Improved Classification of Ultrasound Breast Cancer RF Data with Nakagami and Derived Nakagami Parameters Using Advanced Machine Learning Algorithm

by

Ahmad Shafiullah 160021055
Ahmad Chowdhury 160021067
Rezwana Rosen 160021142

A Thesis Submitted to the Academic Faculty in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh

March 2021

Improved Classification of Ultrasound Breast Cancer RF Data with Nakagami and Derived Nakagami Parameters Using Advanced Machine Learning Algorithm

Approved by:

Prof. Dr. Md. Ruhul Amin

Head of the Department
Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Boardbazar, Gazipur-1704.

Date: 3/15/2021

Declaration of Authorship

This is to certify that the work in this thesis paper is the outcome of research carried out by the students under the supervision of Dr. Md. Ruhul Amin, Professor and Department of Electrical and Electronic Engineering (EEE), Islamic University of Technology (IUT).

Authors

Ahmad Shafiullah
ID-160021055

Ahmad Chowdhury
ID-160021067

Rezwana Rosen
ID-160021142

Table of Contents

List of Tables

List of Figures

List of Acronyms

BI-RAD	Breast Imaging Reporting and Data System
MRI	Magnetic Resonance Imaging
QUS	Qualitative Ultrasound
BSC	Backscattered Coefficient
ESD	Effective Scatterer Diameter
ROC	Receiver Operating Characteristics
SNR	Signal to Noise Ratio
PCA	Principal Component Analysis
ICA	Independent Component Analysis
SVM	Support Vector Machine
RF	Radio Frequency
ROI	Region of Interest
MLP	Mean Lesion Pixel
SLP	Standard Deviation of Lesion Pixel
FNPA	Four Neighborhood Pixel Algorithm
HTC	HT Coefficient
RFE	Recursive Feature Elimination
DT	Decision Tree
RFE_CV	Recursive Feature Elimination with Cross Validation
RF	Random Forest
FP	False Positive
TP	True Positive
TN	True Negative
FN	False Negative

FPR	False Positive Rate
TPR	True Positive Rate
TNR	True Negative Rate
FNR	False Negative Rate
AUC	Area Under Curve

Acknowledgements

We are very much thankful to our honorable supervisor, Dr. Md. Ruhul Amin sir for his support and patience all throughout our thesis work. He helped us with his vast knowledge in every field of engineering and guided us properly along the way. We feel privileged to get him as our supervisor and this thesis research could not have been possible without him.

We would also like to relay our heartfelt gratitude to Dr, Sheikh Kaisar Alam sir, Senior Member, IEEE, Visiting Research Professor at the Centre for Computational Biomedicine Imaging and Modelling, Rutgers University, the State University of New Jersey, for providing us with the invaluable medical dataset based on which our research was conducted.

2021

Abstract

Ultrasound imaging is one of the medical imaging modalities that has become popular among the researchers for breast cancer diagnosis because of its radiation free and non-invasive nature compared to other screening procedures like X-ray Mammography, CT-scan etc. In ultrasound studies, various quantitative analysis has been shown to be capable of breast cancer diagnosis. This research utilizes the statistical parameters from the well-known Nakagami distribution and investigates their potential in improvement of noninvasive semi-automated identification of breast cancer. The dataset used in the study has 130 biopsy-proven patients consisting of 104 benign and 26 malignant cases with traced lesion boundaries. In this study, seven types of Nakagami and derived Nakagami images have been generated for each patient from the basic and derived parameters of Nakagami distribution. To determine the suitable window size for image generation, an empirical analysis has been conducted using three window sizes of width 0.2 mm and lengths 0.1875 mm, 0.45 mm and 0.75mm. The images were analyzed quantitatively for feature extraction, followed by feature selection using the machine learning algorithm- Recursive Feature Elimination with Cross Validation. A combination of six features was found for window size 0.75mm which gives a classification accuracy of 92.3% and area under the ROC curve score of 0.95 using Linear Support Vector Machine classifier.

Chapter 1

Introduction:

1.1 Introduction

Breast cancer has been the most common of all cancers in women. It has become a global issue affecting both men and women all over the world. Approximately 30% of all new cancer diagnoses in women is estimated to be breast cancer. 15% of total cancer deaths is due to breast cancer making it the leading cause of cancer death next to lung cancer. In the future years, with an approximation of 1 in 8 women (13%) will be diagnosed breast cancer that is invasive in nature and 1 in 39 women (3%) may die from only breast cancer [1].

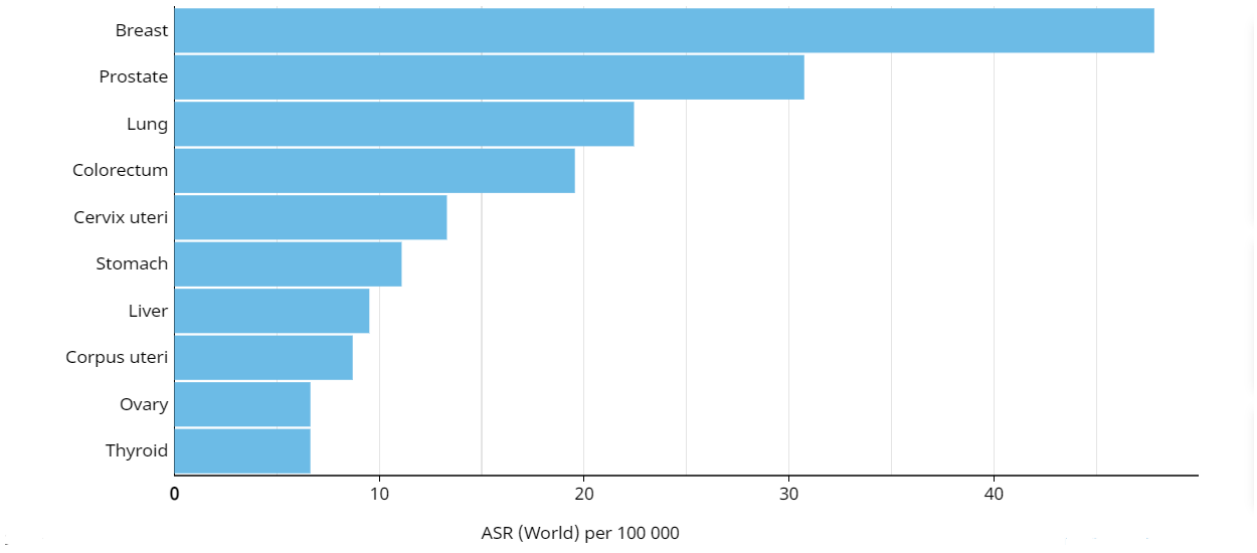


Figure 1.1: Estimated age-standardized cancer incidence rates in 2020 [2]

It is a disease where cells in the breast tissue grow in an uncontrolled amount forming a lump or a mass. It doesn't typically show any symptoms in the early stages and can be easily treated if properly screened. Early detection and increased awareness of breast cancer accounted for an estimated 375,900 fewer deaths. As a result, several screening technologies have been developed for an early detection of breast cancer.

1.2 Breast cancer detection technologies in development

1. Mammography: It is the process of taking X-Ray images of breast for diagnosis. Detection is done by measuring characteristic of breast masses or calcification from the images. An abnormal mammogram is not the definite indicator of cancer, additional tests are required for confirmation.
2. Magnetic Resonance Imaging: It is a process that uses a magnetic field and computer-generated radio waves to create detailed images of the organs and tissues in your body. The resolution of a MRI image is 10 to 100 times better than a mammogram. But, cost of an MRI is multiple times higher than taking a mammography.
3. Ultrasound Imaging: Sound waves are produced form a transducer and by pulse echo system a patient's breast image pattern is produced. Although ultrasound images are not detailed as mammogram or MRI but with proper processing and quantitative analysis ultrasound images can be used as a screening process for breast cancers.

1.3 Significance of the research

Our research focuses on using ultrasound imaging in detection of breast cancer. Even though mammography being the most common of all the screening processes, it's an expensive procedure that exposes the patient to X-Ray and can't be repeated multiple times. MRI is even more powerful tool in this regard but that is also of limited use due to both health and monetary reasons. Ultrasound is a much cheaper alternative compared to these and can be used as a primary approach to differentiate between a cancerous and non-cancerous breast mass. The purpose of ultrasound imaging is not to identify the cancerous lesion but to differentiate between malignant and benign classes.

Women in Bangladesh general are reluctant to use any of existing screening technologies either due to harmful effect or the cost, but if ultrasound imaging is made a viable option it will be indeed a popular one and can cease much of the cancer in the early stage, when it's easily curable.

1.4 Objectives of the Research:

The primary objectives of the research can be stated as following:

I. To generate a new image from Nakagami and derived nakagami parameters by fitting the RF echo signal using Nakagami distribution.

II. To extract morphometric and acoustic parameters from the generated Nakagami images.

III. To use Machine Learning algorithm and find out a combination of features for best classification performance through feature selection algorithm

IV. To classify benign and malignant tumor masses by using a supervised trained model

V. To use performance metrics to analyze the validity of the features and classification model.

1.5 Main Contributions:

In comparison to different scholarly works of quantitative analysis from ultrasound images this thesis provides the following contributions:

1. 7 different types of Nakagami images are generated which can be used for further research work.
2. Several new features are extracted from Nakagami images that can be used in classification techniques.
3. An empirical analysis is made by taking different window sizes and a linear relationship between accuracy and window size is established.
4. A way of combining a group of features via usage of machine learning to generate the best possible outcome is discussed and a perfect combination of features is proposed.
5. A better classification score than any recent published work has been achieved.

1.6 Thesis Outline:

Our thesis is outlined as follows:

Chapter-1 gives idea about the foundation and inspiration behind this study. It also includes the significance, main contributions as well as aims and objectives of this research. At the end of this chapter the thesis outline has also been mentioned.

Chapter-2 presents the relevant research and studies that helped us to get an overall idea about ultrasound imaging, modelling backscattered echo and different machine learning algorithms before developing this thesis.

Chapter-3 focuses on the workflow of our research. It contains the details of the procedures that were followed for image formation, feature extraction, machine learning algorithms used for feature selection and tumor classification.

Chapter-4 introduces the machine learning algorithm called recursive feature elimination technique and different other machine learning models.

Chapter-5 contains in depth information about different performance metrics and scores and their calculation procedures.

Chapter-6 concentrates on the simulation and classification results obtained from the thesis. Also, it contains the performance scores of the individual parametric features that were selected through feature selection.

Chapter -7 summarizes and concludes the work with discussion about future research scopes in this area.

Chapter 2
Literature Review

2.1 Introduction:

Conventional ultrasound imaging procedures mainly utilize qualitative features of the breast tumors such as size, shape, border etc. for classifying the tumors into benign and malignant cases. Through recent studies it has been seen that in conventional ultrasound images, almost all types of breast cancers, including invasive ductal carcinomas (more than 90%) are visible. American College of Radiology (ACR) has developed the Breast Imaging Reporting and Data System (BI-RADS) which provides a glossary of features providing description of the ultrasound properties of breast tumors that plays an effective role in the breast cancer classification [1,2,9]. BI-RADS outlined six different types of possible findings, ranging from Category 0, which states incompetence of the assessment, and necessity of additional imaging evaluation to classify the required tumor, up to Category 5, which suggests that the tumor is highly suggestive of malignancy.

Despite the advantages offered by BI-RADS, the successful application using BI-RADS characteristics is largely dependent on the skills and knowledge of the clinician. Also the classification system through BI-RADS depends on the qualitative features of the tumor, that may be subjective. Hence, quantitative ultrasound (QUS) techniques have received a lot of attention in the recent years, especially various machine learning algorithms along with QUS techniques tend to make the breast cancer evaluation procedure more precise, quantitative, reproducible and less operator dependent.
Usually, three aspects of QUS have been used from the past years till now for breast cancer classification with high accuracy and these are:

1. Classification using physical properties of tissue.
2. Classification using parameters developed from the spectrum of ultrasound echoes.
3. Classification using envelope of backscattered ultrasound with help of statistical analysis.

Though our study has mainly focused on the last two aspects of the above-mentioned aspects of QUS, the relevant research went far beyond. Some of the significant findings and proposed ideas related to these aspects are mentioned below:

2.2 Analysis of Physical Properties of Tissue:

Mamou, et al. [3] showed a new perspective in the field of ultrasound imaging by introducing a new methodology called 3D Impedance map(3DZM), along with quantitative ultrasound imaging to classify three kinds of cancer tumors of rats, fibroadenoma, carcinoma and sarcoma based on the parameters, acoustic concentration and average scatterer size [3]. In their work, using Gaussian FF they could identify significant differences between fibroadenoma and the other two types of cancer [3].

Analysis of physical properties of tissue, which includes backscatter coefficient (BSC) and attenuation coefficient, was first proposed by D'Astous and Foster [4]. In a later study, Nam, et al. [5] attempted to differentiate between tumor types using the attenuation coefficient and BSC parameters, and reported attenuation coefficients that were 20% higher for carcinoma than for fibroadenoma. They also discovered that most carcinomas had lower frequency-average BSC values than fibroadenomas, as well as greater variability in Effective Scatterer Diameter in fibroadenomas (ESD) [5].

Also, Michael L.Oelze [6] in one of his studies, observed spectral based and envelope statistics based techniques consisting of the estimators like, backscatter coefficient, attenuation coefficient, effective acoustic concentration and some other statistical parameters at high ultrasonic frequencies for improved diagnostic QUS imaging. He found, QUS techniques over the frequency range of 5 to 25 MHz based on spectral features and envelope statistics to be capable of differentiating benign from malignant tumors in rodent models of breast cancer and can be used for the quantitative analysis of cancer as the attenuation compensation problem has been overcome [6].

2.3 Spectral and Statistical Analysis of Ultrasound Echo:

Among the many groups who have studied the automated methods of breast-lesion classification through spectral analysis of ultrasound echoes, Lizzi, et al. is one of the first to develop spectral parameters like-midband fit, spectral intercept and spectral slope [7,8]. In different studies it has been found that the capacity of a single parameter to distinguish benign and malignant lesions is not that much reliable. So later, Alam, et al. [9] in one of their studies implemented a multi-feature analysis procedure (semi-automated as the tracing of the boundary of the lesion was done manually) with the aim of non -invasive identification of malignant breast lesions. Alam, et al. [9] processed RF data of 130 biopsy-proven patients collected from three clinical sites, using sliding-window Fourier analysis and extracted acoustic features from the spectral parameters proposed by Lizzi, et al., like, echogenicity, heterogeneity, shadowing etc., as well as morphometric features related to the shape and size of the lesion like, area, aspect ratio and boundary roughness and some hybrid features (consist of both acoustic and morphometric information). Among all the extracted features Alam, et al. found in total four quantitative features, including lesion-margin definition, spiculation and border irregularity to be the most useful ones for breast lesion classification, and reported an area under the receiver-operating-characteristics (ROC)curve of 0.947±0.045.

Different mathematical statistical distributions are used to model the backscattered envelope for the statistical analysis of ultrasound signals. These statistical distributions are applied to the probability density function of the backscattered echoes based on the randomness of ultrasonic backscattering for evaluating properties of tissue scatterers. Nakagami, K and Homodyned K distributions are some of the most notable statistical distributions used for breast cancer classification.

The Nakagami distribution, a two-parameter distribution proposed by Shankar received the attention of the researchers for statistical analysis of ultrasound backscattered signals due to its less computational complexity and the ability to describe all scattering conditions in medical ultrasound including pre-Rayleigh, Rayleigh, and post-Rayleigh distributions [10-12]. Also some Nakagami distributions in compounding form such as, the Nakagami-Gamma [13,14], Nakagami-lognormal [15], Nakagami-inverse Gausssian [15], Nakagami generalized inverse Gaussian [16], and Nakagami Markov random field models [17], also have the potential to

model backscattered envelope of ultrasound signal. From the table below it is clearly visible that nakagami distribution is able to encompass all scattering conditions which is necessary to classify breast lesions accurately [12]:

Table 2.1: Relationship Between Nakagami and Rayleigh/Rician Distributions and the Phase Characteristics [12]

Envelope Designation	Range of m	Phase Statistics	Nakagami designation
Rayleigh	m=1	Uniform	Nakagami
Pre-Rayleigh	$0.5 < m < 1$	Uniform	Nakagami
Pre-Rayleigh	$0 < m < 0.5$	Uniform	Nakagami-Gamma
Pre-Rayleigh (Rician)	m> 1	Nonuniform	Nakagami-Rice
Generalized Rician	$0.5 < m < 1$	Nonuniform	Nakagami

It has been observed in a study [12] that, different values of the shape parameter m represents different statistical distributions. When m=1, the Nakagami density function becomes like Rayleigh distribution [12]. If the scatterers are randomly located but have random scattering cross sections, the distribution becomes like a pre-Rayleigh distribution with uniform phase statistics and the value of m ranges from 0.5 to 1[12]. In addition to randomly located scatters,if there are periodic scatterers in the range cell separated by a spacing of integral multiple of

wavelength corresponding to the frequency of demodulation, the distribution becomes like post-Rayleigh or Rician distribution with nonuniform phase statistics and the value of m is greater than [12]. Lastly, in addition to randomly located scatterers, if there are periodic scatterers in the range cell separated by a spacing of integral multiple of a quarter wavelength corresponding to the frequency of demodulation, the distribution becomes like generalized Rician distribution with nonuniform phase statistics and the value of m ranges from 0.5 to 1 [12].

Shankar, et al. [12] proved the potential of Nakagami parameters, using the two parameters, m (effective number) and α (effective cross section) to classify the breast masses of total 52 patients (14 malignant, 38 benign). In this study, they reported areas under the receiver-operating-characteristics (ROC) curve to be 0.79±0.11 and 0.828±0.10 for the two parameters individually respectively. They also reported that, after using multiple images for the same patients, that is by increasing the total number of patients to 75(37 malignant, 38 benign), ROC curve area increased to 0.838±0.065 and 0.85±0.06.

In another study, to enhance the classification performance Shankar, et al. [18] proposed the concept of compounding Nakagami parameters. For this study, they collected two images for each of the 85 patients (24 malignant, 61 benign) and observed that the discrimination of lesion after combining the normalized Nakagami parameters from two images for each patient was comparatively better than using the parameter from a single image. They found the area under the ROC curve to be 0.8316 after compounding the Nakagami parameter m for each patient.

Shankar, et al. [19] proposed to use K distribution along with Nakagami distribution to model the backscattered ultrasound echo for better classification. They did a five-parameter analysis for in total 99 patients (29 malignant, 70 benign) based on these two distributions at the site, boundary, spiculated region and shadow of the mass. Combination of these features lead to an increased area under the ROC curve which is 0.96±0.02. Addition of the level of suspicion values of the radiologist with these parameters increased the area of the ROC curve to 0.97 ± 0.014.

For ultrasonic tissue characterization, Tsui, et al. [20] proposed an improved form of the Nakagami parameter m called mlog calculated using ultrasonic backscattered envelopes compressed by logarithmic computation. According to their study, this logarithmic transform is simple and less computation is needed to estimate the m parameter, whose sensitivity is

better than the one estimated from the uncompressed envelope for both non-focused and focused transducers [20]. They found the dynamic range between lowest and highest scatterer concentration for non-focused and focused transducer of the m parameter to be 0.14 and 0.45 respectively and for mlog to be 3.7 and 5.14 [20]. This proves mlog to have higher sensitivity, leading it to be a better parameter for differentiating and classifying various scatterer properties [20].

Tsui and Chang [21] developed ultrasonic parametric image based on the Nakagami parameter map and compared its performance with B-mode image for tissue characterization and differentiating scatterer properties. They found Nakagami image constructed with an optimal window size, has the ability to impart both the global and local backscattered statistics of the ultrasonic signals in a tissue leading it to be a very efficient local scatterer concentration detector [21]. They mentioned this optimal window size for constructing the Nakagami image to be square with the side length equal to three times the pulse length of the incident ultrasound [21]. They also concluded that Nakagami image can effectively assist B-mode image in medical diagnosis for the following reasons:

1. The system and operator influence the creation of B-mode images, while the Nakagami image shows reliable results that are independent of dynamic ranges and system gains since it is shaped by the shape of the envelope rather than the magnitude of the backscattered signal [21].

2. Local scatterer concentrations can be quantified using Nakagami images, and backscattering information can be collected from weaker echoes that may be lost in the B-mode image [21].

Further, Tsui, et al. [22] observed the potential of Nakagami image to differentiate different scatterer structures at different signal-to-noise ratios (SNRs) and concluded the following:

1. By quantifying the statistical distribution of backscattered envelope Nakagami images can distinguish different scattering concentrations for single, hypoechoic and hyperechoic structures [22].

