statistical features like Edge Features, Neighborhood Gray Tone Difference Matrix (NGTDM), Statistical Feature Matrix (SFM), and Gray Level Difference Statistics (GLDS) were calculated in this paper.

   a) **Edge Features***:* The information present in the edges is higher as compared to the other parts of the image. The spatial variation in an image is calculated by the gradient value. If there is an instant change in pixels then gradient will be high else it will be low. It calculates two features: absolute gradient and absolute gradient variance.

   b) **NGTDM**: In NGTDM, the calculation of busyness, coarseness, complexity, contrast, strength takes place. In this it considers a difference between the gray levels between pixels.

   c) **SFM:** It calculates contrast, coarseness, periodicity and roughness of pixels at different distance within an image.

d) **GLDS:** The contrast, energy, entropy, homogeneity and mean were calculated by GLDS on the basis of the co-occurrence of the pixel pair that has difference in gray levels separated by particular distance.

Signal Processing Based Features: For Signal Processing Based method, the method is used for feature extraction is known as Laws Mask analysis. For implementation of this method the Laws mask filter is applied to database.

- **Laws' Mask Texture Analysis:** the small convolution mask is used as filter in this method. The ROI's are convolved with these filters so that the texture feature's characteristics are extracted[19-21]. By performing averaging, edge detection, spot detection and wave detection, the properties of texture are determined using these filters. The Laws mask having different lengths (3,5,7,9) are used to calculate five parameters; mean, standard deviation, skewness, kurtosis and entropy form ROI's.

Transform Domain Features: The transform domain is also used for feature extraction over various scales by using various multi resolution schemes. It is logical to compute texture features in the transform domain as human visual system processes images in a multi scale way and scale is considered to be a necessary aspect to analyze texture of image.

These schemes of transform domain are:

- Wavelet Packet Transform (WPT)
- Gabor Wavelet Transform (GWT)
- Fourier Power Spectrum (FPS)

WPT: The WPT is any wavelet transform for which the wavelets are sampled in packets. Like other wavelet transforms, it has a key advantage over the Fourier transform that is known as temporal resolution. In this temporal resolution, it captures both frequency and location information. It has huge number of applications in science, engineering, mathematics and computer science. It is mainly used for signal coding, to represent a signal in a more superfluous form, often as pre-conditioning for

data compression. Practical applications can also be found in signal processing for gait analysis, in digital communications and many others.

GWT: A mathematical too for analyze the mathematical functions and found an optimize solution the Gabor wavelets are widely used. It is used for signal analysis and has found widespread use in computer vision. The applications of GWT are widely used in object tracking, edges detection, extract feature, texture analysis and faces classification. Its properties support the use of gabor wavelets. GWT also have applications in mathematics engineering, biomedical, computer science, radiologically and medicine inspired early vision systems. Gabor wavelets transformation supplied to solve many application problems (image compression, filter design and edge detection).

FPS Features: In FPS feature extraction method, two spectral features are computed from each ROI using Discrete Fourier Transform. These spectral features are:

- Radial Sum
- Angular Sum

4. **Data Partitioning:** It is a process in which the data is partitioned into two sets; testing and training. The techniques used in data partitioning:

- Hold Out
- K-fold cross validation

Hold Out: In this technique the dataset is partitioned in two sets. These sets are known as training and testing sets of data. It uses the testing score as a generalization measures. Hold out technique gives better approximation of data.

K-fold Cross validation: In k-fold cross validation process, the data is train and test at every point of the dataset. It is expensive as compared to hold out techniques.

5. **Classification Module:** The process of grouping the testing samples into the corresponding classes is known as classification of dataset. It is categorized into two types:

- Supervised Classification
- Unsupervised Classification

Supervised Classification: when the classes are already defined for the training sets then the classification is known as supervised classification [22].

Unsupervised Classification: In unsupervised Classification techniques, the classes are not defined to training sets, there are undefined classes.

For classification module, different classifiers are employed to classify the unknown testing instances of various ultrasonic classes based on the training instances. The main classifiers used for classification is k-NN, PNN, SVM, SSVM and ANN etc. In order to avoid any bias by unbalanced features, Min-Max normalization procedure is used to normalize the extracted features.

k-Nearest Neighbor: It is based on the idea of estimating the class of an unknown instance form its neighbors. The basic principle behind k-NN is the assumption of the feature vector lying close to each other belongs to the same class. Because of it tries to group the instances of feature vector into same classes lying close to each other. In the training dataset, by looking among k-nearest the class of an unknown instance is selected. The main advantage k-NN is its ability to handle multiple class problems and it is also robust to the noisy data problem because it averages the k-nearest neighbors[22-24]. The various distance metrics can be used to calculate say Euclidean distance, Cosine distance, city block, Minkowski, Chebychev and correlation. The Euclidean distance is used as distance metric in this classification module. The value of k is key factor in k-NN classifier as classification performance of k-NN is depends on the value of k. the Example for the classification of the unknown instance is shown in Fig 2.3:
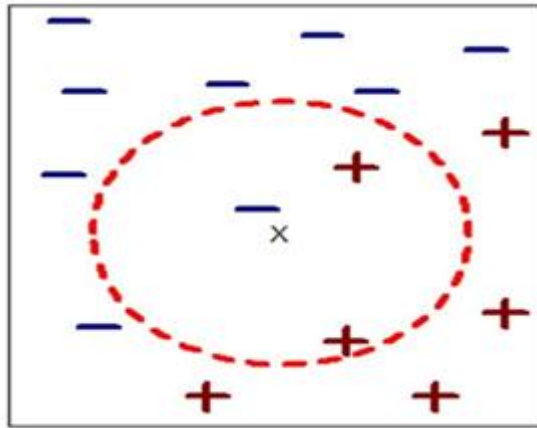
Fig 2.3: Example of k-NN for k = 3

In the given example the test sample (×) is unknown instance. It should be either classified to the class of dash (-) or to the class of cross (+). When $k = 3$, the algorithm looks for three nearest neighbors.

Probabilistic Neural Network: The PNN is a neural network which is commonly used for pattern recognition and classification problems. The basic principle for PNN is probability distribution function (PDF). The PDF od each class is estimated by Parzen window and a non-parametric function. Then, by using PDF of each class, the probability is unknown instance is estimated[25-29]. Then Bayes' rule is applied to allocate the class with highest posterior probability to unknown instance. By using this method, the mis-classification probability can be minimized. It generates accurate estimated target probability scores. It also supports remote sensing classification.