2. Nakagami images may be able to produce artifacts when the region of interest contains anechoic tissues [22].

3. When both the scatterer concentration and SNR are low, the Nakagami image does not perform well in distinguishing regions with high scatterer concentrations [22].

4. Nakagami image is more sensitive than conventional B-scan and can complement B-scan image to identify the region of interest for having larger CNR than B-scan image [22].

5. Nakagami images may have the potential for assessment of treatment of tissues [22].

Tsui, et al. [23] also explored the ability of Nakagami parametric image to distinguish benign and malignant breast tumors. They used the backscattered signals to form Nakagami parametric image (m parameter) of breast tumor to classify breast cancer. In their study, they collected data for in total 100 patients (50 malignant,50 benign) and obtained an area under the curve of 0.81 ± 0.04, accuracy of 82 percent, sensitivity of 92 percent and specificity of 72 percent [23]. They added malignant tumors to have backscattered signals with more complex composition leading to be more pre-Rayleigh distributed than those from benign tumors [23]. They also concluded that Nakagami image is relatively independent of signal and image processing built in ultrasound systems which may assist B -mode image for texture analysis of the tumor [23]. Larrue and Noble [24] in one of their studies, suggested ways for error factor correction for calculation of Nakagami parameters.

Dutt and Greenleaf [25] proposed the Homodyned K distribution for modelling ultrasound echo envelope and later it was modified by Hruska [26] and Hruska and Oelze [27]. Hruska and Oelze [27] proposed an improved estimation algorithm based on SNR, skewness and kurtosis of fractional order moments, to estimate the parameters, μ (number of scatterers per resolution cell) and K (the ratio of coherent to incoherent backscatter signal energy) obtained from Homodyned K model and used angular compounding (120 angles of view) for reducing the variance within the estimates, and thus tried to classify two types of tumors. The Homodyned K parameter along with BI-RADS classifiers was later applied by Trop et al. [28], who demonstrated their ability to reduce the number of biopsies performed on breast cancer patients by 25%, whilst maintaining 100% sensitivity.

2.4 Computer Aided Diagnosis (CAD):

Cheng, et al. [29] in their survey, introduced the four stages of the CAD system and their advantages and disadvantages for breast cancer detection and classification. They discussed different filtering techniques for image enhancement and segmentation algorithms like, Active Contour Model, Markov Random field etc. [29]. They also discussed different feature extraction algorithms like Principal Component Analysis (PCA), Independent Component Analysis (ICA) and feature selection algorithms for dimensionality reduction of datasets with a large number of features [29]. They also suggested machine learning classifiers for distinguishing lesion from non-lesion like, Linear classifiers, Bayesian Neural network (BNN), and also for distinguishing malignant from benign lesions like, LDA, LOGREG, ANN, BNN, decision tree, SVM [29].

From the above discussion, it is evident that Nakagami statistical parameters, parametric images and different machine learning classifiers have high capability to make breast lesion analysis and classification procedure fully automated. This thesis utilizes this ability of Nakagami distribution, Nakagami parametric images to be specific to classify malignant and benign breast lesions.

Chapter 3

Methodology

3.1 Introduction

Ultrasound imaging process starts with collection of RF data from the patients with tumors (biopsy-proven) followed by tracing of the boundary of the lesions by knowledgeable non-clinicians. The demarcation of the regions of analysis was done through a semi-automated procedure. Based on the tracings and demarcations made on the boundary of the lesion, the regions-of-interest were analyzed in this study. This study utilizes the features extracted through the analysis of almost nine regions-of-interest for different window sizes and were used for breast lesion classification through linear SVM. Figure 3.1 represents the flowchart of the algorithm followed in this study.

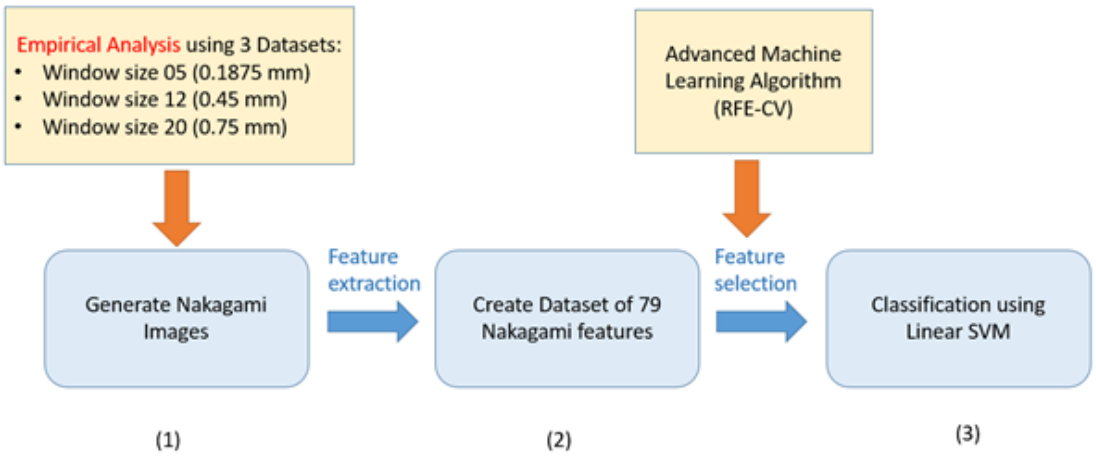


Figure 3.1: Overview of methodology

3.2 Ultrasound Imaging

Diagnostic ultrasound usually uses frequencies ranging from 2MHz to 20 MHz (million cycles per second). Since ultrasound can be concentrated into tiny, well-defined beams that can probe the human body and interact with tissue structures to create images, it is commonly used as a diagnostic tool. When electrical pulses are applied to the transducer (one of the main parts of ultrasound imaging system), it produces ultrasound pulses as its first feature. When RF echo pulses return to the body surface after a very short period of time, the transducer picks them up and converts them back into electrical pulses, which are then processed by the device and shaped into a B-mode image.

Most part of the ultrasound energy is absorbed and some part of the signal is attenuated as an ultrasound pulse beam is passed through the body. Internal body structures reflect some of the pulses, sending echoes back to the surface, where they are captured by the transducer and used to shape the image. As a result, the general ultrasound picture is a representation of the body's structures or reflecting surfaces. In our study also the RF-echo signal data was acquired in this process.

The general grey and white background seen in the ultrasound picture is produced by a mixture of various tissue types and several surfaces of anatomical parts. The picture is dark since there are no reflective surfaces inside a fluid, such as a cyst. As a consequence, the general ultrasound picture is a reflection of the anatomical area's echo producing sites.

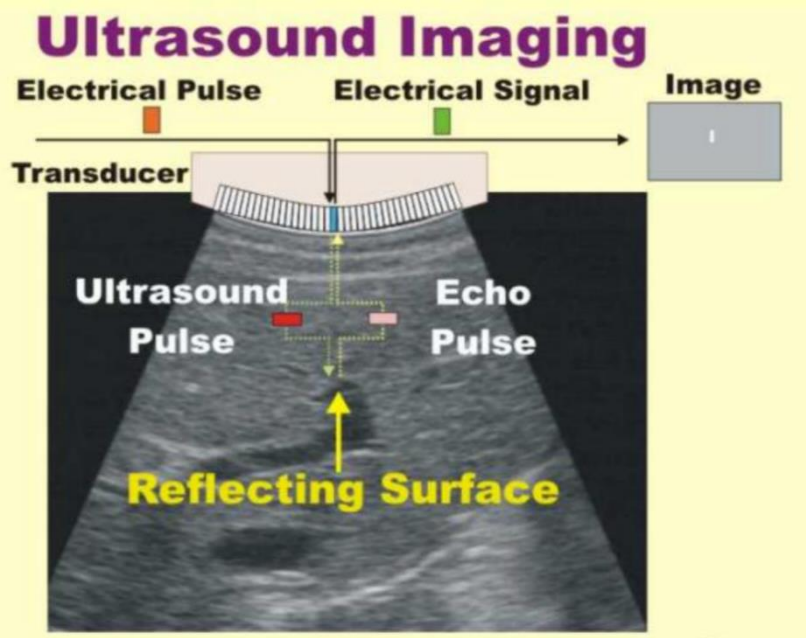


Figure 3.2: Basic Ultrasound Image Processing [31]

3.3 Data Acquisition

The dataset used for our research has been collected from ATL's PMA study approved by IRB (IRB: institutional review board) undertaken in 1994. The type of the data in the dataset is RF data and comprises biopsy proven analysis results of 130 patients of which 104 are benign and 26 are malignant. The data were collected from three clinical sites (Thomas Jefferson University, University of Cincinnati and Yale University). The exclusion criteria of our dataset included patients under age 18 for consent restrictions, breast implant, previous breast

carcinoma biopsy, simple cyst, pregnant women, males, transsexuals, micro calcifications not associated with a mass on sonography. The masses were imaged using a Philips Ultrasound (Bothell, WA) UM-9 HDI scanner with an L10-5 (7.5 MHz) linear-array transducer. RF echo-signal data was digitally acquired by interfacing a Spectrasonics Inc. (King of Prussia, PA) acquisition module with the scanner. For every lesion at least one radial and one anti-radial view were obtained. The operator chose a single transmit focal length and a default power level for the L10-5 transducer. The RF data was sampled at a rate of 20 MHz with a 14-bit effective dynamic range. For TGC (time-gain-control) obtained for every scan and the collected data were corrected before processing [9].

3.4 Introduction to Backscattered Echo Envelope

All conventional medical ultrasonic imaging systems are based on a simple pulse-echo technique utilizing the backscattered echo waveform. In ultrasonic imaging system, the transducer generates a short and relatively broad-band pulse that travels into the tissue for perhaps about a microsecond with a velocity of about 1.5 mm/μsec and as the pulse propagates in tissue, it will be attenuated due to absorption and scattering. The backscattered waveform (backscatter is the reflection of waves, particles, or signals back to the direction from which they came) whose amplitude will be considerably reduced from the incident pulse, then recorded by the same transducer.

The transducer is a phase sensitive device, so the output of the transducer (produced by the pressure of the backscattered signal on the face of the transducer) will be an RF trace (representational of the backscattered signal) recorded as a function of depth within the tissue. Since the backscattered signals span a wide dynamic range (100 to 130dB), some compression techniques are required if the signals are to be viewed. At least some of the compression is done under front panel control known as TGC (or time gain compensation) which increases the gain as some function of depth so that signals from greater depth (which are more attenuated) will be amplified more than signals near the transducer. The envelope of the received signal is then detected at some point of this compression process

3.4.1 A-mode and B-mode Display/Image

The RF data or detected envelope can be displayed on an oscilloscope or an appropriate monitor. This trace, which is an amplitude-modulated display of the backscattered signal, is

known as an A-mode display or scan. To make an image, first the A scan is taken and the amplitude mode display is converted into a brightness mode display along a vertical axis that converts a horizontal axis with waveform spikes into a vertical axis with a series of bright dots. Since the image is formed from a display of bright dots, it is known as B-mode image.

3.4.2 Theoretical Model of Backscattered Echo Envelope

In ultrasonic imaging, when an acoustic pulse travels through tissue, the backscattered echo received at the transducer may be modeled as the algebraic sum of the contributions from the individual scatterers [12]. If there are N scatterers in the range cell, and a_n and θ_n represent the amplitude and the phase of the n^{th} scatterer respectively, the backscattered echo may be written as [12]:

$$s(t) = \sum_{n=1}^{N} a_n \cos(w_o t - \theta_n) \qquad (1)$$

where $w_o = 2\pi f_o$, f_o is the center frequency of insonation **[12]**. In terms of the inphase and quadrature elements, X and Y, the backscattered echo s(t) can be written as:

$$s(t) = X\cos(w_o t) + Y\sin(w_o t) \qquad (2)$$

where,

$$X = \sum_{n=1}^{N} a_n \cos(\theta_n)$$

$$Y = \sum_{n=1}^{N} a_n \sin(\theta_n)$$

The envelope of the backscattered echo, R is given by,

$$R = \sqrt{(X^2 + Y^2)} \qquad (3)$$

The backscattered echo signal may have two types of components based on scatterer positions; if there exists a periodicity pattern in the position of the scatterers or if there exists strong specular reflections, then a coherent (or deterministic) component appears in the received signal, because of a long-range organization (relative to the wavelength) and usually it is represented by ϵ [12,30]. The power of the coherent signal component is called coherent signal power. The remaining power (from the total signal power) is called the diffuse signal power and corresponds to the diffuse (or random) component, made of a diffuse collection of scatterers [12,30]. The diffuse signal component is usually represented by σ [12,30].

Throughout the past years, to model the first order statistics of the amplitude of the envelope of backscattered echo, various statistical distributions have been suggested. In many cases, parameters of these statistical distributions have been used to discriminate between benign and malignant breast masses. This thesis studies the use of Nakagami Distribution to model the backscattered echo envelope and formation of Nakagami image from the parameters we get from Nakagami Distribution. These parametric images assisted us in visualizing and analyzing the scatterer properties of breast tumors, as well as examining clinical success in identifying benign and malignant tumors.

3.5 Introduction to Nakagami Distribution

The Nakagami distribution is one of the simplest models for modelling backscatter envelope. It is a two-parameter distribution first introduced in Nakagami (1943, 1960) in the context of wave propagation. It can be viewed as an approximation of Homodyned K-distribution.

3.5.1 Nakagami Distribution as an Approximation of Homodyned K-Distribution:

The compound representation of two-dimensional Homodyned K-distribution is given below:

$$PHK\ (A|\epsilon, \sigma 2, \alpha) = 0PRi\ (A|\ \epsilon, \sigma 2w)\ G\ (w|\ \alpha,1)\ dw \qquad (4)$$

where, PRi denotes the Rice distribution which has been modulated by the Gamma distribution G (w| α,1) with mean and variance equal to α and $\sigma^2>0$,

The Nakagami probability density function is given by [12],

$$N\ (m, \Omega) = \frac{2m^m}{\Gamma m \Omega^m} A^{2m-1} e^{-\frac{mA^2}{\Omega}} \qquad (5)$$

Here, Γ represents the Euler gamma function. The real numbers:

m> 0 $is\ the\ shape\ parameter\ and\ \Omega > 0$ are the scaling parameter.

The cumulative distribution of the Nakagami distributed envelope F(r) is given by [12],

$$F(r) = \int_0^r \frac{2m^m}{\Gamma m \Omega^m} A^{2m-1} e^{-\frac{mA^2}{\Omega}} dA \qquad (6)$$

The following expressions are for the mean intensity of the Nakagami distribution and its SNR **[30]**:

$$E[I] = \Omega;$$

$$SNR^2 = m$$

This intensity SNR parameter should not be confused with the amplitude SNR.

These two parameters can also be expressed by,

$$m = \frac{[E(R^2)]^2}{E[R^2 - E(R^2)]^2}$$

$$\Omega = E\ (R^2)$$

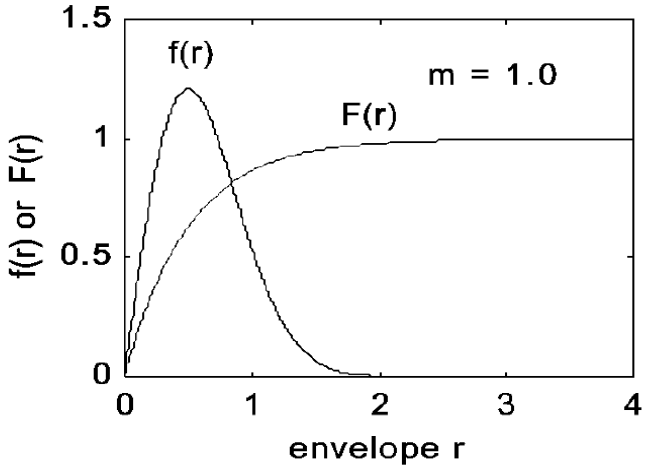


Figure 3.3: Nakagami probability density function f(r) and cumulative distribution function F(r) [12]

3.6 Generation of Nakagami and Derived Nakagami Images

Nakagami Image is the local map of Nakagami parameters generated from B-mode image where the local backscattered envelopes are fed into Nakagami distribution to get corresponding m and values. The following steps describes the procedure to generate Nakagami images:

1) Local backscattered envelopes are collected using a column window within the envelope image for estimating local Nakagami parameters, m and omega, which are allocated as new pixels in the corresponding m and omega matrices.

2) Step 1 is repeated with the window moving in one-pixel increments around the entire envelope image, yielding the Nakagami images, m image, and image as the map of local m and omega values.

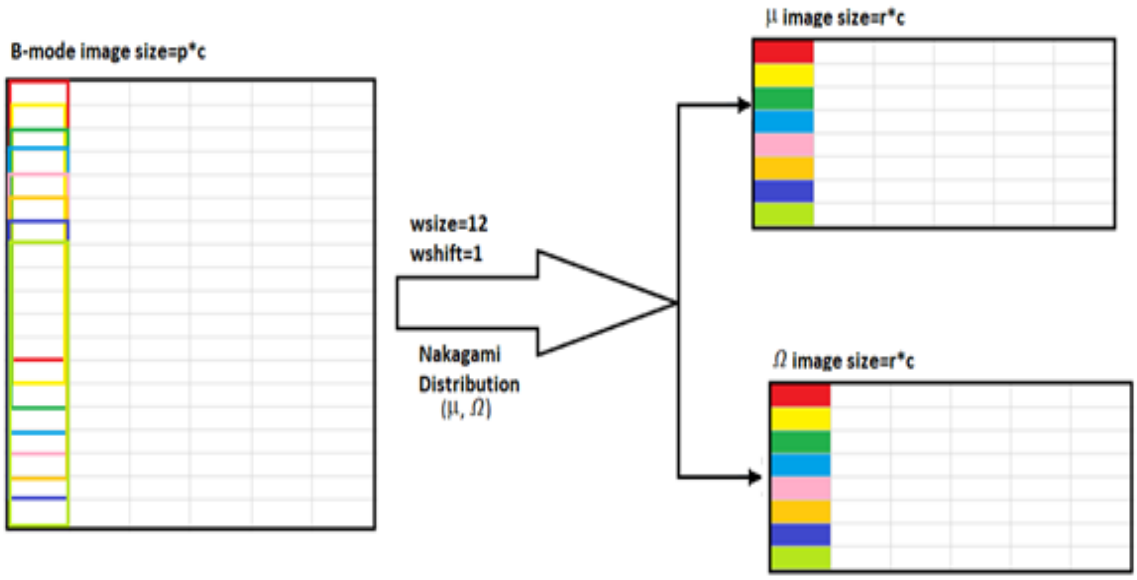


Figure 3.4: Windowing process

Derived Nakagami images namely Pre-alpha, Real alpha, Imaginary alpha, Phase alpha and Absolute alpha are generated from the mu and omega images as the local map of derived Nakagami parameters from the corresponding pixel locations of the mu and omega images.

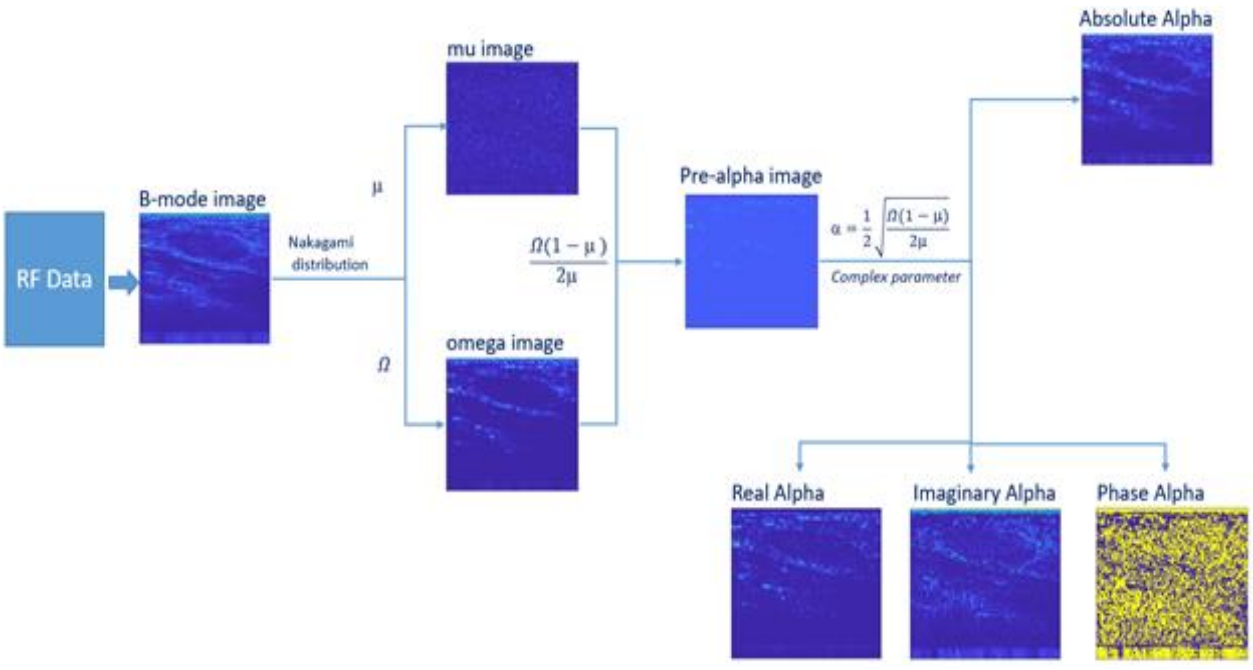


Figure 3.5: Parametric image generation flowchart

In order to conduct empirical analysis, three different datasets are generated using three different sizes of column windows: window size 5 pixels (0.1875 mm), window size 12 pixels (0.45 mm) and window size 20 pixels (0.75 mm). Empirical analysis is conducted to determine the best size of column window for classification by observing accuracy and area under curve (AUC) for varying window sizes.

3.7 Description of Nakagami Images

The parameters that have been used in this study for analysis through parametric image formation can be divided into two types:

1) Nakagami Basic Images

2) Nakagami Derived Images

3.7.1 Nakagami Basic Images:

A description of the parameters is provided below:

1. Mu Image: local map of the parameter m is known as Shape parameter and gives information about envelope statistics. Its value is such that,

$$m \geq 0.5$$

It is also a measure of the degree of heterogeneity or lack of homogeneity in the range cell. The more degree of homogeneity in the tissue, the closer the value of m to 1[12].

2. Omega Image: local map of the parameter Ω is known as Scaling parameter.[12]

3.7.2 Nakagami Derived Images:

We determined the Nakagami parameters by fitting the backscattered envelope in Nakagami distribution and derived some parameters from these Nakagami parameters, m and Ω. Alpha parameter, α is known as Alpha parameter or the effective cross section of scatterers in the range cell and is given by,

$$\alpha = \frac{1}{2}\sqrt{\frac{\Omega(1-m)}{2m}}$$

It provides information about the degree of variation in the amplitude of cross sections and on the level of attenuation in the range cell. [12] It is a complex value. The derived Nakagami images are generated calculating the local values of the derived Nakagami parameters:

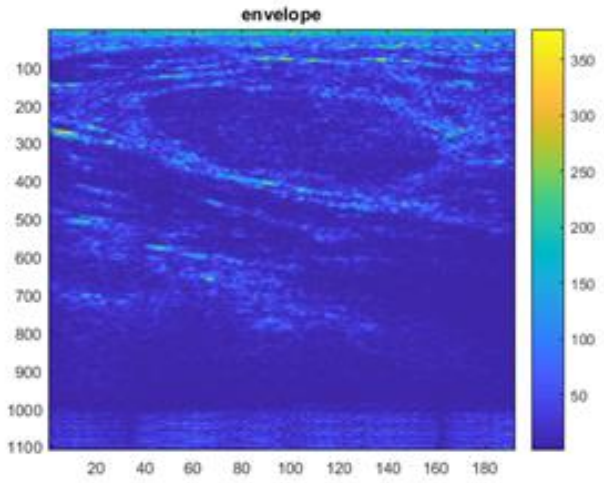
1. Pre-alpha Image: local map of the parameter pre-alpha which is a derived parameter and is given by,

$$\text{pre-}\alpha = \frac{\Omega(1-m)}{2m}$$

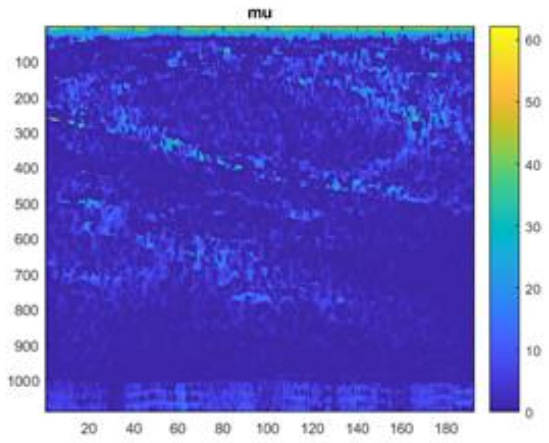
2. Alpha-absolute Image: local map of the absolute value of the Alpha parameter.
3. Alpha-phase Image: local map of the phase value of the Alpha parameter.
4. Alpha-real Image: local map of the real part of the Alpha parameter.
5. Alpha-imaginary Image: local map of the imaginary part of the Alpha parameter.

Nakagami images and derived Nakagami images

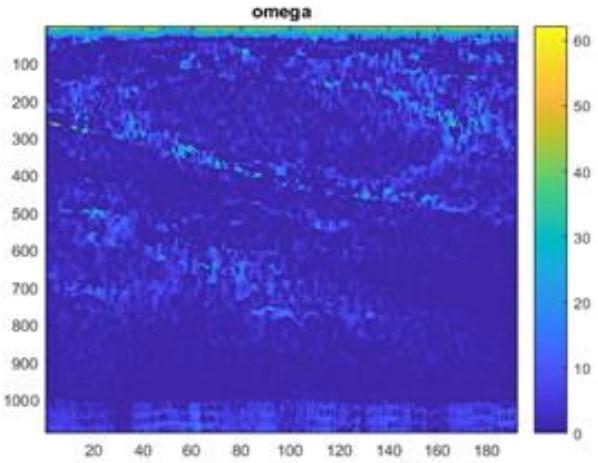
(a)



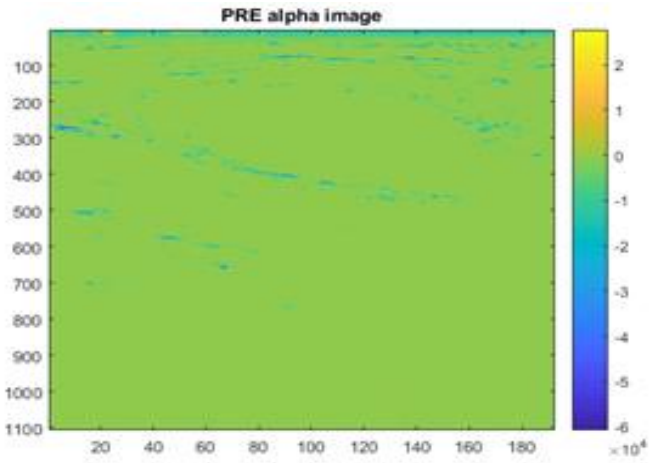
(b)


mu
100
200
300
400
500
600
700
800
900
1000
20
40
60
80
100
120
140
160
180
60
50
40
30
20
10
0

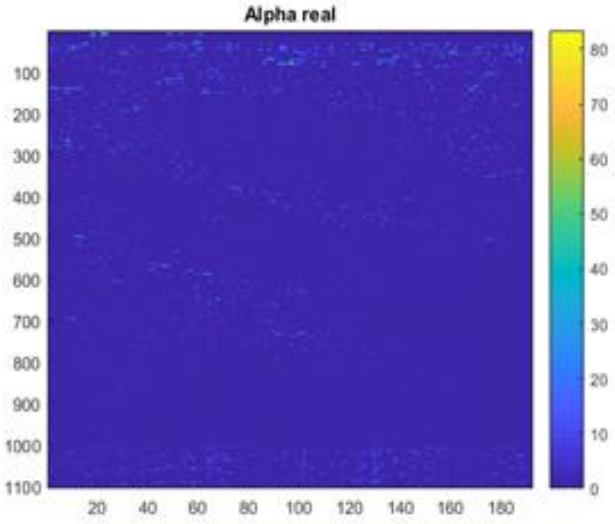
(c)


omega
100
200
300
400
500
600
700
800
900
1000
20
40
60
80
100
120
140
160
180
60
50
40
30
20
10
0

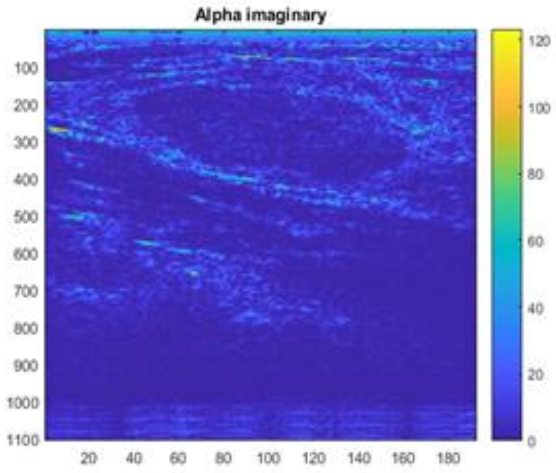
(d)


PRE alpha image
100
200
300
400
500
600
700
800
900
1000
1100
20
40
60
80
100
120
140
160
180
2
1
0
-1
-2
-3
-4
-5
-6
×10^4

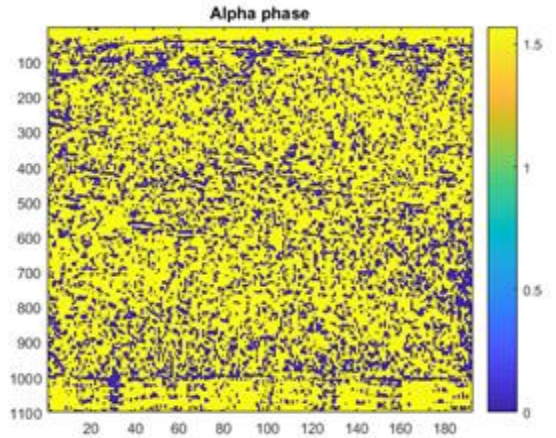
(e)


Alpha real
100
200
300
400
500
600
700
800
900
1000
1100
20
40
60
80
100
120
140
160
180
80
70
60
50
40
30
20
10
0

(f)


Alpha imaginary
100
200
300
400
500
600
700
800
900
1000
1100
20
40
60
80
100
120
140
160
180
120
100
80
60
40
20
0

(g)


Alpha phase
100
200
300
400
500
600
700
800
900
1000
1100
20
40
60
80
100
120
140
160
180
1.5
1
0.5
0

(h)

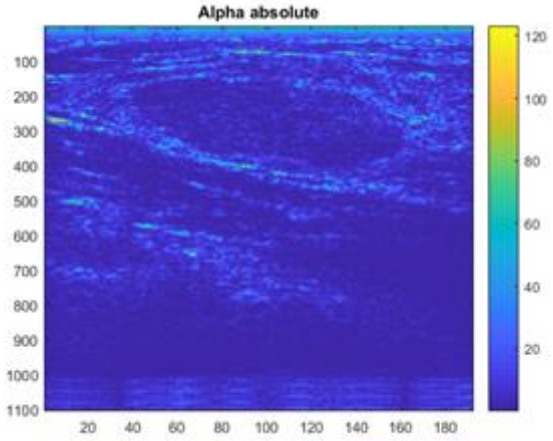


Figure 3.6: Visualization of all Nakagami images: (a) envelope. (b) mu. (c) omega, (d) pre alpha. (e) alpha real. (f) alpha imaginary. (g) alpha phase. (h) alpha absolute.

3.8 Feature Extraction from Region of Analysis

All feature processing software have been built in MATLAB^{TM} (The Mathworks, Inc., Natick, MA). The first step of the whole procedure was to obtain the frequency response of the RF data for each tumor. Then Nakagami images were generated through windowing throughout the envelope image and a multi-feature-analysis procedure was applied on the images. This procedure uses the same BI-RADS criteria (the Breast Imaging Reporting and Data system-A system developed to improve the accuracy of breast ultrasound diagnosis) followed by the clinicians.

3 types of features have been generated from the Nakagami basic and derived parameters and they are:

1) Morphometric Features

2) Acoustic Features

3) Hybrid Features

3.8.1 Morphometric Features

These features represent the shape or boundaries of the lesion. These features have been calculated based on the lesion boundaries traced on the Nakagami images. These features are described below:

a) Aspect Ratio: It is the ratio of maximum vertical lesion dimension to maximum horizontal lesion dimension. In breast carcinomas with small lesions, the aspect ratio typically reaches 0.8 [9]. It is derived from the region of the lesion as,

$$Aspect\ ratio = \frac{Maximum\ vertical\ lesion\ dimension}{Maximum\ horizontal\ lesion\ dimension}$$

b) Compactness: It is the ratio of square root of lesion area and its maximum diameter. It represents the compactness of the shape of the traced boundary. It is derived from the lesion area as,

$$Compactness = \frac{\sqrt{Lesion\ Area}}{Maximum\ Diameter}$$

c) Roundness: It is the ratio of the lesion area and its maximum diameter squared. It represents the roundness of the shape of the traced boundary. It is derived as,

$$Roundness = \frac{Lesion\ Area}{(Maximum\ Diameter)^2}$$

d) Border Irregularity: It is the ratio between border perimeter and actual lesion perimeter which quantifies the border property. This parameter is very sensitive to lesion spiculation and thus it is adept at quantifying border property. With the increase of spiculation, this parameter usually decreases and becomes very low for borders with high irregularity and spikes. It is derived as,

$$Border\ Irregularity = \frac{Border\ \ Perimeter}{Lesion\ Area}$$

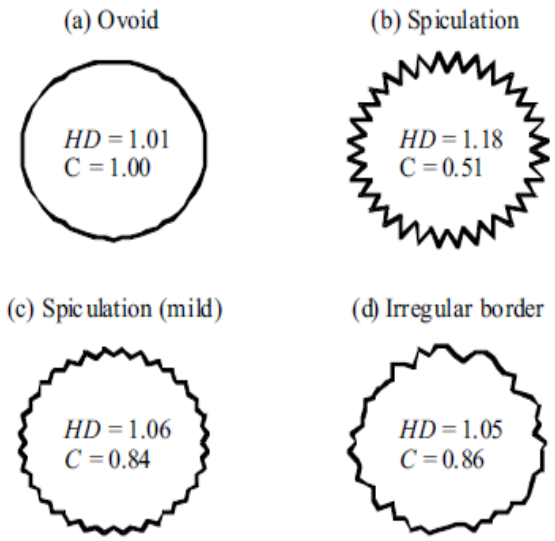


Figure 3.7: Illustration of border irregularity [9]

e) Form Factor: It is the ratio of the lesion area and perimeter squared. It is derived a

$$Form\ Factor = \frac{Lesion\ Area}{(Perimeter)^2}$$

f) Solidity: It is the ratio between lesion area and convex area. It is derived as,

$$Solidity = \frac{Lesion\ Area}{Convex\ Area}$$

g) Kolmogrov Fractal Dimension: Computes the slope of the line that plots the number of grid squares through which the lesion boundary passes versus the grid size on log-log axes.

h) Minkowski Fractal Dimension: It computes slope of the line that plots the area swept out by circles vs. their diameter on log-log axes.

i) Hausdorf Fractal Dimension: It is used to compute lesion border irregularity and roughness. Usually, as spiculation increases fractal dimension increases and vice versa.

3.8.2 Acoustic Features

Acoustic features are related to the quantitative measures of the region of interest. These features have been calculated using the pixel values in the region of interest (ROI) of parametric images. Each image is analyzed extracting various regions such as the left-anterior, left-lateral, left-posterior, tumor-anterior, tumor, tumor-posterior, right-anterior, right-laterals and right-posterior, and their pixel values are fed into various mathematical equations.

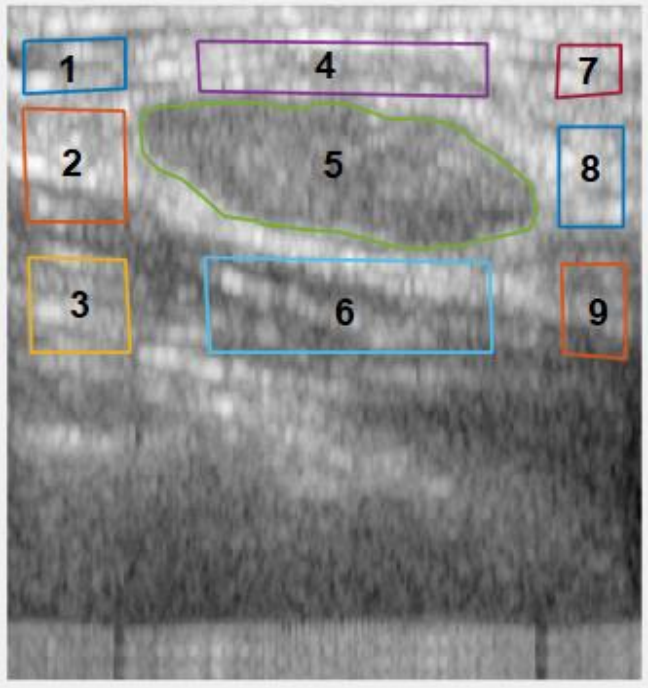


Figure 3.8: Nakagami mu image with analysis region traces superimposed: ROI(1): left-anterior, ROI(2): left-lateral, ROI(3): left-posterior, ROI(4): tumor-anterior, ROI(5): tumor, ROI(6): tumor-posterior, ROI(7): right-anterior, ROI(8): right-laterals, ROI(9): right-posterior and ROI(10): surface.

Analyzed acoustic features are described below:

a) MLP: It is the mean of the pixel intensity values within the lesion and it is represented by μ_L. For this feature no attenuation correction is necessary.

b) SLP: It is the standard deviation of pixel intensity values within the lesion and it is represented by σ_L.

c) CNC: Computation of the contrast for the co-occurrence matrix of an image.

d) FNPA: It is the abbreviation of the feature four-neighborhood pixel algorithm that computes 4-neighborhood-pixels texture of any parameter within the lesion area [9]. FNPA for an image of size $m \times n$ with pixel values x(k,l) is given by,

$$FP2 = FP1 \,/\mu$$

where,

$$FP1 = \sum_{l=1}^{n} \sum_{k=1}^{m} \frac{1}{4}[|x(k,l) - x(k-1,l)| + |x(k,l) - x(k+1,l)| + |x(k,l) - x(k+1,l)|]$$

And, μ is the mean value of FP_1.

e) HTC: It computes the fractal dimension coefficient that characterizes surface roughness in an image. It uses 7x7 sub-images for the computation.

f) Shadow: It's the difference between mean nakagami parameter values in comparable shadowed and unshadowed regions posterior to the lesion (normalized by lesion thickness). The average of the variations between left-lateral and left-posterior and right-lateral and right-posterior is compared to the difference between tumor and tumor-posterior. It is defined as follows:

$$SN = M_{pl} - 1/2(M_{pn} + M_{pl})$$

where, Mpl is the mean Nakagami parameter posterior to lesion, Mpnr and Mpnl are the mean Nakagami parameters in normal tissue right and left lateral posterior to lesion respectively [9].

g) Relative-Absorption: It is a compounded feature and given by,

$$RA = \frac{1}{d1}(M_{pn} - M_{an}) - \frac{1}{d2}(M_{pl} - M_{al})$$

where, Mal is the mean Nakagami parameter inside the lesion, Mpl is the mean Nakagami parameter posterior to the lesion, Man is the mean Nakagami parameter normal tissue next to the lesion, Mpn is the mean Nakagami parameter in normal tissue lateral posterior to the lesion, d2 is the distance between Mpl and Mal centroids, and d1 is the distance between Mpn and Man centroids [9].

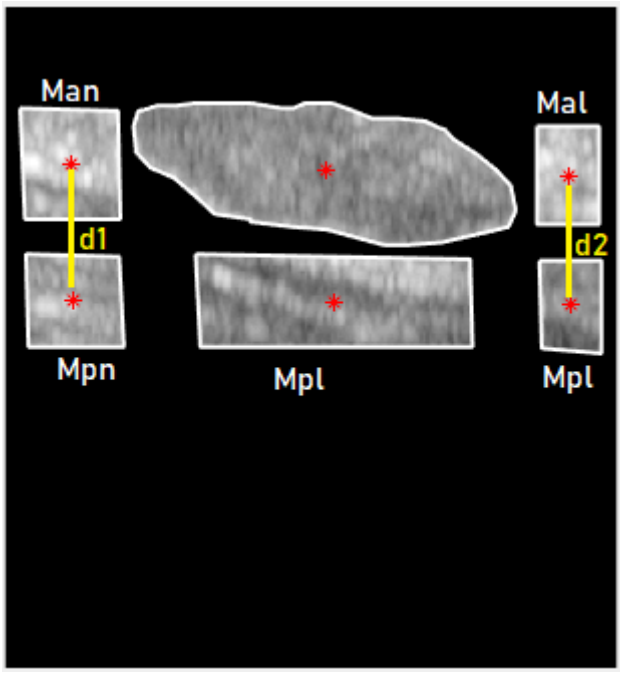


Figure 3.9: Relative absorption

3.8.3 Hybrid Features

These features are a combination of both morphometric and acoustic features. They are calculated utilizing both the lesion contour and acoustic features.To eliminate dependence on contour length and Nakagami parameter values, normalization is necessary. These features are described below:

a) MG Area: It is the ratio of the gradient around the lesion boundary of one pixel width to the area of the boundary region.

b) MG Gradient: It is the sum of the magnitudes of the lesion contour's gradients normalized by the sum of the magnitudes of the lesion contour's gradients.

c) OB: It is the ratio of the square root of the difference to the sum of the omega parameter around the lesion boundary and inside the lesion boundary.