Support Vector Machine: The SVM classifier comes under the class of supervised learning machine and works on the basis of statistical theory. SVM classifier can classify both linear and non-linear classification. With the help of the available training data it creates the hyper plane between the classes which results in good separation achievement intuitively but the sets that are available to discriminate are not linearly separable in the space. In non linear classification module, the data is mapped from input space to higher dimensional feature space by using the input data which is mapped into the kernel function[30]. The Gaussian radial basis function has been used for classification of the data. Present algorithms occurs the sub gradient and coordinate descent methods that

have a big advantage of having large and sparse datasets. For implementation of the SVM classifier for classification the library is used, named as LibSVM.

Smooth Support Vector Machine: SSVM is the advanced and improved version of SVM classifier. The purpose of SSVM classifier is to obtained unconstraint smoothing and unconstrained optimization reformulation which could not obtained by using SVM classifier as it associated with the traditional quadratic program. SSVM works on the concept of smoothing unconstrained optimization reformulation the problem related to SVM classifier for the pattern classification[31-32]. The toolbox of SSVM has been designed by Laboratory of Data Science and Machine Intelligence, Taiwan for implementation of SSVM. In case of SSVM implementation, it is similar to SVM implementation. In implementation, the ten-fold cross validation is carried out on training data for each combination. The optimum value of C is given by the procedure of grid search in parameter space for which training accuracy is maximum.

Artificial Neural Network: The ANN are the computational model which are based on the large collection of simple neural network units. Each neural unit is connected with many other units and links can enhance the activation state of adjoining neural units. By using summation function, each neural unit can be computed. The ANN is combination of many artificial neurons that are linked together according to a specific architecture of network. Its prime goal is to transform inputs into meaningful outputs. The ANN is used to achieve the goal to controlling the movement of a robot based on self- perception and other information. These systems are self learning and trained and work in the areas where the feature detection is difficult to extract.

CHAPTER 4

CAD SYSTEM FOR NON SMALL CELL LUNG CARCINOMA USING LAWS' MASK ANALYSIS

In CAD system, there are various techniques used for detection and diagnosis of lung cancer. This system is based on the analysis of quality of input images (ultrasound images) by detecting lesions that may indicate the presence of abnormal lung tissues. In CAD system there are four main steps: preprocessing, feature selection, feature extraction and feature classification[33-37]. After preprocessing and feature selection of ultrasonic images, the Laws' mask analysis is used for texture feature extraction. The SVM classifier is used for classification purpose in this work. Classifiers are used to allocate the cells into benign and malignant. The working of system is based on different norm such as Specificity, Sensitivity and accuracy. The results of classifiers are used for CAD system for early analysis of lung cancer to improve the survival rate. The work sequence of this work is shown in figure 4.1:
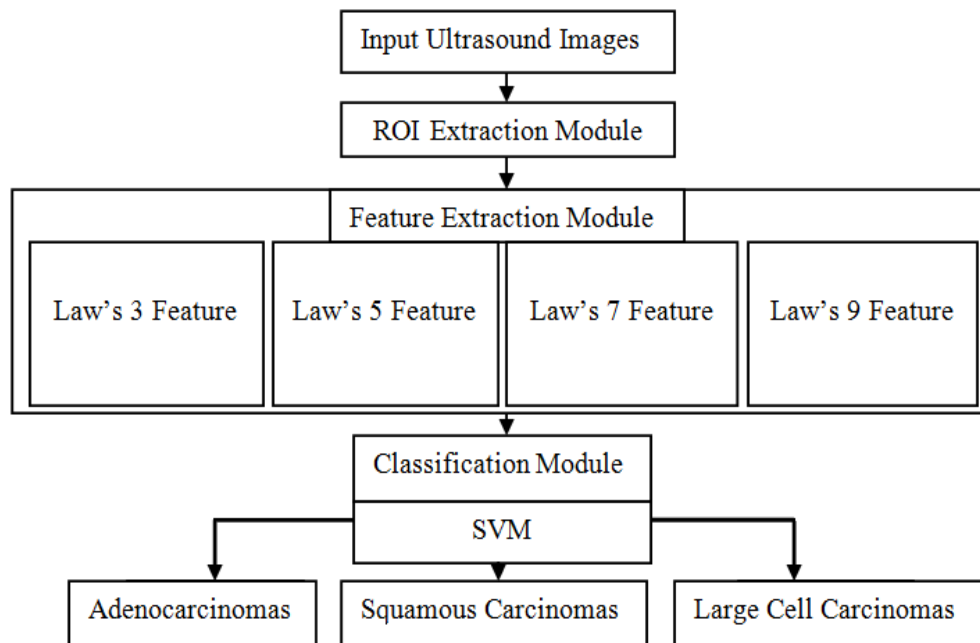
Fig 4.1: Overview of work

Database Ultrasound Images: The data is collected in the form of ultrasound images. For this work the data of 92 cases of NSCLC is taken and discarded the cases of biopsy and cases having blood vessels. The figure 4.2 represents the input ultrasound image:
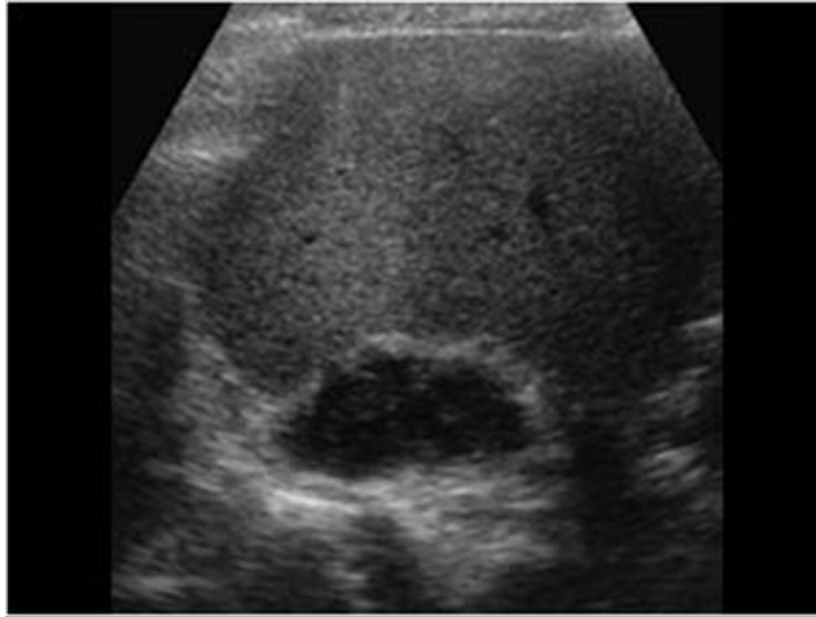
Fig 4.2: Input ultrasound image [15]