For each patient seven Nakagami parametric images have been generated, nine morphometric features and for each image seven acoustic and three hybrid features have been calculated yielding to in total seventy-nine features, i.e., 7 Nakagami Images*(7 acoustic features + 3 hybrid features) + 9 morphometric features which totals up to 79 features.

Chapter 4
Feature Selection and Classification

4.1 Introduction

For making classification in a dataset that consists of lower number of training data and higher number of features, feature selection is an essential tool. The advantage of using feature selection algorithms on small sample set is to overcome the obstacle of overfitting data which invariably increases classifier prediction capability [44]. Feature selection refers to the technique where from a large number of features, a subset of features is selected that yield the most optimum classification [45]. It works by eliminating features that provide little to no improvement in classification and goes on excluding the weak features until the performance score falls. Fewer numbers of features not only make a dataset viable for practical use but it also allows machine learning algorithms to run in an efficient manner with less time and space complexity. A perfect example for use of feature selection is in the microarray-based cancer classification that contains several thousand features [46-47]. For this research from all the acoustic and Nakagami parameters a set of features has been found by help of feature selection algorithm that provides a very good performance score and the number of features are optimum for practical implementation of further classification of new patient data.

4.2 Feature selection methods

For using classification techniques on a supervised dataset, there are 3 types of feature selection methods: filter based, wrapper type and embedded methods [47].

1. Filter based method: This method does not work using any learning algorithm. In filter-based models, the general characteristics of data and statistical methods are used to evaluate the performance of the feature. Correlation between a set of features and the target feature and the correlation value is used to determine the ranking of the target variable and a decision is made based on whether to eliminate the feature or not [5,6]. Filter based approaches are independent of classifiers and due to low computational complexity, they are much faster than other types of feature selection algorithms [50].
2. Wrapper based method: The wrapper wraps around a learning algorithm and evaluates and selects based on primarily the accuracy of the target algorithm, but it can be also

modified to take other performance metrics as selection criteria. It excludes a subset of features from the total number of features and analyzes the impact of those feature exclusion on performance metrics using the particular learning algorithm. The features that bring significant change due to elimination are considered high quality features and on the other hand the features with less impact are eliminated.

3. Embedded method: Unlike the previous methods, embedded method is different in terms of feature selection and learning interaction. Embedded method combines the principle of feature and wrapper-based model by integrating the learning part with the feature selection part.

4.3 Basic principles of feature selection

As defined before feature selection is basically the process of selecting a subset from the original set by reduction of features. What would be the optimal number of features is calculated by certain evaluation criteria. A feature selection process contains four basic steps (as shown in figure 4.1) such as [51]:

1. Subset generation
2. Subset evaluation
3. Stopping criteria
4. Result validation

The most important step is the first one, subset generation and it consists of two basic part:

1. The search starting point can be done using one of the three ways:
   (i) Forward selection
   (ii) Backward selection
   (iii) Stepwise selection

In forward selection the process is started with no features and each feature are added one by one until the addition of new features does not change the improvement factor significantly
In case of backward selection, the process starts with all the features and each feature are removed one by one. Until the reduction of any feature decreases the performance metric drastically this process is repeated.
In stepwise selection, the process is started from both of the ends and features are added and removed at the same time.

2. The search strategy also follows 3 different routes:

(i) The complete search: Tries out all the possible combinations, guarantees to find out the most optimum subset of features. Takes longer computational time
(ii) The sequential search: Finishes searching whenever the performance criteria is met. Lacks completeness and not suitable for finding optimum subset of features.
(iii) Random search: Starts randomly with a subset of features and proceeds in two different paths.

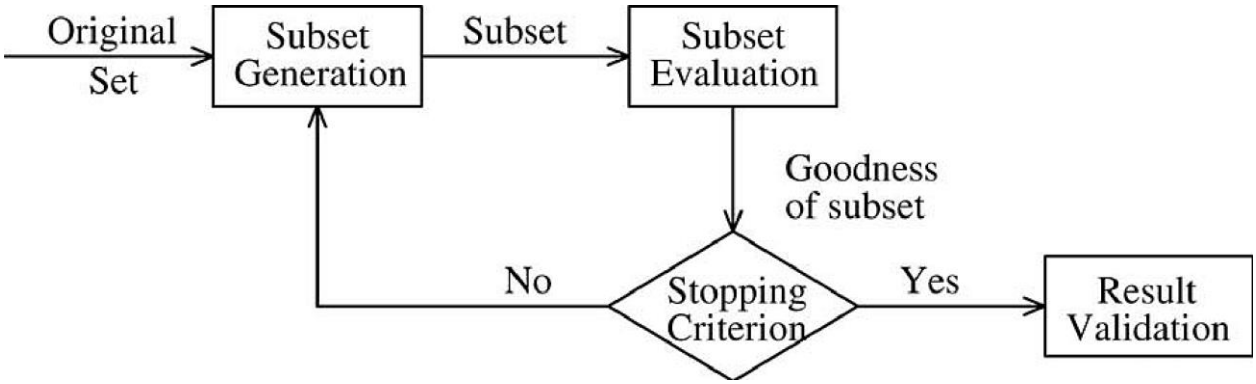


Figure 4.1 Feature Selection Process

4.4 Recursive Feature Elimination (RFE)

Among different kinds of feature selection method RFE is one the most recent ones suitable for classification of dataset with small sample size [52]. It is an ideal tool for the dataset for this research consisting of 130 patient data.

RFE is a feature selection algorithm that works by eliminating the least important features whose exclusion would have the least effect on performance metrics. It can compute the redundancy of a feature and retains only the important features for classification. It is closely related to SVM (support vector machine) which are recognized for their performance in classifying small sample dataset.

RFE is a wrapper type feature selection algorithm that uses filter-type feature selection processes at its core. It starts by searching for a subset of features taking the whole training dataset as input and discards less important features until a desired number of features remain. To do this, RFE uses the machine learning algorithm given in the wrapper model and ranks features by importance. With each iteration the model is refitted with fewer and fewer number

of features and performance is evaluated. The main advantage of RFE is flexibility by proper tuning of its hyperparameters. RFE hyperparameters that can be tuned to desired model performance:

1. Number of features: An arbitrary number of features may give good classification but is not practical. So, the desired number of features can be set as a hyperparameter of RFE and RFE will find the perfect combination of that many features with best classification. But it's not possible to know exact number of features that yield the best classification so it's better to test the algorithm by setting different number of features
2. Automated selection of features: The number of selected features can also be made to choose automatically by RFE algorithm. But for that we need RFE_CV which is a better improved version of RFE that performs cross validation evaluation of different numbers of features and selects the features with the best classification score.
3. Selected features: It is possible to know which features were selected and which were rejected by RFE after the feature selection process is done. From the support attribute of the algorithm the feature names can be obtained and from the ranking attribute ranking scores of each feature can be known. This ranking score calculation and elimination process are done inside the algorithm.

4.5 Recursive Feature Elimination with Cross-validation (RFE_CV)

The drawback of RFE is that it needs a specific number of features as input but it is not possible to know beforehand how many features would result in an optimum subset. RFE_CV uses cross-validation to score different features of subsets. From these subsets of different size, the one with the best one is chosen for classification. The mean value of different cross-validated test sets is taken as the overall score of that collection of features. As a result, RFE_CV does not require any kind of input for number of feature selection, it makes the process automatic and tries every possible combination of features. RFE_CV is more accurate than RFE in terms of feature selection as it leaves no subset untested, but it requires a greater computation time and it is much slower than the regular RFE algorithm.

4.6 Classification

A classification issue is separating the lesion into benign and malignant groups. Classification in machine learning is a two-step process, with the first step being learning and the second

being prediction. The model is built based on given training data in the learning stage, and it is used to predict the output for given data in the prediction step.

There are many popular machine learning models which are used for feature selection and classification. Some of them which we have used for our study and analysis are described in the next section.

4.7 Different Classification Algorithms

4.7.1 Logistic Regression:

Logistic regression has been named after the logistic function which is used at the core of the system. A logistic function is used to model a binary dependent variable in this statistical model. Sigmoid function, the other name for logistic function, was created by statisticians to explain the properties of population growth in ecology, such as how it rises rapidly and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can map any real-valued number to a value between 0 and 1, but never exactly between those two points.

$$1 / (1 + e^{-X})$$

Here, the base of natural logarithms (Euler's number or the EXP () function) is e, and the actual numerical value to be transformed is X. The following is a plot of numbers between -6 and 6 transmuted into the range 0 to 1 using the logistic function:

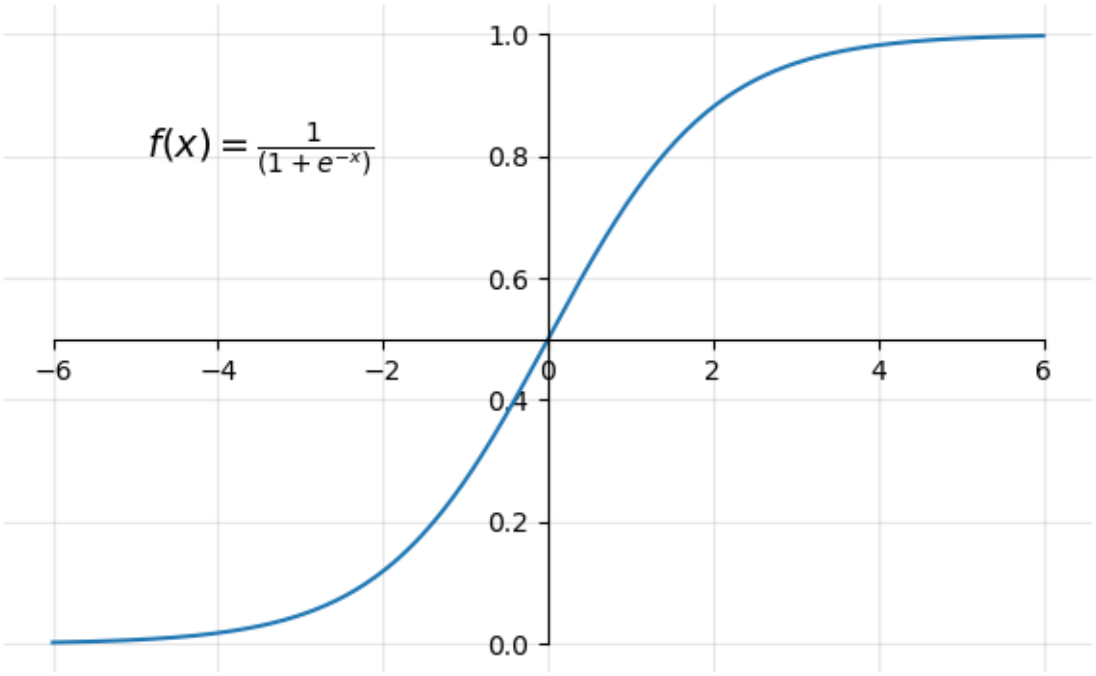


Figure 4.2: Logistic Function [32]

Similar to linear regression, logistic regression uses an equation as a representation.

To anticipate an output (y), linear combination of inputs(X) is done using weights or coefficient values (referred to as the Greek capital letter Beta) (y). The output value being modeled is a binary value (0 or 1) instead of a numeric value, which is a crucial distinction from linear regression.

An example of a logistic regression equation is shown below:

$$y = e^{\wedge} (b0 + b1*x) / (1 + e^{\wedge} (b0 + b1*x))$$

Here the predicted output is y, the bias or intercept term is b0, and the coefficient for a single input value is b1 (x). The b coefficient (a constant real value) for each column in the input data must be learned from the training data. The coefficients in the equation (the beta value or b's) are the actual representation of the model that will be stored in memory or in a register.

Logistic Regression as Predictor of Probabilities:

If we want to model the likelihood that an input (X) belongs to the default class (Y=1), we can write it like this:

$$P(X) = P(Y=1|X)$$

Though logistic regression is a linear procedure, the logistic function is used to transform the predictions. As a consequence, the predictions can no longer be understood as a linear combination of the inputs. Continuing from the previous example, the model can be stated as:

$$p(X) = e^{\wedge} (b0 + b1*X) / (1 + e^{\wedge} (b0 + b1*X))$$

From this equation it can be concluded that logistic regression is basically a linear combination of the inputs.

The following assumptions can be made on logistic regression:

- For two-class classification problems, logistic regression is used. It forecasts the possibility of an instance belonging to the default class, which can be noted as 0 or 1.
- The output variable in logistic regression is supposed to be error-free, so outliers or misclassified instances from training data should be removed.
- Logistic regression establishes a linear relationship between input and output variables.
- If there are several strongly correlated inputs, the model will overfit, much like linear regression.
- If the dataset contains several highly correlated inputs or the data is scattered, the probability estimation process can fail to converge.

4.7.2 Decision Tree

Decision tree is one of the most popular non-parametric supervised learning classification algorithms that can be used for both classification and regression. The goal of a decision tree is to learn basic decision rules from data features to construct a model that predicts the value of a target variable. The decision rules become more complicated as the tree grows deeper, and the model becomes more accurate.

A decision tree is a tree structure that looks like a flowchart, with an internal node representing a feature, a branch representing a decision law, and each leaf node representing the output.The topmost node in a decision tree is known as the root node. The tree is divided on the basis of attribute value in a recursive manner. Its representation, such as a flowchart structure, closely resembles human thought, which is why decision trees are simple to comprehend and perceive.

Decision Tree shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network and its training time is also faster compared to the neural network algorithm. It is a non-parametric or distribution-free approach and can also accurately deal with high-dimensional data.

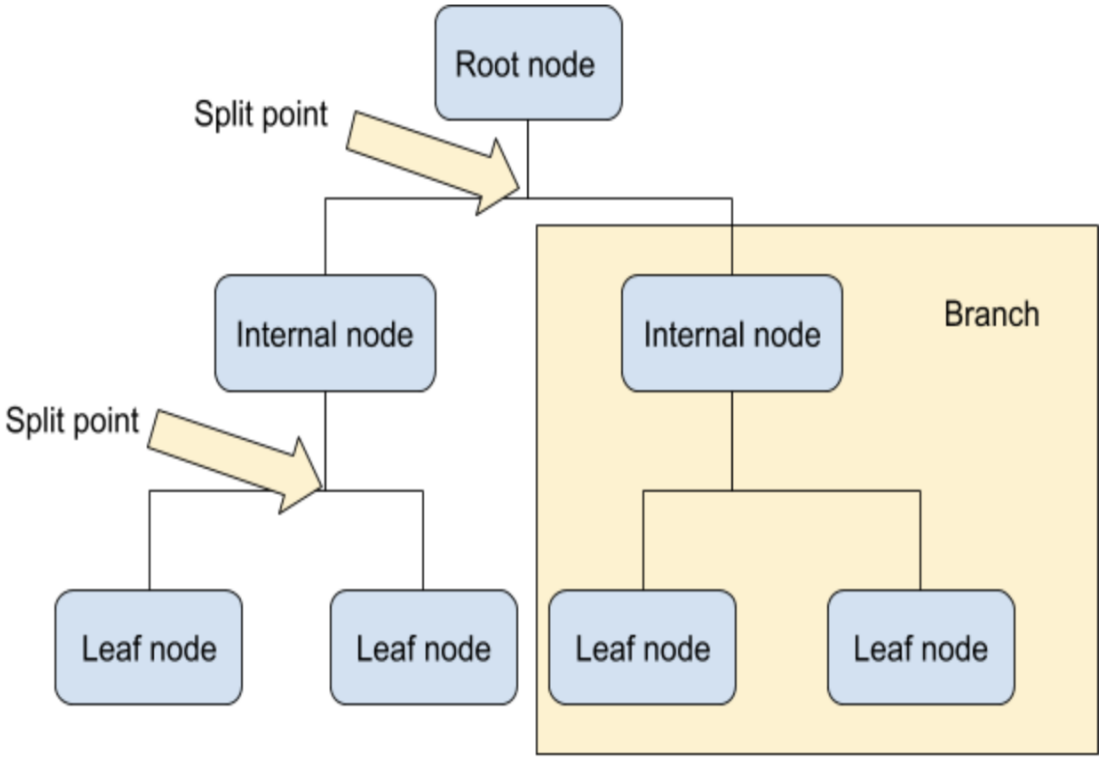


Figure 4.3**:** Flowchart of Decision Tree [33]

Decision Tree Algorithm:

The following is the basic concept behind every decision tree algorithm:

1. Attribute Selection Measures (ASM) is used to break the records and select the best attribute.

2. The dataset is divided into smaller subsets using that attribute as a decision node.

3. The tree is built by recursively repeating this process for each child until one of the conditions is met:
    - The tuples are all associated with the same attribute value.
    - There are no more attributes available.
    - There aren't any more instances.

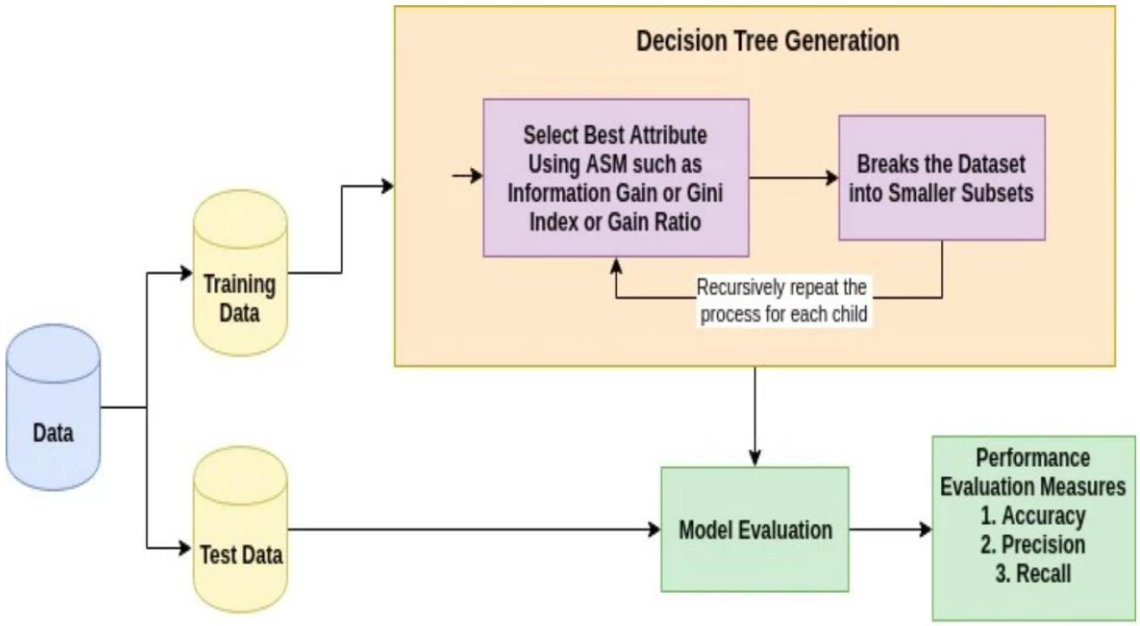


Figure 4.4**:** Decision Tree Algorithm [34]

The most popular attribute selection measures of decision tree algorithms are Information Gain, Gain Ratio, and Gini Index.

Advantages:

- Easily understandable
- Very less amount of data preparation required
- Capable of dealing with both numerical and categorical data.
- Capable of dealing with problems with multiple outputs.
- Easily explainable by Boolean algebra logic
- For model evaluation statistical tests can be done.
- Even if the true model from which the data were produced slightly contradicts its assumptions, it still performs well.

Disadvantages:

- Overly complex trees can be generated by decision-tree learners, who do not generalize the data well. This is referred to as overfitting.
- Since minor changes in the data can result in a completely different tree being created, decision trees can be unstable.
- Under many aspects of optimality and even for simple concepts, the problem of learning an optimal decision tree is considered to be NP-complete. Some principles, such as XOR, parity, and multiplexer problems, are difficult to understand because decision trees can not easily articulate them.
- If certain classes dominate, decision tree learners build biased trees and so the dataset should be balanced before using the decision tree to suit it.

4.7.3 Random Forest

Random forest is the most flexible and easy to use supervised learning algorithm used for classification and regression both. A forest is made up of trees, and it is said that the more trees there are in a forest, the more robust it is. Like this random forest is an ensemble method composed of decision tree classifiers. Random forests generate decision trees from randomly chosen data samples using attribute selectors like, information gain, gain ratio, and Gini index, extract predictions from each tree, and vote on the best solution.

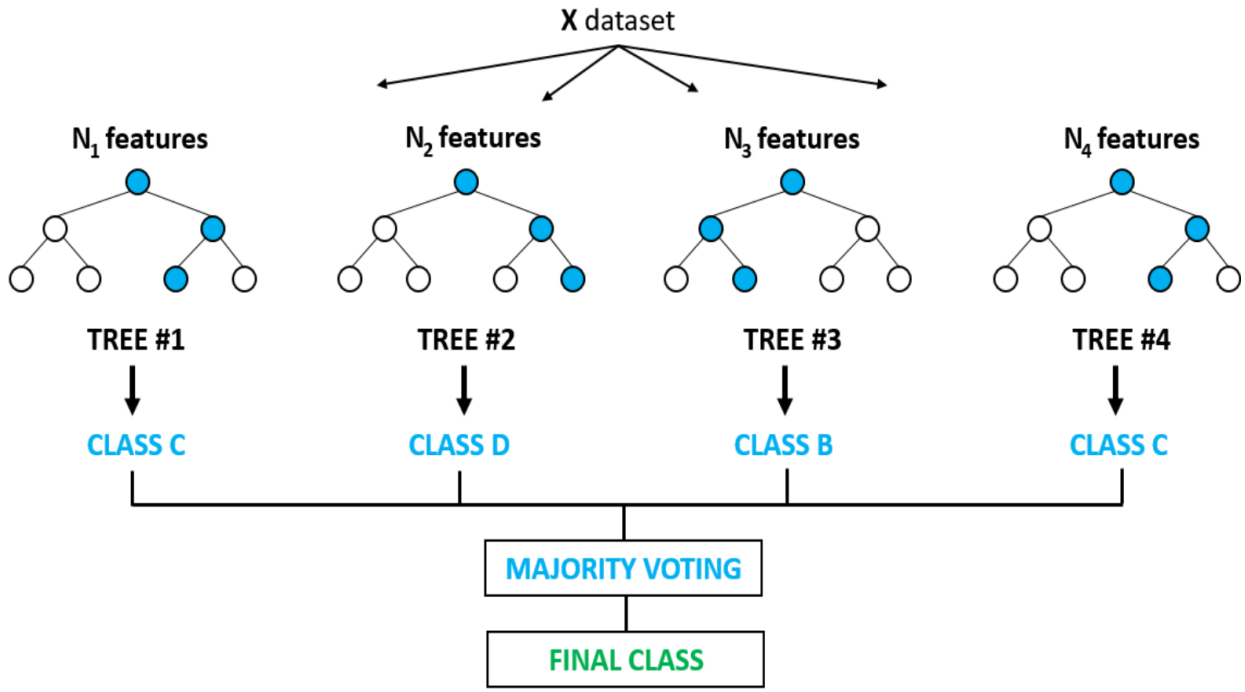


Figure 4.5: Flowchart of Random Forest [35]

Random Forest Algorithm:

The following is the basic concept behind random forest algorithm:

1. Random samples from a dataset are chosen.
2. For each sample, a decision tree is built, and a prediction result is obtained from each and every decision tree.
3. For each expected outcome, a vote is taken.

4. The prediction with the most votes is chosen as the final prediction.

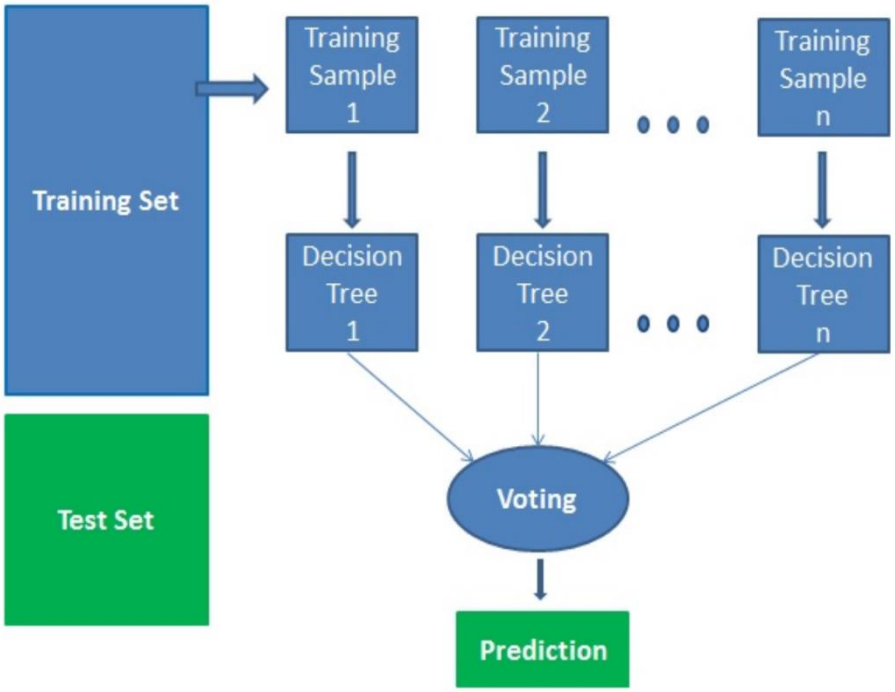


Figure 4.6: Random Forest Algorithm [36]

Advantages:

- Because of the large number of decision trees involved in the procedure, random forests are considered a highly accurate and robust system.
- It is not affected by the issue of overfitting as it averages all of the forecasts, canceling out any biases.
- Applicable for both classification and regression problems.
- Capable of dealing with missing values.
- The relative feature importance can be calculated.

Disadvantages:

- Time consuming as to generate outcome every tree has to give prediction and voting has to be performed on it.
- In comparison to a decision tree, where a decision can be taken simply by following the path in the tree, the model is difficult to explain.

4.7.4 Perceptron:

The Perceptron is a linear machine learning algorithm for two-class classification tasks. It is considered as the building block of artificial neural networks or deep learning to be specific. For binary classification tasks, it can easily learn a linear separation in feature space. It uses the stochastic gradient descent optimization algorithm to learn but does not predict calibrated probabilities.

PR Algorithm:

A single neuron in a perceptron takes a row of data as input and predicts a class label. The weighted summation of the inputs and a bias (set to 1) are used to obtain this. The activation is defined as the weighted sum of the model's input.

Activation = Weights * Inputs + Bias

The model will output 1 if the activation is greater than 0; otherwise, it will output 0.

Predict 1: If Activation > 0

Predict 0: If Activation <= 0

In models like linear regression and logistic regression, where the inputs are multiplied by model coefficients, it's a good idea to normalize or standardize data before using the model.

Since Perceptron is a linear classification algorithm, it learns a decision boundary that divides two classes using a line in the feature space called a hyperplane. The stochastic gradient descent optimization algorithm is used to train the model's coefficients, which are called as input weights.

The model is fed one example from the training dataset at a time, the model makes a prediction, and the error is calculated. According to the Perceptron update rule, weights are updated to reduce the error for each of all the examples of the training set and the process is known as an epoch. To prevent premature convergence the hyperparameter called learning rate is kept small while updating weights. Following is the equation followed to update the weights:

$$\text{weights}(t+1) = \text{weights}(t) + \text{learning\_rate} * (\text{expected\_i} - \text{predicted\_}) * \text{input\_i}$$

When the model's error falls to a low level or no longer improves, or when the maximum number of epochs is reached, training is terminated. Initial random weights are also kept small. Higher learning rates helps the model to learn fast but may cross the converging point while low learning rate gives better performance but the learning time can be lengthy.

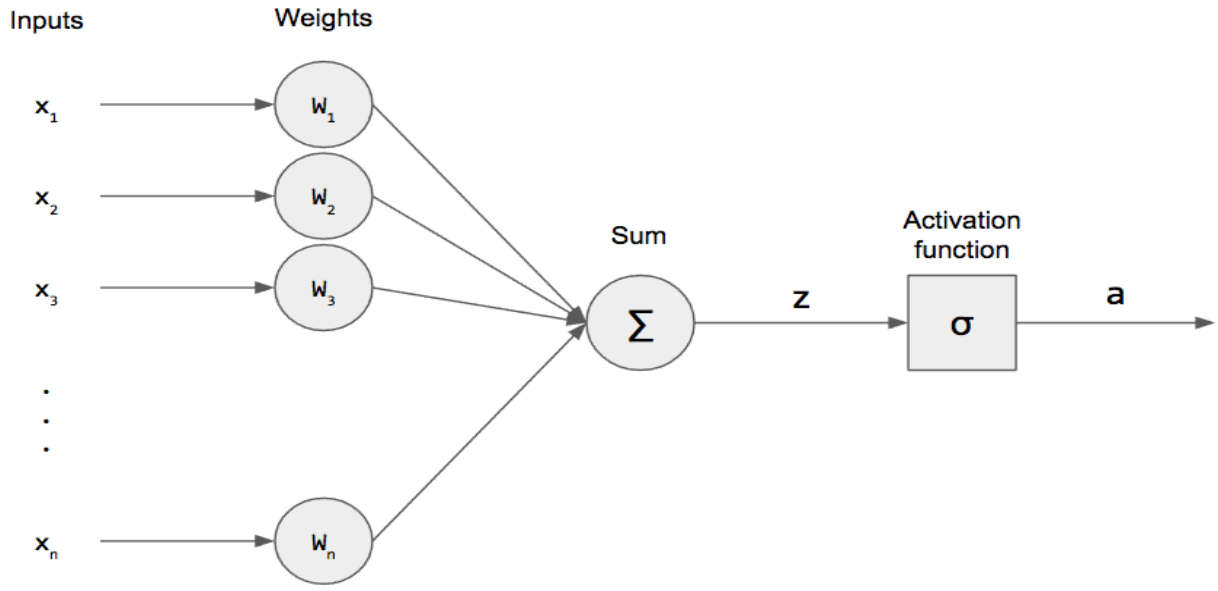


Figure 4.7: Perceptron Algorithm [37]

Advantages:

- There is no requirement for a learning rate.
- It is not regularized.
- It only updates its model when there is error. This property implies that the Perceptron is slightly faster in training, resulting in sparser models.

4.7.5 Gradient Tree Boosting

Gradient boosting-one of the most powerful techniques for creating predictive models is a variant of boosting that applies to any differentiable loss function. AdaBoost and related algorithms were recast in a statistical framework known as ARCing algorithms, and later developed as Gradient Boosting Machines, and later simply known as gradient boosting. The goal of this algorithm is to add weak learners to reduce model loss using gradient descent.

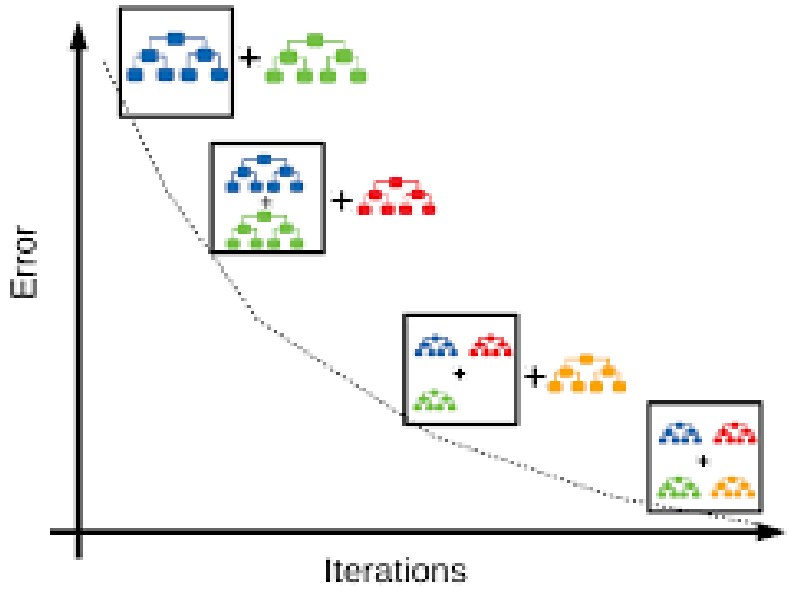


Figure 4.8: Gradient Tree Boosting Model [38]

Gradient Tree Boosting Algorithm:

The basic idea behind any gradient tree boosting algorithm is given below:

1. A differentiable loss function is to be optimized. Usually, gradient descent algorithm is used to minimize the loss function.

2. Weak learners, usually decision trees, are needed and used to make predictions. Regression trees are used specifically as they generate real values for splits and their outcomes can be added together, allowing subsequent model outcomes to be added and the predicted residuals to be corrected.
3. To add the weak learners or trees one at a time an additive model is used.

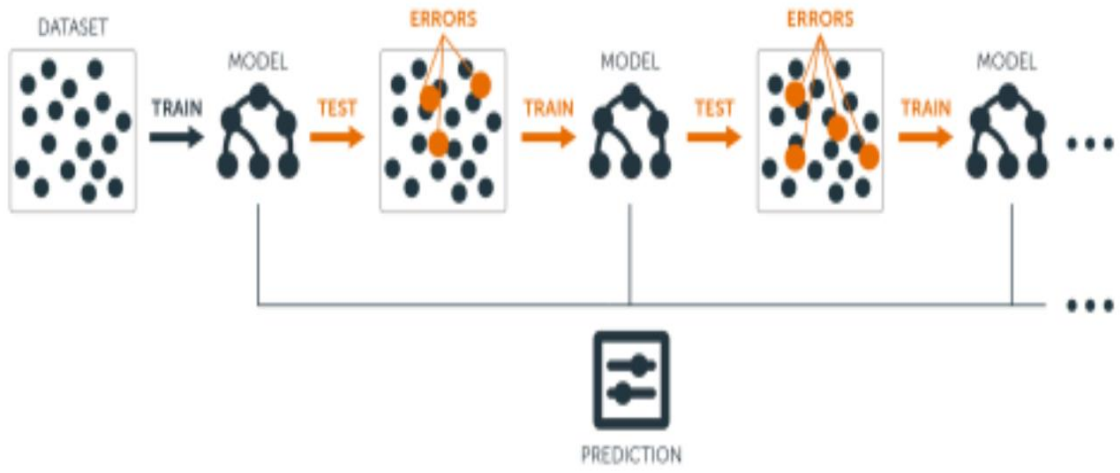


Figure 4.9: Sequential Ensemble Approach [39]

Advantages:

- Frequently provides unrivaled predictive accuracy.
- Capable of optimizing different loss functions and provides flexibility in hyperparameter tuning.
- Data preprocessing not needed.
- Often capable of dealing with numerical and categorical data.
- Capable of dealing with missing data.

Disadvantages:

- GBMs can be prone to overfitting due to its nature of improving for error minimization. To prevent this cross-validation is needed.
- Often necessitates a large number of trees, which can take a long time and consume a lot of memory.
- Tuning requires a large grid.
- Less comprehensible

4.8 Support Vector Machine (SVM)

A support vector machine (SVM) is a supervised machine learning classification algorithm that performs exceptionally well with a small amount of data to analyze. It can be used to solve both classification and regression type problems and can handle both continuous and categorical variables with ease.

To distinguish different classes, SVM creates a hyper-plane in multidimensional space. It iteratively produces the best hyper-plane, which is then used to optimize an error. The aim of SVM is to find a maximum marginal hyperplane (MMH) that divides a dataset into classes as evenly as possible.

4.8.1 Support Vectors

The data points nearest to the hyperplane are called support vectors. These points, which are more applicable to the classifier's construction, describe the separating line by measuring margins.

Hyper-plane: A hyper-plane is a decision plane that distinguishes between a group of objects that belong to different classes.

Margin: The distance between the two lines on the closest class points is known as a margin. The perpendicular distance from the line to the support vectors or closest points is determined to find the margin. A greater margin between the groups is considered a decent margin, whereas a smaller margin is considered a poor margin.

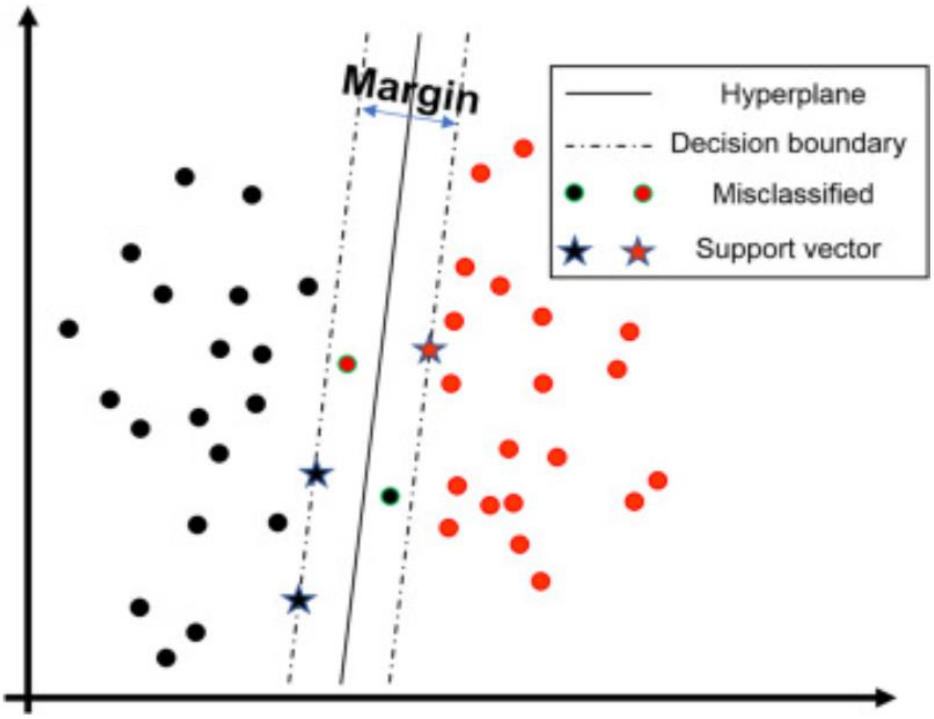


Figure 4.10: SVM Model [40]

4.8.2 SVM kernels

Kernel is used to implement the SVM algorithm. An input data space is transformed into the appropriate form by a kernel. The SVM kernel converts a low-dimensional input space into a higher-dimensional space that transforms non-separable problems into separable problems by adding more dimensions, making it particularly useful in non-linear separation problems.

SVM mainly uses three kinds of kernel:

- Linear Kernel
- Polynomial Kernel
- Radial Basis Function Kernel

As ours is a binary classification type problem, linear kernel SVM or linear SVM has been used to classify data by finding the best hyper-plane that separates all the data points of the two

classes from each other. For any two observations, a linear kernel can be used as a standard dot product. The sum of the multiplication of each pair of input values is the product of two vectors.

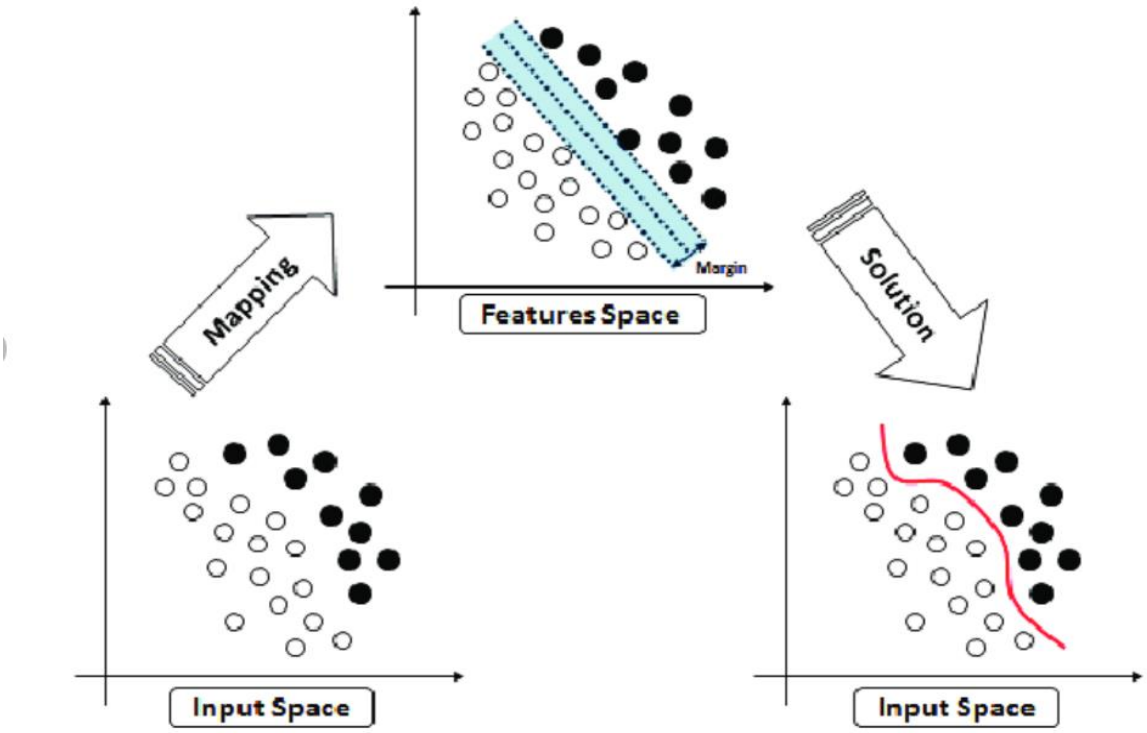


Figure 4.11: SVM Kernel Method [40]

4.8.3 SVM Algorithm

SVM's main goal is to segregate a given dataset as effectively as possible. SVM searches for the maximum marginal hyper-plane to separate the support vectors in the given dataset in the following steps:

1. It creates hyper-planes that effectively divide the groups. Figure on the left-hand side depicting three hyper-planes: black, blue, and orange. The blue and orange classes have higher classification errors, but the black class correctly separates the two classes.