ROI (Region of Interest) Extraction Module: The abnormal areas in the ultrasonic images are detected and mark with the help of expert radiologist. The marked areas are then segmented using *ImageJ* software. This software helps to load the ultrasound image, mark the suspected area and segment it in ROI's. Further the ROI is enclosed into rectangular box adjoining the boundaries of abnormalities. The Fig 4.3 represents the ROI extracted from ultrasound image shown in figure 4.2:
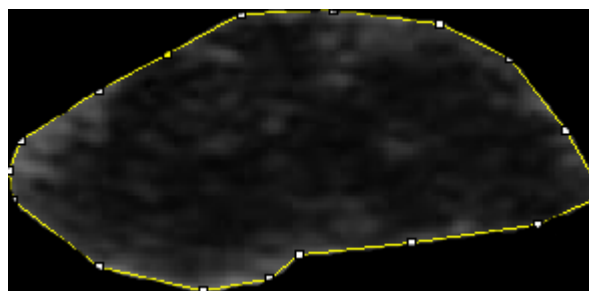
Fig 4.3: ROI of ultrasound images[49]

Feature Extraction Module: The lesion is that abnormal or infected area which has been damaged by any disease. The feature extraction module for this work contains Texture Features which are classified as:

- Signal Processing Based Methods

- Statistical Methods

- Transform Domain Methods

In this work the lesions are categorized with the use of Signal Processing Based methods.

Signal Processing Based methods: A set of coherent masks are used in Laws' based texture features. In texture feature extraction the ROI images are convolve with these masks. The filters are made up by the combination of two or more one dimension kernel vectors. These vector can be of dimensions d =3,5,7,9. The texture features are determined by performing Edge detection, Level Detection, Ripple Detection, Spot Detection and Wave Detection by using filters of different dimensions.

- **Laws' mask d=3:** The resolution of dimension 3 vector is:

    $L3$ is 1, 2, 1; $E3$ is -1, 0, 1 and $S3$ is -1, 2, -1.

The masks of different dimensions are used to extract the different features. The mask of dimension d= 3 is shown in figure 4.4:

L3L3	E3L3	S3L3
L3E3	E3E3	S3E3
L3S3	E3S3	S3S3

Fig 4.4: Mask of dimension 3.

In Laws' mask of dimension 3, there are 9 two dimensional masks which includes 3 masks of identical pairs, so rotational invariant texture images will be 6. There are 5 descriptors derived from ROI, so feature vector length is 6×5=30. The algorithm followed by Laws' mask 3 is given below:

  Step 1: Collection of dataset.

  Step 2: Data preprocessing was done by selection of ROI's

  Step 3: Import ROI images.

  Step 4: Conversion of ROI to gray level images.

Step 5: Set the mask of dimension 3 as shown in Fig 4.4.

Step 6: Convolve the mask with gray level images.

Step 7: Calculation of features (Contrast, Texture Image, Energy).

- **Laws' mask d=5:** For dimension 5 vector the resolution is:

$L5$ is 1, 4,6,4,1; $E5$ is -1,-2,0,2,1; $S5$ is -1,0,2,0,-1; $W5$ is -1,2,0,-2,1 and $R5$ is 1,-4,6,-4,1

For Laws' mask of dimension 5 is shown in figure 4.5:

L5L5	E5L5	S5L5	R5L5	W5L5
L5E5	E5E5	S5E5	R5E5	W5E5
L5S5	E5S5	S5S5	R5S5	W5S5
L5R5	E5R5	S5R5	R5R5	W5R5
L5W5	E5W5	S5W5	R5W5	W5W5

Fig 4.5: Mask of Dimension 5.

In Laws' mask of Dimension 5, there are 25 two dimensional masks in which rotational invariant texture images of identical paring are 10[38-41]. The total invariant texture images will be 15. There are 5 descriptors which are derived from ROI, so the numbers of features extracted from Laws' mask of dimension 5 are 15×5=75. The algorithm of dimension 5 is given below:

Step 1: Collection of dataset.

Step 2: Data preprocessing was done by selection of ROI's

Step 3: Import ROI images.

Step 4: Conversion of ROI to gray level images.

Step 5: Set up mask for dimension 5 as shown in Fig 4.5.

Step 6: Convolve mask with gray level images.

Step 7: Detect edges, ripples, spots, level and wave of images.

Step 8: Compute the features (Contrast, Texture Image, Energy).

- **Laws' mask d=7:** The Laws' mask filter of dimension 7 will be of 3×3 dimension. The filter with dimension 7 will have same feature vector length as dimension 3. The algorithm is given below:

    Step 1: Collection of dataset.

    Step 2: Data preprocessing was done by selection of ROI's

    Step 3: Import ROI images.

    Step 4: Conversion of ROI to gray level images.

    Step 5: Set up mask for dimension 7.

    Step 6: Convolve mask with gray level images.

    Step 7: Detect edges, ripples, spots, level and wave of images.

    Step 8: Calculate the features (Contrast, Texture Image, Energy) of images.

- **Laws' mask d=9:** The filter with dimension 9 will have same feature vector length as dimension 5. The Laws' mask of dimension 9 will be of 5×5 dimension. The algorithm is given below:

    Step 1: Collection of dataset.

    Step 2: Data preprocessing was done by selection of ROI's

    Step 3: Import ROI images.

    Step 4: Conversion of ROI to gray level images.

    Step 5: Set up mask for dimension 9.

    Step 6: Convolve mask with gray level images.

    Step 7: Detect edges, ripples, spots, level and wave of images.

    Step 8: Compute the features (Contrast, Texture Image, Energy).

There is a sequence to analyze the Laws' mask. The sequence for Laws' mask of dimension 3,7 and 9 are also same as given in eq (1-4). The only difference is to change the value of Mask's dimension.

The sequence for Laws' mask of dimension 5 is:

a) The Texture Image (*TI*) is obtained by convolving the 2-D mask with input image I (i, j)

$$TI_{E5E5} = I_{i,j} \otimes E5E5 \tag{1}$$

b) The contrast of the texture image obtained from equation (1) which is normalized

$$Normalize\ (TI_{mask}) = \frac{TI_{mask}}{TI_{L5L5}} \tag{2}$$

c) The Texture Energy Measurement (TEM) filters are used to pass the Texture image

$$TEM_{i,j} = \sum_{u=-5}^{5} \sum_{v=-5}^{5} Normalize(TI_{i+u,j+v}) \tag{3}$$

d) To obtain 15 rotationally invariant TEM's that are denoted as *TR* are obtained by collaborating the 25 TEM descriptors

$$TR_{E5L5} = \frac{TEM_{E5L5} + TEM_{L5E5}}{2} \tag{4}$$

e) Five statistical parameters are determined, that are the Mean, Standard Deviation, Skewness, Kurtosis, Entropy. Here M×N is the dimension of the image.

1. *Mean (m):* It describes the mean intensity value with in texture image.

$$Mean = \frac{\sum_{i=0}^{M} \sum_{j=0}^{N} (TR_{i,j})}{M \times N} \tag{5}$$

2. *Standard Deviation (SD):* It is used to measure the variability.

$$SD = \sqrt{\frac{\sum_{i=0}^{M} \sum_{j=0}^{N} (TR_{i,j} - Mean)^2}{M \times N}} \tag{6}$$

3. *Skewness:* It measures of the asymmetry of the probability distribution of a

random variable that is real valued.

$$Skewness = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N}(TR_{i,j}- Mean)^3}{M \times N \times SD} \qquad (7)$$

4. *Kurtosis:* It measures of the probability distribution's shape of a random variable that is real valued.

$$Kurtosis = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N}(TR_{i,j}- Mean)^4}{M \times N \times SD^4} \qquad (8)$$

5. *Entropy:* It measures the randomness of the elements of the image.

$$Entropy = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N}(TR_{i,j})^2}{M \times N} \qquad (9)$$

Classification Module: There are two types of characterizations of classification in classification module.

- Supervised classification
- Unsupervised classification

Supervised classification is the one for which the classification classes are defined for training sets. But in unsupervised classification the classes are not defined for training sets. In this work the supervised classification are obtained. The classifier used here is SVM classifier. For implementation of SVM classifier, there is a library named as LibSVM is used. It works on both linear and non-linear classifications. SVM works on basic approach of decision plane.

Results: The results are obtained by using Laws' mask of different dimensions (3,5,7,9) of signal processing based method. The result of Laws' mask using SVM classifier is shown in table 4.1:

Table 4.1: Laws' mask of dimension 3					
TFV(*l*)	**CON**			**Sensitivity**	**OCA**
	AC	**SC**	**LC**		
TFV(30) **AC**	16	0	0	100%%	
SC	1	14	0	93.3%	95.65%
LC	1	0	14	93.3%	

Results of Laws' mask of dimension 5 for classification using SVM classifier are shown in table 4.2:

Table 4.2: Laws' mask of dimension 5					
TFV (*l*)	**CON**			**Sensitivity**	**OCA**
	AC	**SC**	**LC**		
TFV(75) **AC**	15	0	1	93.8%	
SC	2	12	1	80%	89.13%
LC	1	0	14	93.3%	

Results of Laws' mask of dimension 7 for classification using SVM classifier are shown in table 4.3:

Table 4.3: Laws' mask of dimension 7					
TFV(*l*)	**CON**			**Sensitivity**	**OCA**
	AC	**SCM**	**LC**		
TFV(30) **AC**	14	0	2	87.5%	
SC	3	12	0	80%	84.78%
LC	2	0	13	86.6%	

Results of Laws' mask of dimension 9 for classification using SVM classifier are shown in table 4.4:

Table 4.4: Laws' mask of dimension 7					
TFV(*l*)	**CON**			**Sensitivity**	**OCA**
	AC	**SC**	**LC**		
TFV(75) **AC**	14	0	2	87.5%	
SC	2	12	1	80.%	82.6%
LC	2	1	12	80%	

AC: Adenocarcinomas, SC: Squamous Carcinomas, LC: Large cell Carcinomas, TFV: Texture Feature Vector, CON: Confusion Matrix, OCA: Overall Classification Accuracy, Sensitivity, *l*: length of TFV

Different researchers have investigated using different methods, so for better observation of work, a comparison is displayed in Table 4.5.

Table 4.5: Comparison of our work with existing papers			
Author	**Classifier**	**Accuracy**	**Feature**
Land WH, Jr. *et. al*	SVM	72%	Morphological Features (seeded watershed transforms)
Pawar PD *et. al*	SVM	79.2%	Morphological Features
Gujral H *et.al*	FFT	85.1%	Texture Features (Gabor)
Proposed work	SVM	95.65%	Texture Features (Laws' mask)

The highest accuracy achieved in this work is 95.66% by using Laws' mask of dimension 3. The classifications obtained were from SVM classifier. The results by using SVM classifier for Laws' mask of dimension 5 were 89.13%. We got 84.78% accurate results for Laws' mask of dimension 7 whereas for dimension 9 the results were 82.6%.

Conclusion and Future Work: The highest result obtained is 95.67% and is coming out by Laws' mask of dimension 3. There are some results which are misclassified; result of laws' mask of dimension 9. So, in future authors will try to improve the results of Laws' mask of dimension 9 and also focus to improve the overall performance of the system. To achieve higher accuracy in this work, the author will use various filters and classifiers for better results.

CHAPTER 5

CAD FOR TWO CLASS CLASSIFICATION OF LUNG CANCER USING STATISTICAL FEATURES

A large number of techniques are being used for detection and diagnosis of Lung Cancer. The Computer Aided Diagnosis (CAD) is the most common and accurate technique for early detection of abnormal cells which can cause cancer to healthy lung tissues. CAD system works on the basis of analysis of condition of ultrasound images. CAD system follows different steps: Data collection (ultrasound image), Data Preprocessing (ROI Selection), Feature Extraction, Data Partitioning (hold- out method), Feature Classification and Result Calculation. To classify input ultrasound images into benign and malignant, different classifiers were used. The system work is based on the calculation of parameters such as individual accuracy, overall accuracy and sensitivity. These benchmarks are obtained by calculating the matrix of Support Vector Machine (SVM).

The results were obtained by using various features using Statistical Methods. The best results achieved were having accuracy of 91.4% by using Gray Level Difference Statistics (GLDS). CAD system will use these results for detection of Lung Cancer cells in initial stage to enhance the capability of survival of patient. In this research work, the results obtained clearly demonstrate a promising accuracy and sensitivity of classes of lung cancer.

The system follows the following steps for the processing of data in order to extract results from its input data. The algorithm is developed to give the basic methodology of classification of Lung Cancer.

Algorithm:

Step 1: *Data Collection*: Create a database of ultrasound images for Lung Cancer.

Step 2: *Data Preprocessing*: Selection of Region of Interest (ROI).

Step 3: Convert the selected ROI to Gray level images.

Step 4: *Feature Extraction*: Extract the texture features from gray level images using statistical methods.

Step 5: *Data Partitioning*: Partitioning the dataset into two sets (Training and Testing) using hold-out partitioning method. Hold-out data Partitioning is to reserve half

data for training and other half for testing or to reserve 1/3 for testing and 2/3 for training.

Step 6: *Classification*: Apply SVM to partition database on the bases of classes of cancer.

Step 7: Calculate the Sensitivity and Accuracy of the data.

Database Ultrasound Images: The data is taken for analyzing the lung carcinoma whether the cancer is of SCLC type or NSCLC type. The data is taken of 92 cases and abandon the cases having blood crafts. The database of 92 cases, out of which total 11 cases of SCLC and 81 cases of NSCLC are taken for this work.