2. As shown in the right-hand diagram, it chooses the right hyper-plane with the highest segregation from the nearest data points.

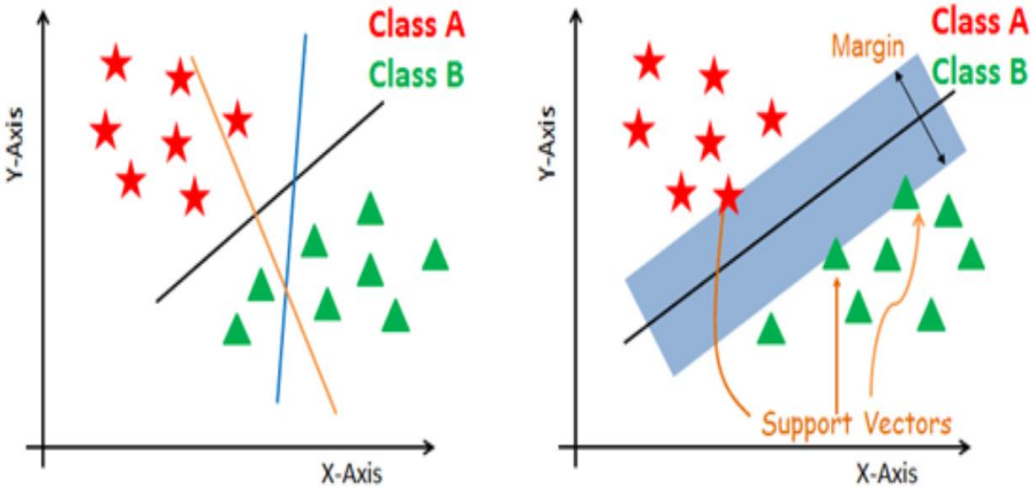


Figure 4.12: SVM Algorithm [41]

Advantages:

- Works efficiently in high dimensional spaces.
- When the number of dimensions exceeds the number of samples, the method is still accurate.
- It is memory effective since it uses a subset of training points (called support vectors) in the decision function.
- Kernels are as adaptable as they are customizable.
- Can deal with both separable and non-separable data.

Disadvantages:

- When the number of features exceeds the number of samples, it's crucial to prevent over-fitting when selecting kernel functions and regularization terms.
- Probability estimates are calculated using an expensive five-fold cross-validation method, which is not explicitly provided by SVMs.

Chapter 5
Performance Metrics

5.1 Introduction

In classification problems a classifier is evaluated based on some performance metrics after the model has been trained with some training dataset. There exists numerous performance metrics on which a classifier's performance can be evaluated and there is set rule to use any specific one [54,55]. It is very common for a classifier to perform well in some performance metrics while showing poorer results when tested on some other performance metrics [53]. A classifier has to be tested on multiple performance metrics where each depicts a unique side of the classifier's performance capability. So, to validate our result a classifier performance should be done on several different performance metrics and several datasets rather than choosing only a handful with desired outcomes.

5.2 Performance metrics for binary classifier

In a binary classification problem like this cancer stage detection in this research, four outcomes are possible:

1. True positive: The classification model declared a sample as positive class and it's verified as positive by the ground truth data
2. True negative: The classification model declared a sample as negative class and it's verified as negative by the ground truth data
3. False positive: The classification model declared a sample as positive class and it's verified as negative by the ground truth data
4. False negative: The classification model declared a sample as negative class and it's verified as positive by the ground truth data

Every classification algorithm works using a parameter called threshold t (where $0 < t < 1$), and this threshold is used decide class of any sample. Any test case with threshold value greater than a specific value is classified as certain class and any value below that is classified as the other class (in case of binary classification). The decision threshold is changed to make classification in favor of any either positive class or the negative class depending upon which class is more important should be classified with less error margin. Lowering decision threshold usually increases the FP while reducing the number of FN errors. Several kinds of

performance metrics exist in case of binary classification and the ones used in this research are discussed in details.

5.3 Accuracy measurement of prediction values

Accuracy is defined by the ratio of number of correct prediction instances by overall number of samples present in the training dataset. If the total number of test samples in a training dataset is N, then accuracy for a specific threshold t is formulated by:

$$\text{Accuracy} = \frac{TP(t)+TN(t)}{N}$$

For the dataset of this research N = 130 as the dataset consists of 130 patients among them 104 are with benign cancer which are treated as positive classes and 26 are with malignant cancer, treated as the negative class.

Although accuracy gives an overall measure for a classifier and it's the simplest one to understand, but it might not be viable in case of biomedical classification due to some class being more important than other and error margin for all the classes are not treated as same.

5.4 Confusion Matrix

Confusion matrix is a table showing TP, TN, FP and FN values and used to depict the performance of a classification model on the set of data of which the ground truth values are known. From a confusion table, all other important performance parameters can be easily calculated including accuracy. For binary classification a typical confusion matrix looks like below:

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 5.1: Confusion matrix

5.5 ROC curve

An ROC curve (receiver operating characteristic curve) is used to evaluate the performance of a classifier on different threshold levels. By changing the decision threshold, the true positive rate and false positive rate of a classification model changes. ROC curve is a graph of TPR vs FPR at various levels of classification values with different thresholds. The related terms can be defined as:

- TPR (True Positive Rate) = $\frac{TP}{TP+FN}$
- $FPR\ (False\ Positive\ Rate\ ) = \frac{FP}{TP+FN}$

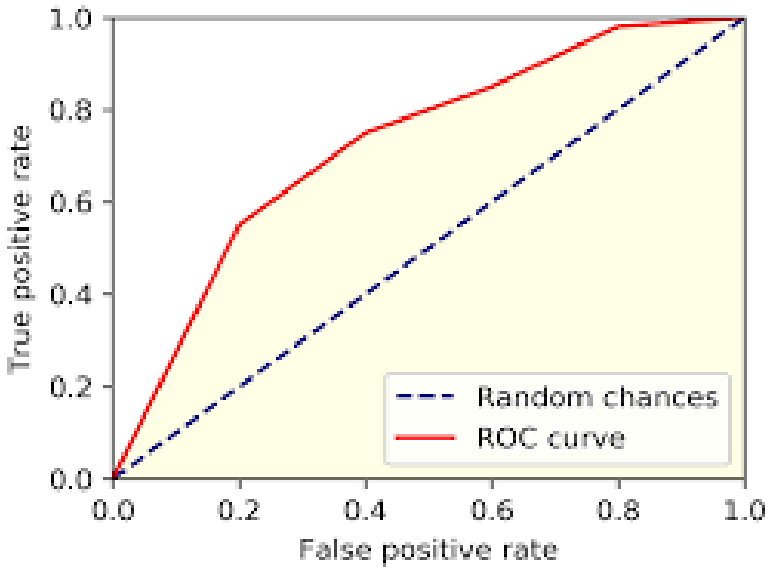


Figure 5.2: ROC curve

A typical ROC curve has N+1 distinct value for a training set of N unique samples. Because a classifier finds out N+1 threshold values and changes the threshold level between these values and plots the ROC curve. Any point on the ROC curve can be chosen and that would yield a result that the classifier is able to produce. At different point on an ROC accuracy can change.

5.6 AUC score

The measurement of an ROC curve is done by AUC (Area under curve) score. It is the entire area underneath the ROC curve and is a very important measurement for a classifier that provides an accumulated performance measure across all different classification thresholds. The better the AUC score, the better the classification is without taking the cost of error or class distribution into consideration. A higher value of AUC score indicates that even with the change in threshold value, the classification stays accurate, the TPR and FPR rate doesn't change drastically. For biomedical purposes, this AUC score is very important and for cancer cell classification typical AUC score should be higher than .95.

5.7 Cross validation

Cross validation is a resampling procedure used to justify the skill of a machine learning model on a dataset with limited data. It is a statistical method for evaluating machine learning model. It is used to estimate how well the classification algorithm will perform for an unseen data. It justifies whether a model is practically useful for new patient data which will be tested with the model and how well can it be expected to perform.

K-fold cross validation:

K is the parameter that refers to in how many numbers of individual groups the dataset is split into. The total number of samples must be divisible by K, otherwise it will result in an uneven distribution of data in each of the groups. A specific value for K is chosen, like for K=5 it is termed as 5-fold cross validation and dataset is divided into 5 sections with equal samples. This K value must be chosen carefully because a poorly chosen K value will not represent the correct cross validation result. If K value is much larger compared to a dataset or very small, it will either result in a bias or overfitting problem and the results cannot be justified.

Process of cross validation

Cross validation process follows these following steps:

1. Random shuffling of the dataset
2. Splitting the dataset into number of K sub-groups
3. For each of the subgroup, target subgroup is taken as test dataset
4. Rest of the subgroups are taken as test dataset

5. The model is trained on selected training datasets and evaluated on selected test dataset
6. The score is obtained by evaluation and the particular trained model is discarded
7. The overall score is taken as mean or average of K different runs

For this research the dataset consists of 130 patient data and 5-fold cross validation was used. So, the whole dataset was divided into five sets of 104 sample training dataset and 26 sample test datasets.

5.8 Summary

Classification results based on a single performance metric is not acceptable. To justify the results, scores of multiple performance metrics are adopted in this research and in the result section we can see the application of this performance metrics for a given model. To further justify the performance metric scores, cross validation has been used to ensure that future data can be also predicted by the model with significant assurance.

Chapter 6
Result

6.1 Introduction

Result analysis consists of two major sections, first part is feature selection using the RFE algorithm and second part is classification using a suitable classifier. Our dataset consists of features calculated with 3 different window sizes. After making classification among the same set of features an empirical analysis can also be made to know which window size is best for classification purpose. For each classification, accuracy, confusion matrix with TPR (True positive rate) and FPR (False positive rate) and the ROC curve with AUC score was observed.

6.2 Classification method

Each dataset was prepared as an excel file where 130 patient data was labeled as M for malignant cases and B for benign cases. For classification, Matlab's Classification learner application was used and the excel file was imported. The classifier app can read the label column and distinguish both of the classes, Benign classes were taken as positive classes and the Malignant classes were taken as negative classes. 5-fold cross validation was chosen and each fold the whole dataset was divided into 104 sample training set and 26 sample test set. For feature selection in RFE the logistic regression algorithm was used and for classification in Matlab the SVM algorithm was used. Every classification has 3 same steps repeated for window size: 5, 12 and 20.

6.3 Classification result without feature exclusion

The first attempt was made on the entire dataset with all the features. In this case no feature was eliminated. Even though this may result in a good classification score due to the potency of the classifier, this is not practically implementable because a total number of 79 features is not feasible to generate every time for a new patient nor it's computationally fast enough to run a real time detection algorithm. But this classification score provides an overall idea about the consistency of the whole dataset.