ROI Extraction Module: The deformity in the ultrasound images are exposed with the help of expert radiologist and then marked using MATLAB 14.7 software. In order to mark these areas, the x, y coordinates of the interested regions were calculated. Further segmented areas of dimensions 32×32 pixels were obtained. The dimensions of the ROI pixels may vary. The ultrasound image of carcinoma is shown in Fig. 5.1 and its segmented ROI is presented in Fig. 5.2:
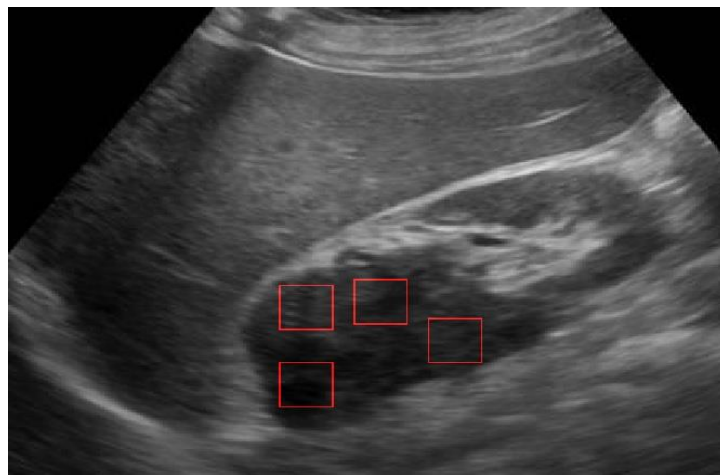

Fig 5.1: Ultrasound Image of Lung Cancer

Fig 5.2: ROI's of ultrasound image

Feature Extraction Module: The infected field which has been destructed by any problem or illness is called lesion. Feature extraction module is divided into Morphological methods, Signal Processing Based Methods, Transform Domain Methods and Statistical Methods. This work is focused on Statistical Methods to extract the Texture features of database.

Statistical Methods: The methods used for extraction of texture features based on pixel's gray level intensities of an ultrasound image, is known as statistical methods. The statistical methods can be classified on the bases of the number of pixels used to calculate the texture features. The statistical features are divided into three types: First Order Statistics (FOS), Second Order Statistics (SOS) and Higher Order Statistics.

1. *FOS Feature:* The FOS is derived from the gray level intensities histograms of the image. The features like standard deviation, kurtosis, average gray level, smoothness, and entropy were calculated for every ROI.

2. *SOS Feature:* In second order statistics, there are calculations with the Gray Level Co-occurrence Matrix (GLCM). In GLCM the different combinations of pixel gray level occurs in an image which are used for second order statistics calculation. The relationship between groups of 2 pixels is considered in GLCM which are called reference and neighbor pixel. These pixel pair having different gray level occurring in image having separation of different dimensions in different directions. The GLCM includes features of calculating contrast, entropy, sum entropy, difference entropy, correlation, inverse difference moment, sum average, variance, sum variance, difference variance angular moment etc. GLCM runs on the bases of following algorithm:

Step 1: Different extracted ROI images (database) were imported.
Step 2: Conversion of database images to gray level images.
Step 3: Assume initial value of co-occurrence matrix as zero.
Step 4: Assign index to co-occurrence matrix.
Step 5: Calculation of co-occurrence matrix.
Step 6: Calculation of pixels (reference pixel and neighbor pixel) of the row

matrix.

Step 7: Normalization (Z Normalization) of data. Normalization in GLCM is to transform the matrix into a close approximation of probability table. The normalization is only an approximation because the gray levels are integer value so they are discrete in nature but probability work on the continuous values.

$$Z \text{ normalization} = \frac{\text{basic value} - \text{mean}}{\text{standard deviation}} \qquad (10)$$

Step 8: Calculation of GLCM features. There are different features of GLCM: entropy, sum entropy, correlation, inverse difference moment, variance, contrast etc. Formulae used for calculation of features are:

1. Entropy is calculated by:
$$\text{Entropy} = -\sum_{i,j} P_{i,j} \log(P_{i,j}) \qquad (11)$$

2. For calculation of Sum Entropy:
$$\text{Sum Entropy} = -\sum_{i=2}^{2Ng} P_{i,j}(i) \ \log(P_{x+y}(i)) \qquad (12)$$

3. For calculating correlation:
$$\text{Correlation} = \sum_{i,j} P_{i,j} \left[\frac{(i - \mu_i)(j - \mu_j)}{\sigma_i \sigma_j} \right] \qquad (13)$$

4. For calculating Inverse Difference Moment:
$$\text{Inverse Difference Moment} = \sum_{i,j} \frac{P_{i,j}}{1 + (i - j)^2} \qquad (14)$$

5. For calculation of variance:
$$\text{Variance} = \sum_{i,j} P_{i,j}(i - \mu_i)^2 \qquad (15)$$

6. For calculation of contrast:
$$\text{Contrast} = \sum_{i,j} P_{i,j}(i - j)^2 \qquad (16)$$

where *i* and *j* are horizontal and vertical coordinates respectively. μ is expected value and *P* is Probability Mass Function (the function that gives the probability that a discrete random variable is exactly equal to some value).

3. *Other Statistical Features:* Other statistical features like Edge Features, Neighborhood Gray Tone Difference Matrix (NGTDM), Statistical Feature Matrix (SFM), and Gray Level Difference Statistics (GLDS) were calculated in this paper.

a) **Edge Features***:* The information present in the edges is higher as compared to the other parts of the image. The spatial variation in an image is calculated by the gradient value. If there is an instant change in pixels then gradient will be high else it will be low. It calculates two features: absolute gradient and absolute gradient variance. The algorithm followed for Edge feature is shown below:

Step 1: Collection of dataset.

Step 2: Data preprocessing was done by selection of ROI's

Step 3: Import ROI images.

Step 4: Conversion of ROI to gray level images.

Step 5: Calculation of horizontal and vertical edges using sobel filter. (Sobel Filter is used to calculate the approximation of the gradient of the image intensity function. It creates an image emphasizing edges in image processing and computer vision).

Step 6: Compute probability distribution of gray level images.

Step 7: Compute mask to selected ROI's.

Step 8: Discard elements which counts pixels outside ROI.

Step 9: Normalization of matrix.

Step 10: Calculation of the features (absolute gradient). For calculation of absolute gradient formula is given in eq 17.

$$\nabla_j A^k \equiv \frac{\partial A^k}{\partial x^j} + A^i \Gamma_{ij}^k \tag{17}$$

where Γ_{ij}^k is Christoffel symbol, *A* is basis vector and *k* is free index.

b) NGTDM: In NGTDM, the calculation of busyness, coarseness, complexity, contrast, strength takes place. In this it considers a difference between the gray levels between pixels. The steps of algorithm are :

Step 1: Collection of dataset.

Step 2: Data preprocessing was done by selection of ROI's

Step 3: Import ROI images.

Step 4: Conversion of ROI to gray level images.

Step 5: Calculation of NGTDM vector by comparing pixels to its neighborhood gray tone values of pixels.

Step 6: Calculation of normalization coefficient.

Step 7: Define neighborhood kernels (Kernel is the small matrix which is useful for blurring, sharpening, edge detecting etc).

Step 8: Convolve the kernels of ones to select pixels for which the kernel lies entirely within ROI.

Step 9: Compute NGTDM matrix by convolving these kernels with gray level images.

Step 8: Calculation of the features (busyness, coarseness, complexity, contrast   (same as in eq.16)).

Formula for calculating image complexity is:

$$\text{Complexity} = \frac{\text{RMSE}}{\text{CR}} \qquad (18)$$

where RMSE = root mean square error between lossy compressed image and original image.

CR = Compression ratio.

c)   SFM: It calculates contrast, coarseness, periodicity and roughness of pixels at different distance within an image. For statistical feature matrix calculation the algorithm is shown below:

Step 1: Collection of dataset.

Step 2: Data preprocessing was done by selection of ROI's

Step 3: Import ROI images.

Step 4: Conversion of ROI to gray level images.

Step 5: Define coordinates of statistical matrix.

Step 4: Shift images by deleting rows and columns.

Step 6: Define ROI in original and shifted images.

Step 7: Common pixels in both areas are considered.

Step 8: Compute statistical feature matrix.

Step 9: Calculation of the features (contrast (same as in eq.16), periodicity and absolute roughness).

Formulae for feature calculation:

$$\text{Roughness} = \frac{1}{n}\sum_{i=1}^{n}|y_i| \qquad (19)$$

where y_i = the vertically distance from i^t data point to the mean line.

n = the order of equally spaced point in the pixels.

d) GLDS: The contrast, energy, entropy, homogeneity and mean were calculated by GLDS on the basis of the co-occurrence of the pixel pair that has difference in gray levels separated by particular distance.

Step 1: Collection of dataset.

Step 2: Data preprocessing was done by selection of ROI's

Step 3: Import ROI images.

Step 4: Conversion of ROI to gray level images.

Step 5: Calculate the probability distribution of gray level differences for different dimensions and angles.

    a. Compute mask to select ROI.

    b. Setup gray level difference vector.

    c. Initialize gray level difference count vector.

    d. Calculate gray level difference.

Step 6: Calculate gray level difference at 0, 45, 90 and 135 degree.

Step 7: Normalize gray level difference matrix for all angles.

Step 8: Calculation of the features (contrast (same as in eq.16), homogeneity, energy, entropy (same as in eq. 11), and mean).

1) The homogeneity is calculated by:

$$\text{Homogenity} = \sum_{i,j} \frac{P_{i,j}}{1+(i-j)^2} \qquad (20)$$

2) Energy is computed by following formula :

$$\text{Energy} = \sqrt{P_{i,j}^2} \qquad (21)$$

3) Mean is calculated using:

$$\text{Mean} = \frac{1}{m} \sum_{i,j} iP_{i,j} \qquad (22)$$

where *i, j* and *P* are same as mentioned for eq. 11 to eq.16.

Classification Modules: There are two type of characterization of classification; Supervised Classification and Unsupervised Classification. When the classes are defined for the training sets in classification, then it is supervised classification and opposite is for unsupervised classification. The classifier used in our work is SVM classifier. The SVM is supervised type of classifier. For implementation of SVM classifier a library is used in MATLAB, called LibSVM. SVM classifier proceeds on the principle of decision planes, where the boundaries of decision are defined. It works on both linear and non-linear classification with higher speed and accuracy. Kernel Based Classifier; the training data of non-linear aligning to higher dimensional feature space from input space has been done using kernel functions. There are different types of kernels. For classification task in this paper we have used the *Gaussian radial basis function*. The appraisal of kernel parameter γ and parameter of regularization C is always a diagnostic step for having desired abstract performance. By doing the expanded search that is carried out in the parameter space for the values of C ϵ {2-4, 2-3... 215}, γ ϵ {2-12, 2-11... 24}, the admirable values of γ and C are obtained.

Results and Discussions: In this paper, we have calculated Sensitivity and Accuracy of Lung Cancer by using confusion matrix. This confusion matrix is basically a table that often describes the performance of classifier using set of test data for which the true values are given shown in Table 5. The confusion matrix is obtained by SVM classifier.

The parameters used for computation of matrix are shown in Table 5.1.

Table 5.1: Confusion matrix for calculation		
	Estimated No	**Estimated Yes**
Real No	TN = 83	FP = 7
Real Yes	FN = 2	TP = 13

True Positive (TP): The true positives are those cases of matrix, when we predict positive result and result is also positive.

True Negative (TN): A true negative occurs in matrix when our predication is negative and outcome is also negative.

False Positive (FP): These are the cases when we predict positive result but result is opposite of it.

False Negative (FN): The predicted value is negative but the result is positive, this is the false negative parameter for confusion matrix.

The formulae for parameters used in calculation of sensitivity and overall classification accuracy are:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{23}$$

Sensitivity is also known as individual classification accuracy or true positive rate.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Number of Cases}} \qquad (24)$$

Accuracy is known as Overall Classification Accuracy (OCA).

The results obtained by calculating the co-occurrence matrix using various features of Statistical Methods and SVM classifier is shown in Table 5.2:

Table 5.2: Two-Class Lung Cancer Classification using SVM						
Features	**Confusion Matrix**			**ICAM**	**ICAB**	**OCA**
		M	**B**			
GLDS	**M**	83	7	92.2%	86.6%	91.4%
	B	2	13			
GLCM	**M**	85	5	94.4%	46.6%	87.6%
	B	8	7			
NGTDM	**M**	90	0	100%	0%	85.7%
	B	15	0			
SFM	**M**	71	19	78.8%	33.3%	72.3%
	B	10	5			
Edge	**M**	90	0	100%	0%	85.7%
	B	15	0			
Note: M:Malignant, B:Benign, ICAM: Individual Classification Accuracy of Malignant, ICAB: Individual Classification Accuracy of Benign, OCA: Overall Classification Accuracy						

Table 6 shows the results of various features. EDGE and NGTDM achieve highest accuracy of 100% but has lower individual accuracy in case of benign class so the accuracy achieved with GLDS feature is considered as best accuracy because it acquires overall accuracy of 91.4% along with the individual accuracy (Sensitivity) of benign and malignant as 86.6% and 92.2% respectively. The GLCM achieves an OCA of 87.6% with individual accuracies for benign and malignant class as 46.6 % and 94.4 % respectively. Different researchers have investigated using different methods, so for better observation of work, a comparison is displayed in Table 5.3.