Window size = 5

Accuracy = 85.4%

AUC score = .79

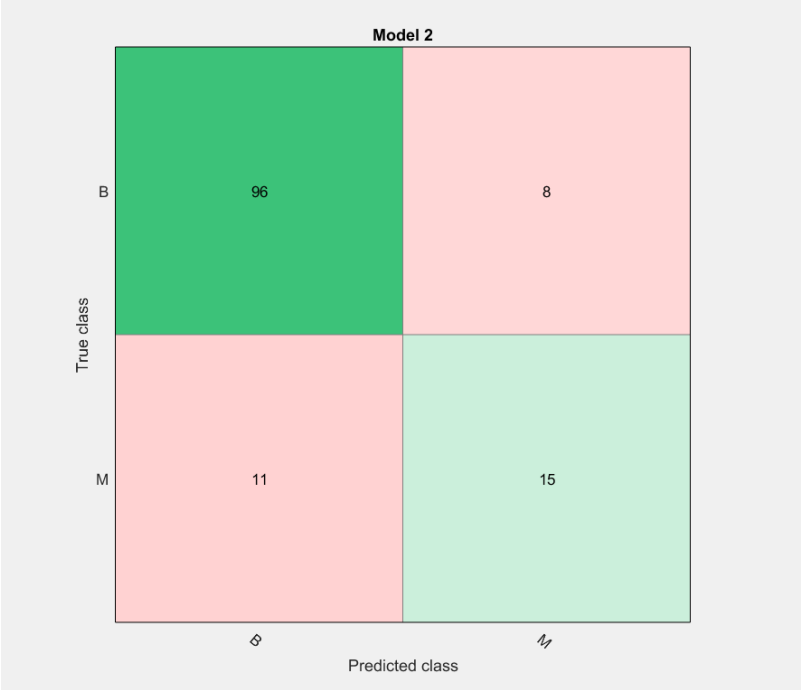


Figure 6.1: Confusion matrix with all features selected and window size = 5

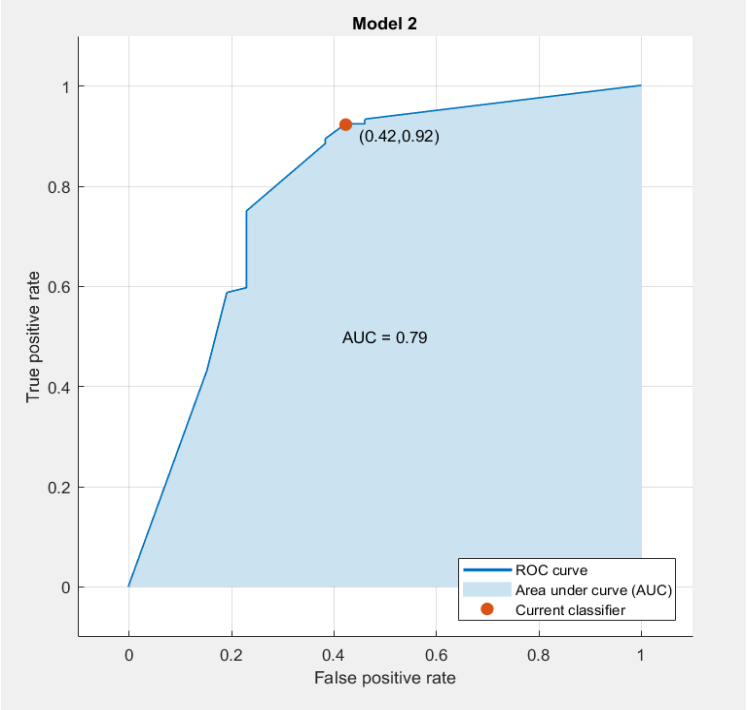


Figure 6.2: ROC curve with all features selected and window size = 5

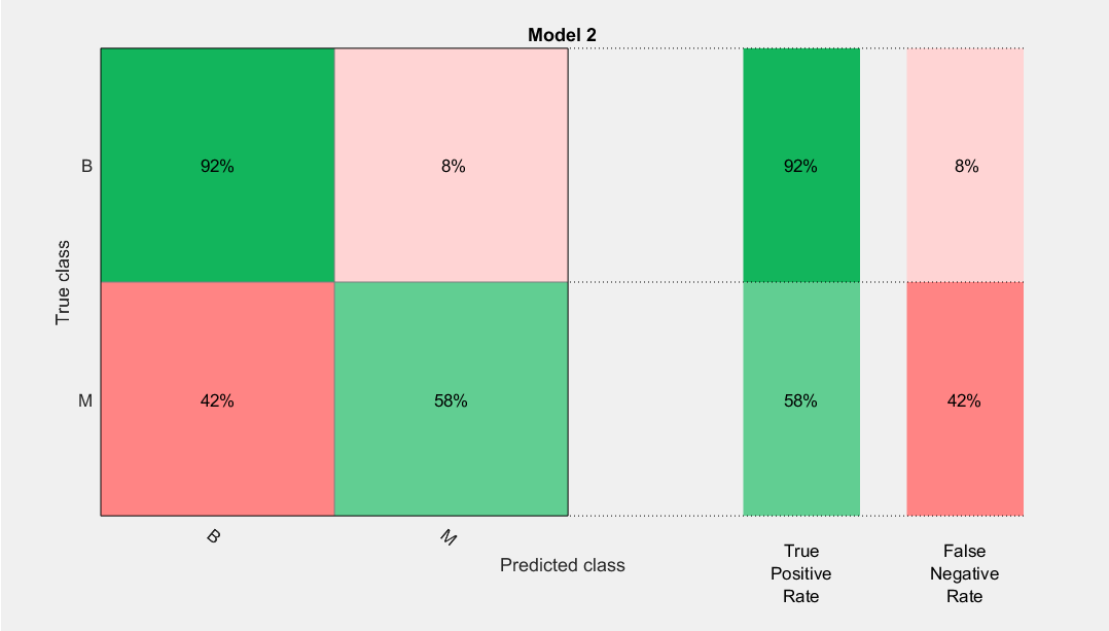


Figure 6.3: TPR and FPR with all features selected and window size = 5

Window size = 12

Accuracy = 90%

AUC score = .82

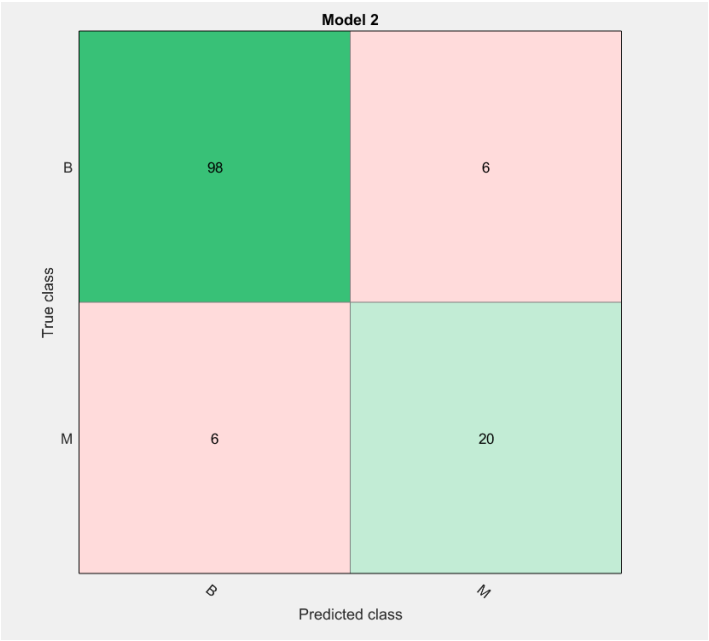


Figure 6.4: Confusion matrix with all features selected and window size = 12

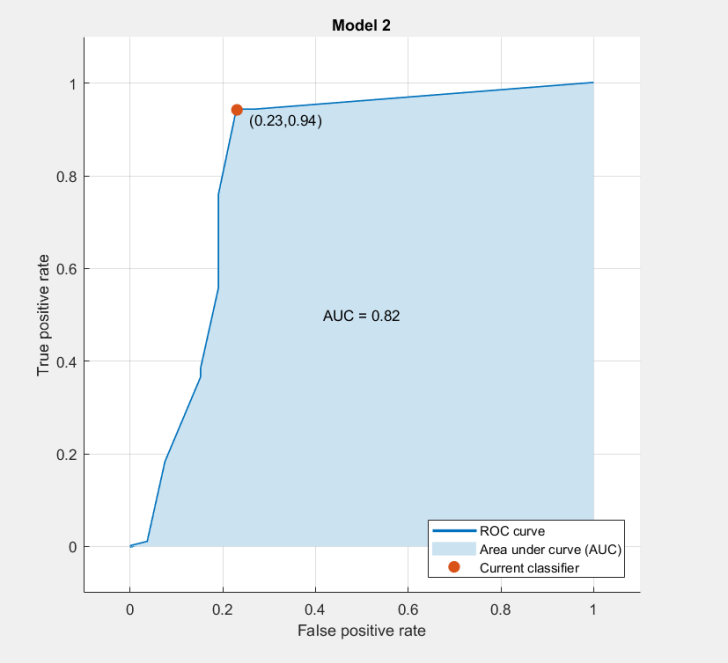


Figure 6.5: ROC curve with all features selected and window size = 12

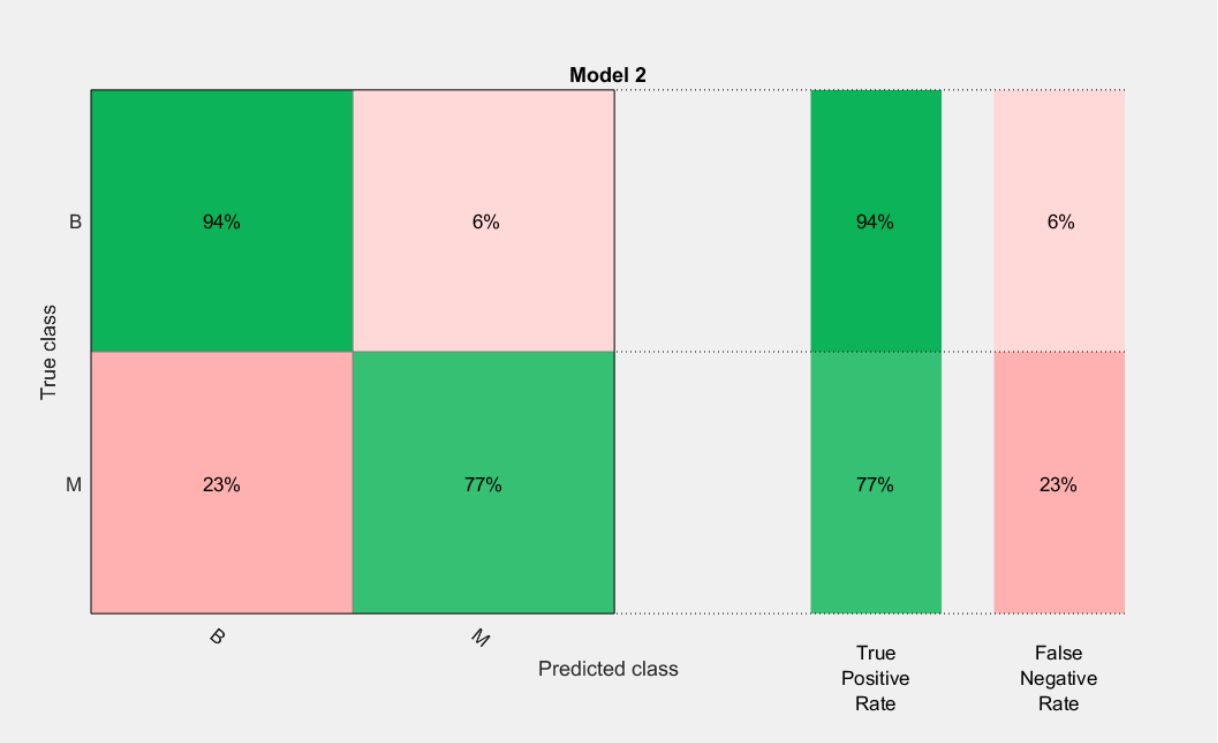


Figure 6.6: TPR and FPR with all features selected and window size = 12

Window size = 20

Accuracy = 88.5 %

AUC score = .91

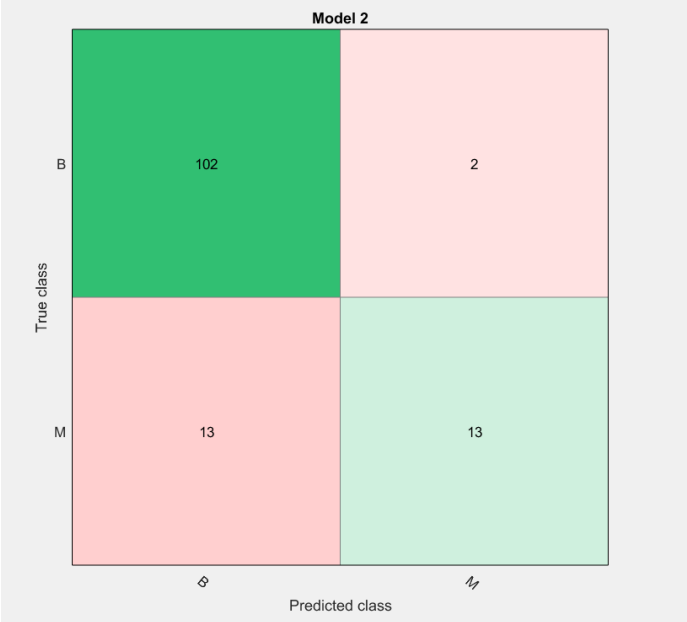


Figure 6.7: Confusion matrix with all features selected and window size = 20

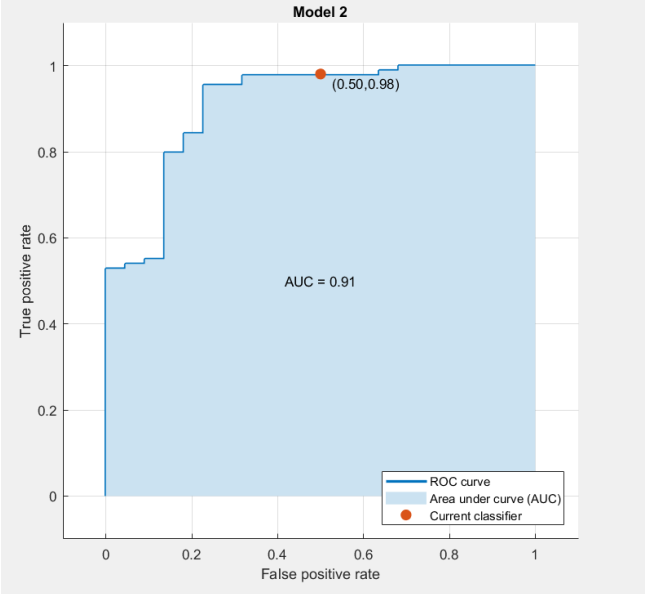


Figure 6.8: ROC curve with all features selected and window size = 20

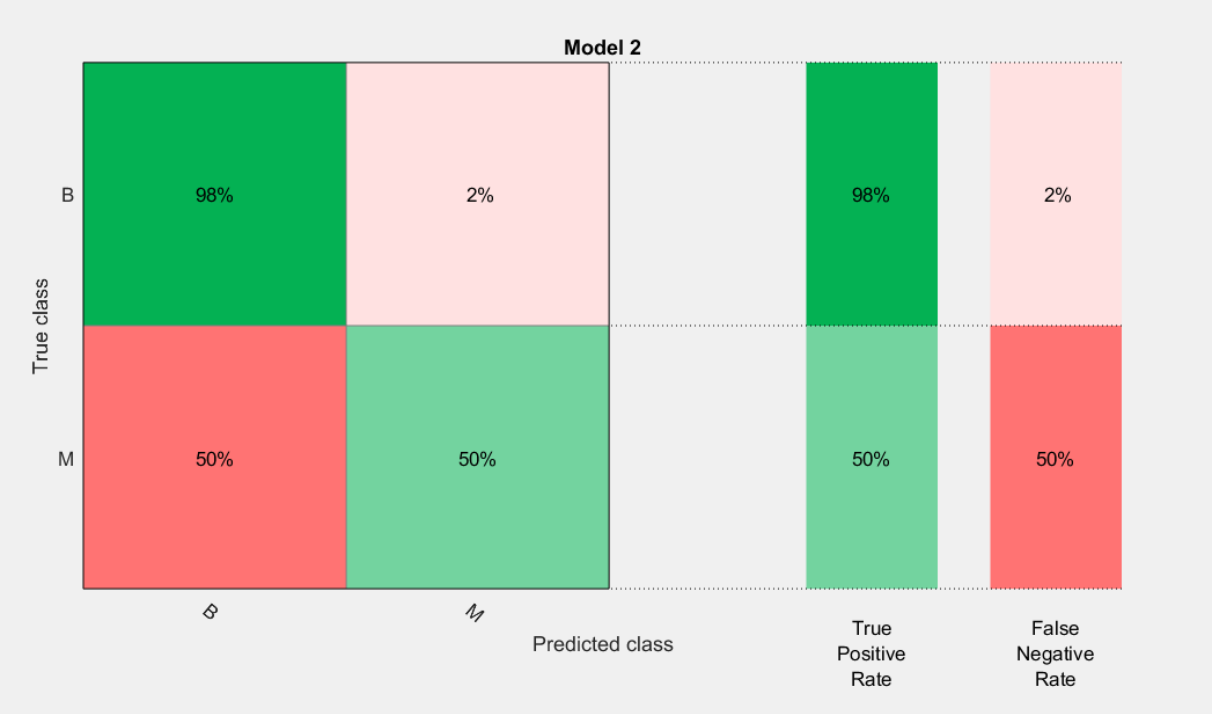


Figure 6.9: TPR and FPR with all features selected and window size = 20

Although from these results we can obtain a good classification accuracy and AUC score, but from the FPR and FNR rate we can see that a lot of malignant classes has been misclassified as benign which is very sensitive and it's even better to get lower accuracy where a classifier misclassifies a benign as malignant.

6.4 Classification result using RFE

To get an improved result and fewer features we now use the RFE algorithm. For RFE algorithms the total number of features has been set to 11. These selected features are: i) compactness, ii) border irregularity, iii)img_htc_nd, iv)img_gradient, v)ph_gradient, vi)real_gradient,vii)mu_cnc, viii)mu_htc, ix)mu_htc_nd, x)pre_fnpa, xi)solidity. For feature selection in the RFE algorithm, logistic regression was used.

Like before the same analysis has been done on these datasets of 11 features from 3 different window sizes.

Window size = 5

Accuracy = 83.8 %

AUC score = .89

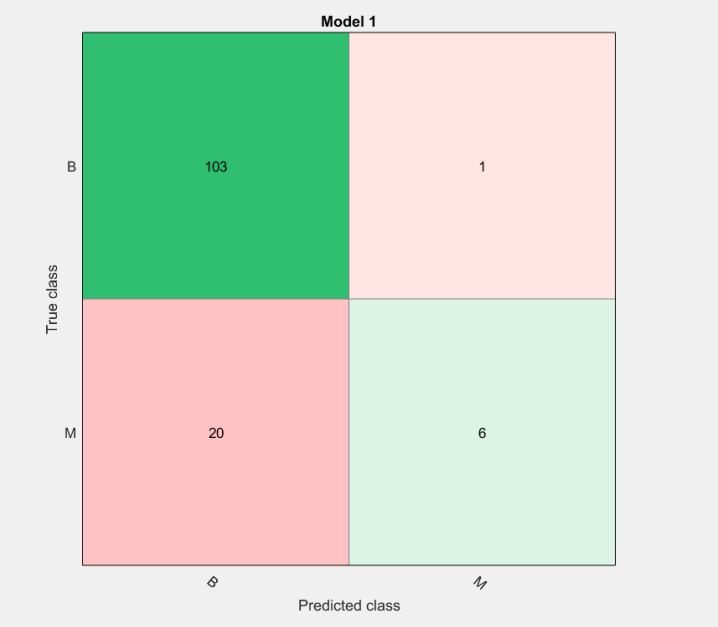


Figure 6.10: Confusion matrix with 11 features selected using RFE and window size = 5

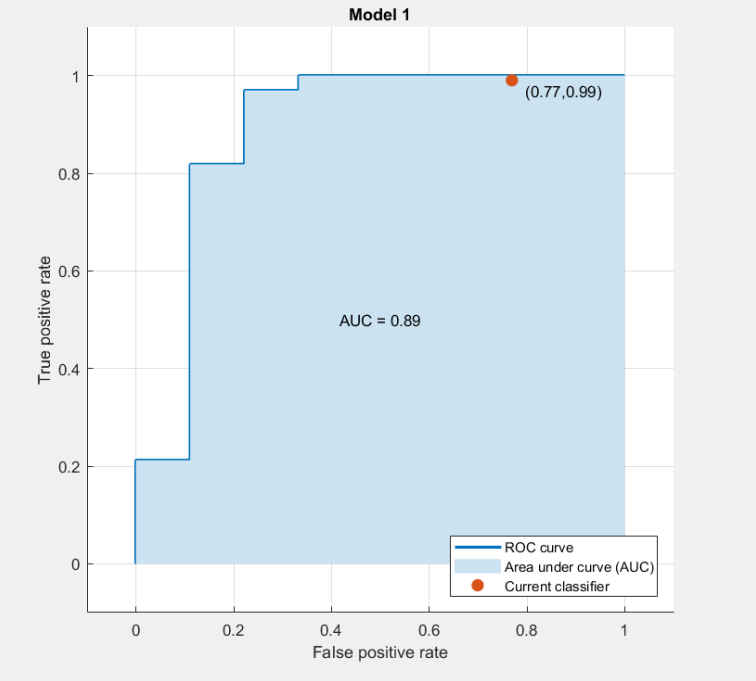


Figure 6.11: ROC curve with 11 features selected using RFE and window size = 5

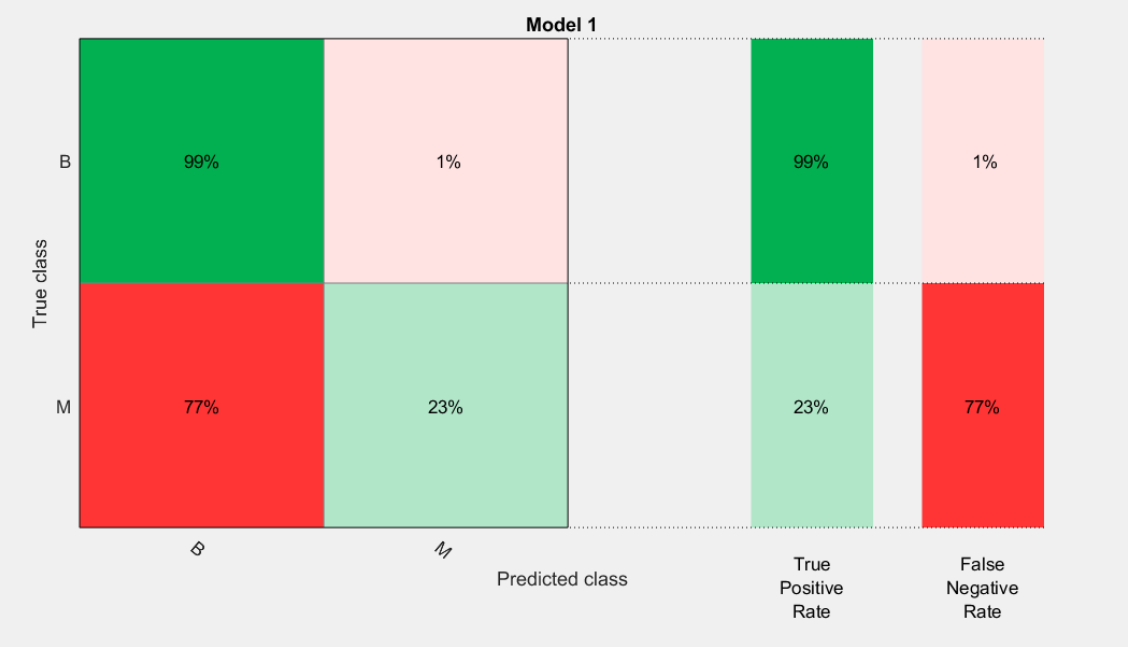


Figure 6.12: TPR and FPR with 11 features selected using RFE and window size = 5

Window size = 12

Accuracy = 90.8 %

AUC score = .89

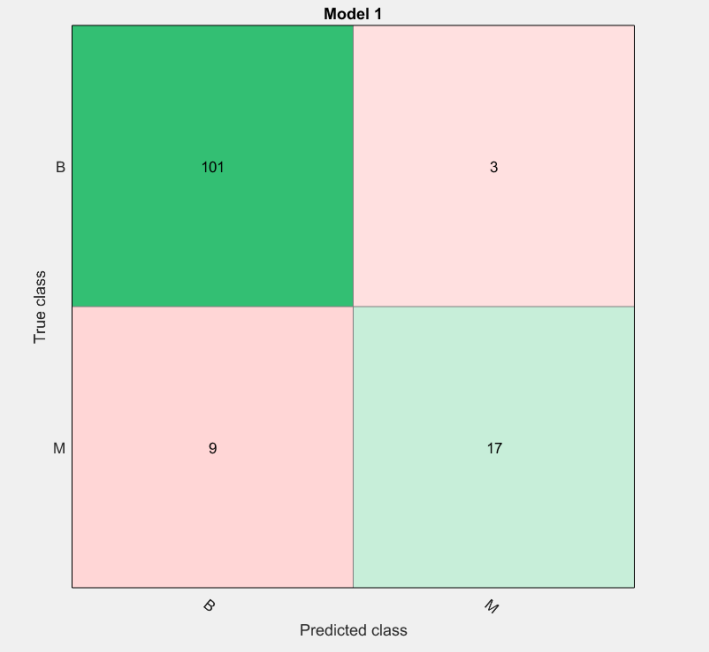


Figure 6.13: Confusion matrix with 11 features selected using RFE and window size = 12

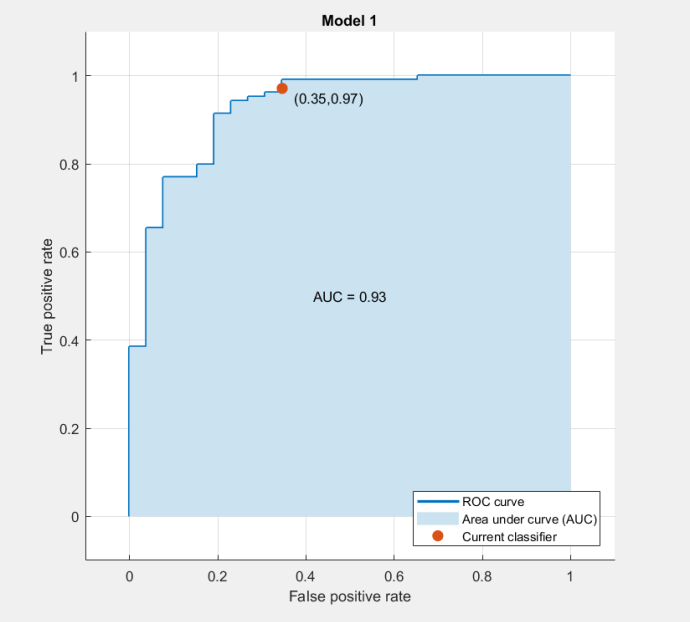


Figure 6.14: ROC curve with 11 features selected using RFE and window size = 12

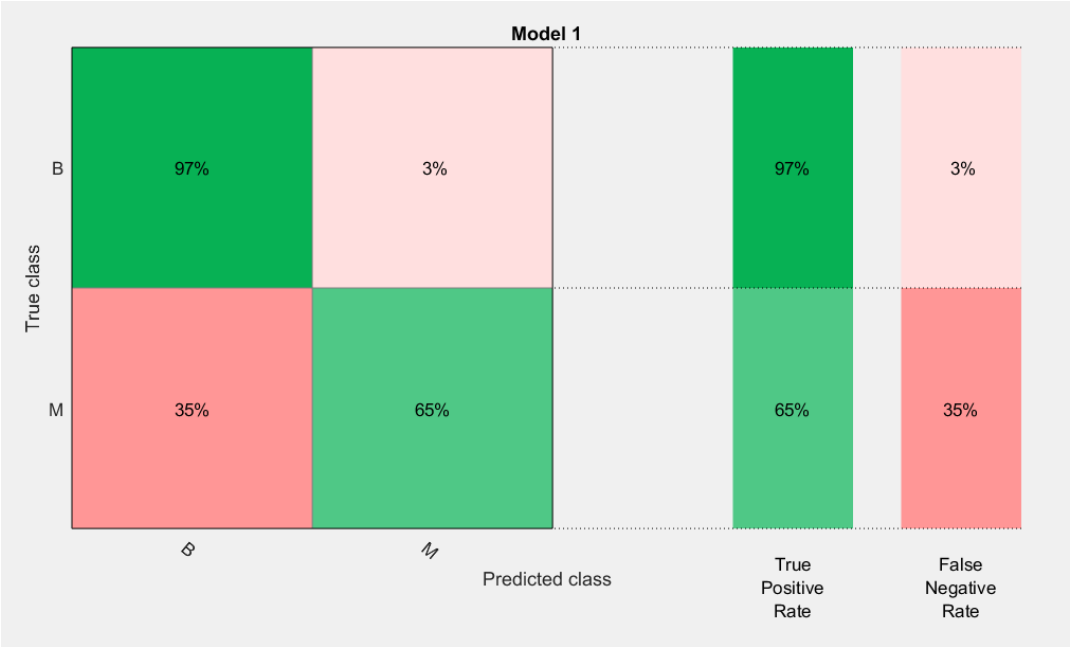


Figure 6.15: TPR and FPR with 11 features selected using RFE and window size = 12