Table 5.3: Comparison of our work with existing papers			
Author	**Classifier**	**Accuracy**	**Feature**
Land WH, Jr. *et. al*	SVM	72%	Morphological Features (seeded watershed transforms)
Pawar PD *et. al*	SVM	79.2%	Morphological Features
Gujral H *et.al*	FFT	85.1%	Texture Features (Gabor)
Proposed work	SVM	91.4%	Texture Features (GLDS)

Conclusion and Future Work: The ambition of this paper is to provide high accuracy and enhanced images for diagnosis of Lung Cancer. This classifies lung cancer into two categories and achieves higher accuracy as compare to former results of different authors. In this work GLDS attains the best results. The cases benign misclassified of NGTDM and Edge feature, so in future the authors will focus to enhance the results of the system with NGTDM and Edge feature and also try to achieve higher overall performance of system. For accomplishment of this target, the author will use various filters and various texture feature extraction methods by using different classifiers for betterment of results.

CONCLUSION AND FUTURE WORK

For designing an efficient CAD system for Lung Cancer detection, various CAD system designs based on Texture based features have been proposed in the present work. For the paper "CAD system for non small cell Lung Carcinoma using Laws' mask analysis" 95.67% result is computed in this paper using Laws' mask as Feature Extraction Module and SVM as a Classifier.

And for the paper **"CAD for two class classification of Lung Cancer using Statistical Features"** the results are obtained with GLDS feature for ICAM is 92.2%, ICAB is 86.6% and OCA calculated is 91.4%. For GLCM, ICAM calculated is 94.4%, ICAB is 46.6% and OCA calculated is 87.6%. For SFM feature the ICAM is 78.8%, ICAB is 33.3%, OCA is 72.3%, for NGDTM and Edge Feature results were misclassified so not considered.

The present work has been worked out for two-class Lung cancer classification of ultrasound cases for enhancement of the images and for higher accuracy. This work focuses on achieving higher accuracy as compared to the results of different authors. Lung Cancer has been classified into two categories and GLDS has attained the best results. Benign is misclassified for Edge and NGTDM feature so the author can focus on enhancing the results for Edge and NGTDM features. Different Texture Feature Extraction methods can be used to attain better results by using these features as well.

Future Scope: The limitation of the present work is that it has been carried out on the database that consists of digitized ultrasound images and not real data. Following are the recommendations for future work:

- The present work has been carried out on images developed using ultrasound as the imaging modality; however images acquired from MRI can also be used in the future to test the proposed algorithms.
- For accomplishment of this target, the author will use various filters and various texture feature extraction methods by using different classifiers for betterment of results.

REFERENCES

1. Haranshvir G and Khushali D: *A Review of Techniques for Lung Cancer Detection.* International Journal of Current Engineering and Technology. June 2015, Vol.5, No.3, pp-1597-1602

2. Kumar R and Srivastava R: *Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features.* Hindawi Publishing Corporation J of Medical Engineering. 2015. Volume 2015. Article ID 457906.

3. Laws KI: *Rapid Texture Identification in Proceedings of SPIE*, Image Processing for Missile Guidance. 1980, pp. 376-380.

4. Shraddha, Kulkarni G and Sahebrao BB: *Techniques for Lung Cancer Nodule Detection: A Survey.* International Research J of Engineering and Technology. 2015. Volume: 02 pp. 2593-2597.

5. https://ufhealth.org/lung-cancer

6. Ajil MV and Sreeram S: Lung *Cancer Detection from CT Image using Image Processing Techniques.* International J of Advance Research in Computer Science and Management Studies Research. 2015. Volume 3. pp. 249-254.

7. Land WH, Jr., Mckee D, Zhukov T, Song D and Qian W: *An End-to-End Process for Cancer Identification from Images of Lung Tissue.* IEEE. 2007. pp. 233-239.

8. Pawar DP: and Gulhane Sanjay.M. *Multistage Approach for Biomedical Image Processing: A Review.* International Journal of Advancements in Technology. 2013. Volume- 5. pp. 28-35.

9. Bhusri S, Jain S and Virmani J: *Classification of Breast Lesions using Laws' Mask Texture features.* Computing for Sustainable Global Development (INDIACom).2016. pp. 1700-1704.

10. Bhusri S, Jain S and Virmani J: *Breast Lesions Classification using the Amalagation of morphological and texture features.* International J of Pharma and BioSciences (IJPBS). Apr-Jun 2016. pp. 617-624.

11. Rana S, Jain S and Virmani J: *Classification of Kidney Lesions using Gabor Wavelet Texture Features.* Computing for Sustainable Global Development (INDIACom). 2016. pp. 2528-2532.

12. Rana S, Jain S and Virmani J: *SVM-Based Characterization of Focal Kidney Lesions from B-Mode Ultrasound Images.* Research J of Pharmaceutical, Biological and Chemical Sciences (RJPBCS). July- Aug 2016. Volume-7(4). Page No. 837.

13. Bhusri S, Jain S and Virmani J: *Classification of breast lesions using the difference of statistical features.* Research J of Pharmaceutical, Biological and Chemical Sciences (RJPBCS). July- Aug 2016. Page No. 1366.

14. Rana S, Jain S and Virmani J: *Classification of Focal Kidney lesions using Wavelet-Based Texture Descriptors.* International J of Pharma and Bio Sciences. July-Sep 2016. pp. 646-652.

15. Ultrasound Cases.info

16. A.Gajdhane V and Deshpande LM*: Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques.* IOSR J of Computer Engineering (IOSR-JCE). 2014. Volume 16. Issue 5. Ver. III. PP 28-35.

17. Rachidi M, Marchadier A, Gadois C, Lespessailles E, Chappard C and Benhamou C.L: *Laws' masks descriptors applied to bone texture analysis: an innovative and discriminant tool in osteoporosis.* Skeletal Radiology. 2008. vol. 37. pp. 541-548

18. Virmani J, Kumar V, Kalra N and Khandelwal N: *Prediction of cirrhosis from liver ultrasound B-mode images based on Laws' mask analysis.* IEEE International Conference on Image Information Processing, ICIIP-2011. Himachal Pradesh, India, 2011. pp. 1-5.

19. Seng H., Chai H.Y. and Swee T.T: *Research on Laws' mask texture analysis system reliability* Reasearch J of Applied Sciences. Engineering and Technology. 2014. vol. 7. pp. 4002-4007.

20. Kaur S: Comparative Study Review on *Lung Cancer Detection Using Neural Network and Clustering Algorithm.* International J of Advanced Research in Electronics and Communication Engineering (IJARECE). 2015. Volume 4. pp. 169-174.

21. Ada and Kaur R: *Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images.* International J of Advanced Research in Computer Science and Software Engineering. March 2013. Volume 3. pp. 187-190.

22. Jain S: *Regression analysis on different mitogenic pathways.* Network Biology. June 2016. pp. 40-46.

23. Jain S: *Mathematical Analysis using Frequency and Cumulative Distribution functions for Mitogenic Pathway.* Research J of Pharmaceutical, Biological and Chemical Sciences (RJPBCS). May - Jun 2016. Volume- 7(3). pp. 262-272.

24. Virmani J, Kumar V, Kalra N and Khandelwal N: *SVM based characterization of liver cirrhosis by singular value decomposition of GLCM matrix.* International J of Artificial Intelligence and Soft Computing. 2013. vol. 3. pp. 276-296.

25. Jain S, Naik PK and Bhooshan SV: *Nonlinear Modeling of cell survival/ death using artificial neural network.* International Conference on Computational Intelligence and Communication Networks (CICN2011), Gwalior, India. Oct 2011. pp 565-568.

26. Jain S. and Chauhan DS: *Mathematical Analysis of Receptors For Survival Proteins.* International J of Pharma and Bio Sciences (IJPBS). 2015. Vol-6. pp. 164-176.

27. Jain S and Chauhan DS: *Linear and Non Linear Modeling of Protein Kinase B/ AkT.* International Conference on Information and Communication Technology for Sustainable Development (ICT4SD - 2015), Ahmedabad, India, pp 81-88.

28. Jain S: *Communication of signals and responses leading to cell survival / cell death using Engineered Regulatory Networks.* PhD Thesis, Jaypee University of Information Technology, Solan, Himachal Pradesh, India, 2012.

29. Jain S and Naik PK: *System Modeling of cell survival and cell death: A deterministic model using Fuzzy System.* International J of Pharma and BioSciences. 2012. pp. 358-373.

30. Pathan A, Saptalkar BK: *Detection and Classification of Lung Cancer Using Artificial Neural Network.* International Journal on Advanced Computer Engineering and Communication Technology. 2016. Vol-1. pp. 62-67.

31. Fatma Taher, Naoufel Werghi and Hussain Al-Ahmad, "*Computer Aided Diagnosis System for Early Lung Cancer Detection*", Algorithms, 8, 1088-1110, 2015

32. Jain S and Chauhan DS: *Implementation of fuzzy system using different voltages of OTA for JNK pathway leading to cell survival/ death.* Network Biology. 2015. pp. 62-70.

33. Poonam Bhayan, Gagandeep Jindal, "*A Segmented Morphological Approach to Detect Tumor in Lung Images*", IJCST Vol. 2, Issue 3, September 2011

34. P. Delogua,, S.C. Cheranb, I. De Mitric, G. De Nunzio, M.E. Fantacci,F. Fauci, G. Gargano, E. Lopez Torres, R. Massafra, P. Oliva, Preite Martinez, G. Raso, A.

Retico,S. Stumbo, A. Tata, "*Pre-processing methods for nodule detection in lung CT*", 2011

35. Jain S: *Design and Simulation of Fuzzy System Using Two Stage CMOS Operational Amplifier*. J of Active and Passive Electronic Devices. 2014. pp. 329-338.

36. Jain S*: Mathematical Analysis and Probability Density Function of FKHR pathway for Cell Survival /Death.* Control System and Power Electronics. 2015 pp 84-93.

37. Amandeep, Jain S and Bhusri S: *CAD System for Non Small Cell Lung Carcinoma using Laws' Mask Analysis.* International Conference on Computing for Sustainable Global Development, BVICAM. 2017. pp. 6285-6288.

38. Sharma S, Jain S, Bhusri S: *Classification of Breast Lesions using Gabor Wavelet Filter for Three Classes.* International Conference on Computing for Sustainable Global Development, BVICAM.2017. pp 6282-6284.

39. Nancy,Paramjit Kaur, "Identifying lung cancer in its early stage using neural network & GA algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 2, 2015

40. Dhiman A, Singh A, Dubey S and Jain S: Design *of Lead II ECG Waveform and Classification Performance for Morphological features using Different Classifiers on Lead II*, Research J of Pharmaceutical, Biological and Chemical Sciences. July-Aug 2016. pp. 1226- 1231.

41. V. Banupriya , J. Bagyamani: "*Lung Cancer Detection using circular hough transform technique",* International Journal of Advance Research In Science And Engineering, IJARSE, Vol. No.4, Special Issue (03), March 2015

42. *J. M. DURGE, PROF. N. P. BOBADE, DR. N. N. MHALA, "Image enhancement to detect the lung cancer at early stage using median filter"*, ISSN, Volume 3 Issue 4: 2015

43. *Georgy George, Nisha J.S.,"A Novel Approach for the Detection of Small Cell Lung Cancer Based on Entropy and PSNR Value",* International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3925-3929

44. Mabrukah Edrees Fadel , Rasim Amer Ali,"*Computerized System For Lung Nodule Detection in CT Scan Images by using Matlab"*. International Conference on Automation, Control, Engineering and Computer Science

45. https://upload.wikimedia.org/wikipedia/commons/thumb/1/11/Pie_chart_of_lung_can cers.svg/2000px-Pie_chart_of_lung_cancers.svg.png

46. http://medicalassessmentonline.com/john/CancerLungSquamouscell2014.jpg

47. http://www.cancertherapyadvisor.com

48. http://medicalassessmentonline.com/john/CancerLargeCellLungCancer2014.jpg

49. Image Processing and Analysis in JAVA ,Image J 1.49 vesion1.6.024
http://imagej.nih.gov/ij/download/win32/ij149-jre6-64.zip

LIST OF PUBLICATIONS

1. Amandeep, Shruti Jain, Sahil Bhusri, "CAD System for Non Small Cell Lung Carcinoma using Laws' Mask Analysis", March 1st - 3rd , 2017, pp 6285-6288, Proceedings of the 11th INDIACom: 4th 2017 International Conference on Computing for Sustainable Global Development, BVICAM, New Delhi.

2. Amandeep, Shruti Jain, Sahil Bhusri, "CAD for Two Class Classification of Lung Cancer using Statistical Features" Indian Journal of Pharmaceutical Sciences [under review]

Submission Info

SUBMISSION ID	806851518
SUBMISSION DATE	29-Apr-2017 12:14
SUBMISSION COUNT	1
FILE NAME	Thesis-Aman_152012.docx
FILE SIZE	1.26M
CHARACTER COUNT	66780
WORD COUNT	12326
PAGE COUNT	68

ORIGINALITY

OVERALL	14%
INTERNET	6%
PUBLICATIONS	9%
STUDENT PAPERS	27%

GRADEMARK

LAST GRADED	N/A
COMMENTS	0
QUICKMARKS	