Window size = 20

Accuracy = 90.8 %

AUC score = .94

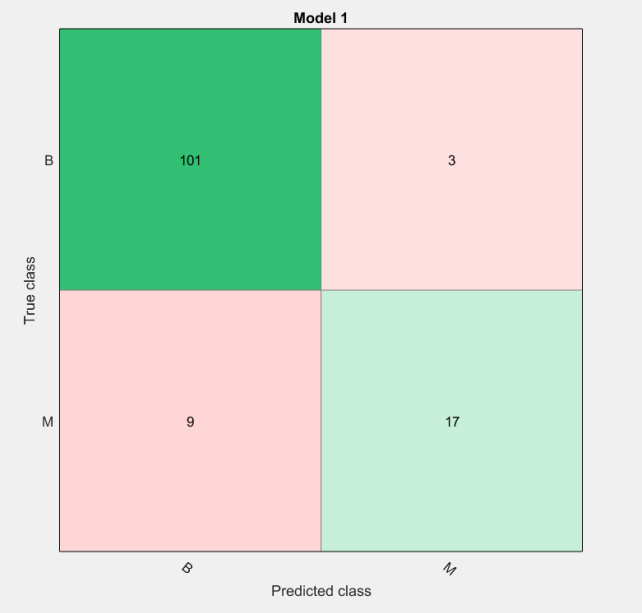


Figure 6.16: Confusion matrix with 11 features selected using RFE and window size = 20

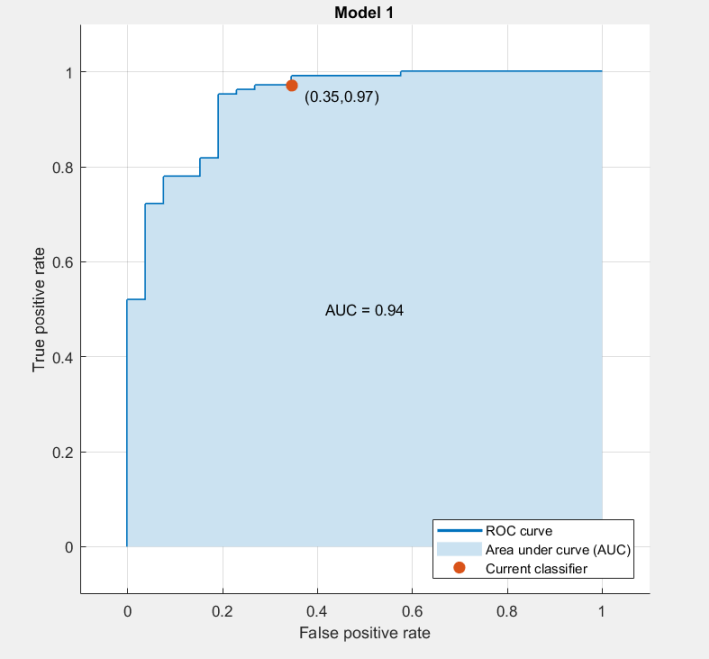


Figure 6.17: ROC curve with 11 features selected using RFE and window size = 20

Figure 6.18: TPR and FPR with 11 features selected using RFE and window size = 20

6.5 Classification result using RFE_CV

RFE_CV is an advanced algorithm that doesn't need a number of features to be selected, but it does require a scoring criterion, based on which the group of features are selected. The scoring criteria was set to accuracy. RFE_CV run on the dataset of 20 window size selected a subset of 6 features for optimum performance and are as follows:

1. Aspect Ratio
2. Border Irregularity
3. Absolute MG Area
4. Imaginary MG Gradient
5. Phase MLP
6. Mu HTC

After finding out the set of features, classification was made on Matlab Classifier Learner using the SVM algorithm. The same set of features was also selected from the dataset of window size 5 and window size 12 and was classified using SVM. Not only these results yield the best performance scores from all the previous subset of features but an empirical analysis show that with the increase in window size the performance metrics are better with the best results for window size 20.

Window size = 5

Accuracy = 87.7 %

AUC score = .88

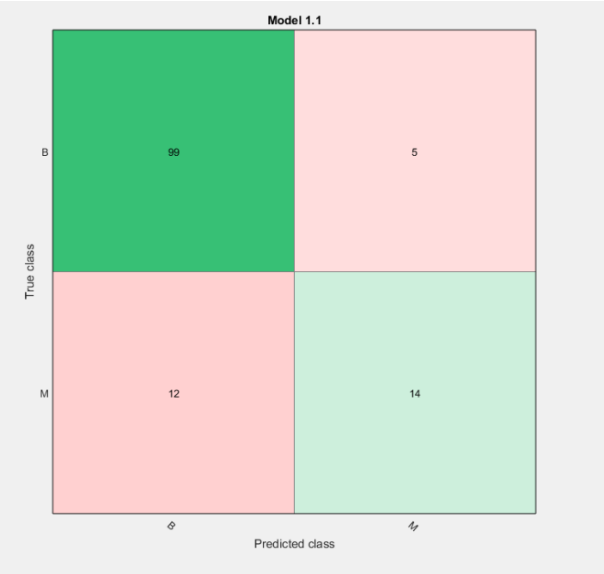


Figure 6.19: Confusion matrix with 6 features selected using RFE_CV and window size = 5

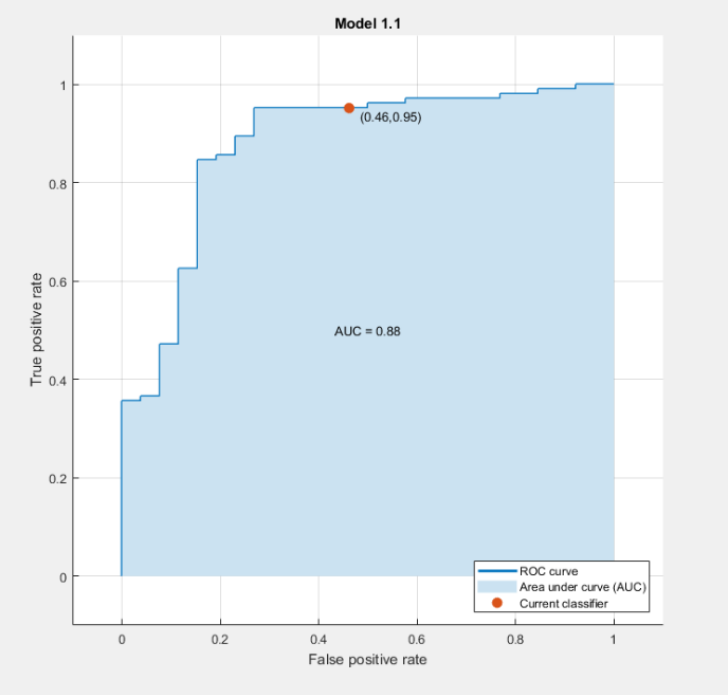


Figure 6.20: ROC curve with 6 features selected using RFE_CV and window size = 5

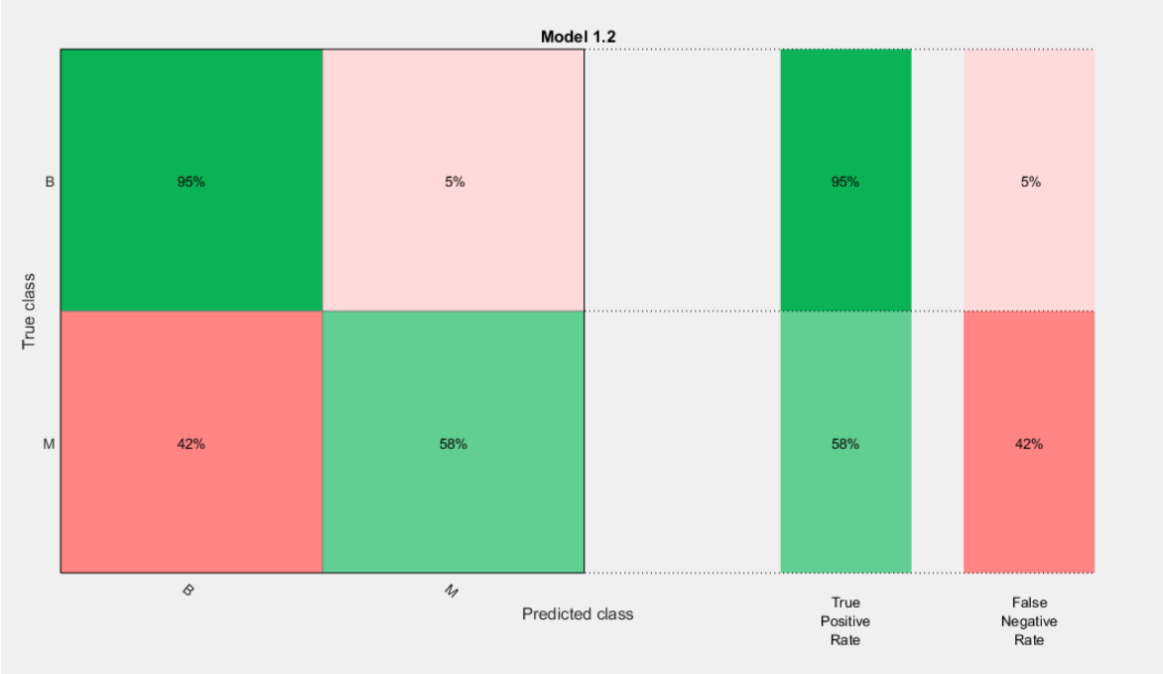


Figure 6.21: TPR and FPR with 6 features selected using RFE_CV and window size = 5

Window size = 12

Accuracy = 90%

AUC score = .93

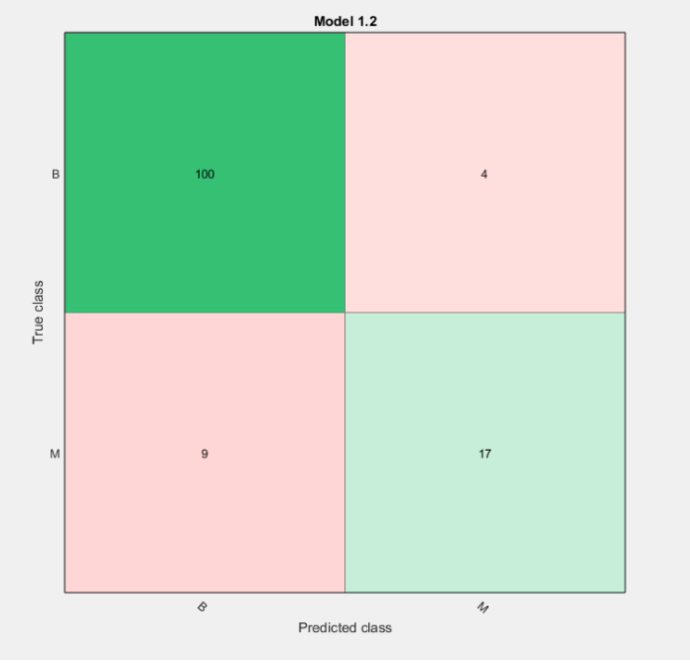


Figure 6.22: Confusion matrix with 6 features selected using RFE_CV and window size = 12

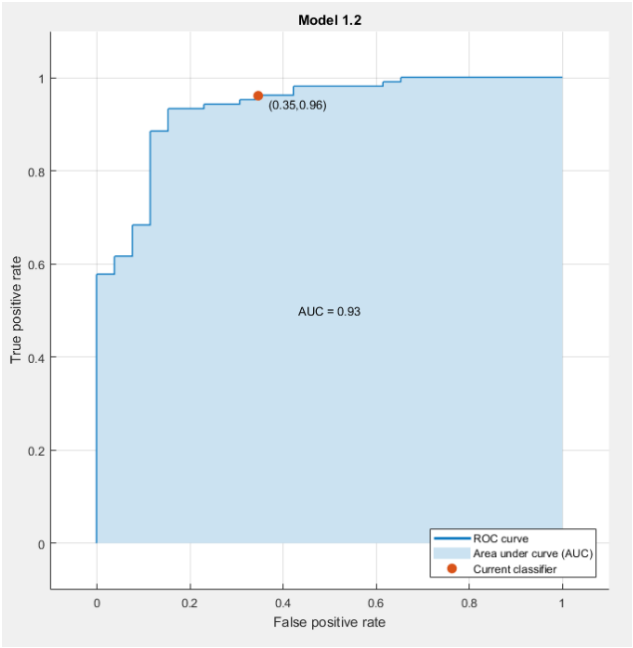


Figure 6.23: ROC curve with 6 features selected using RFE_CV and window size = 12

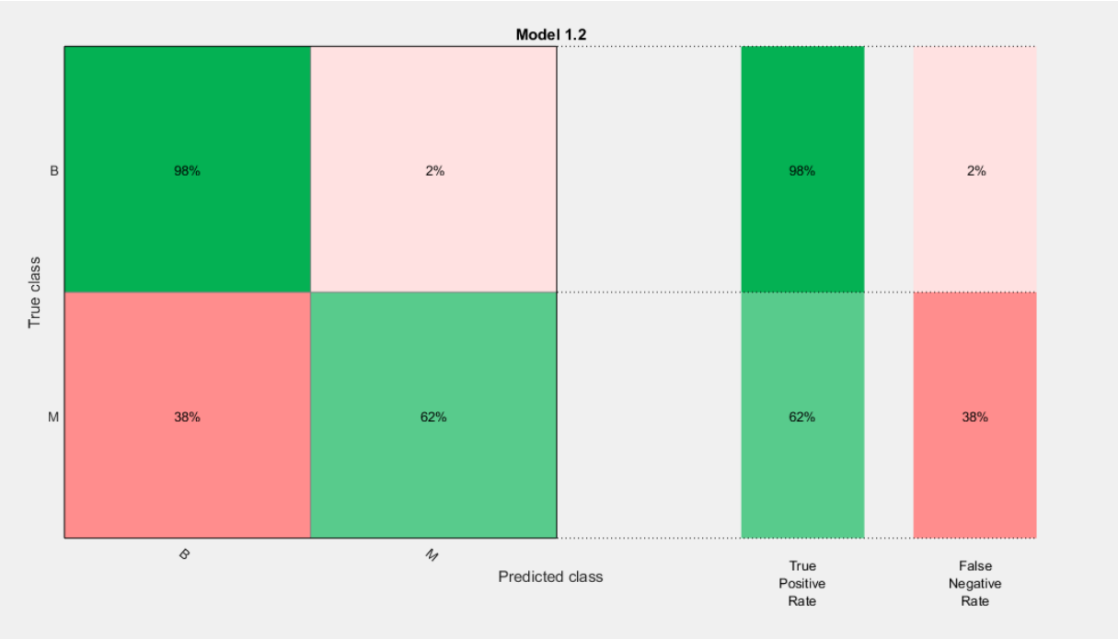


Figure 6.24: TPR and FPR with 6 features selected using RFE_CV and window size = 12

Window size = 20

Accuracy = 92.3%

AUC score = .95

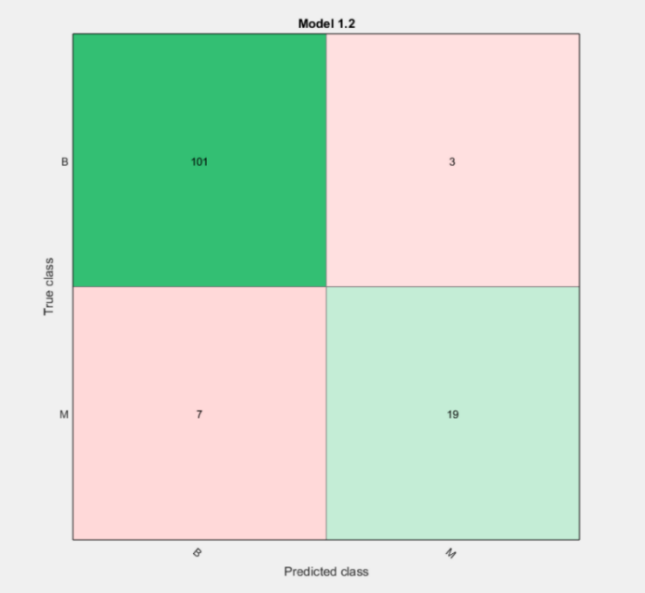


Figure 6.25: Confusion matrix with 6 features selected using RFE_CV and window size = 20

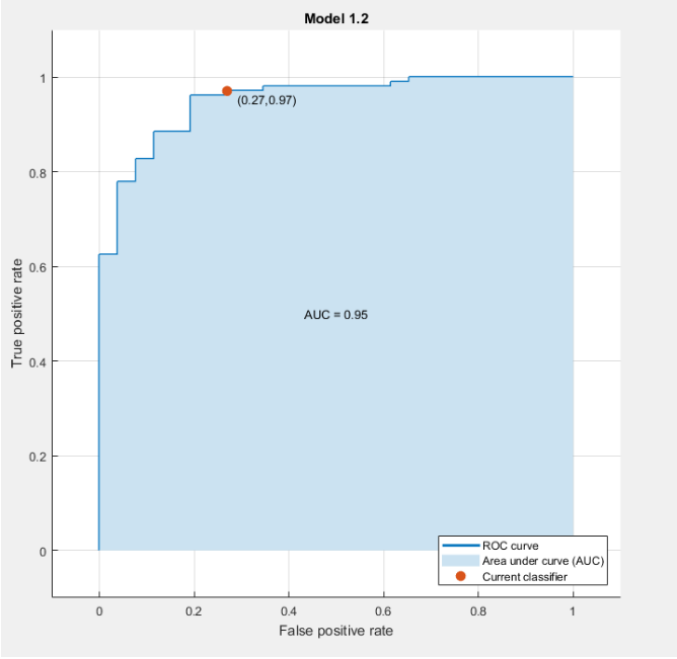


Figure 6.26: ROC curve with 6 features selected using RFE_CV and window size = 20

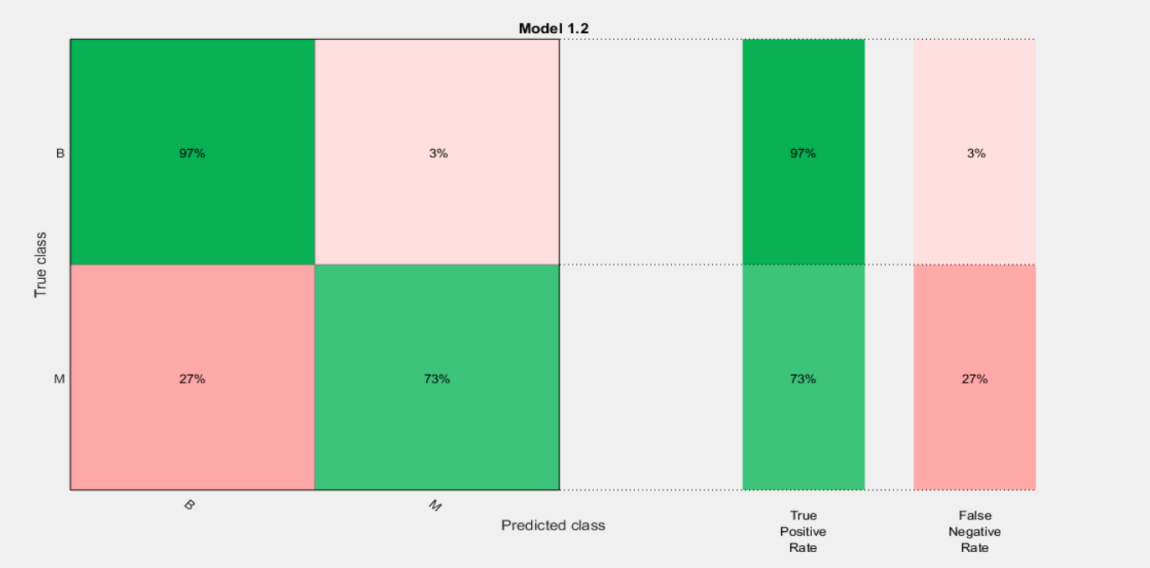


Figure 6.27: TPR and FPR with 6 features selected using RFE_CV and window size = 20

6.6 Individual performance of the six selected features

These 6 features collectively provide the best classification performance but it has to be made sure that these features are individually acceptable and no redundancy exists in them. Otherwise, the whole subset of features can be regarded as inaccurate.
SVM was used taking each individual feature from the dataset of window size 20 and performance scores were evaluated. Some of the features may not show great results on its own but it is the combination of these features that works as a better classification tool.

Aspect Ratio:

Accuracy: 80%

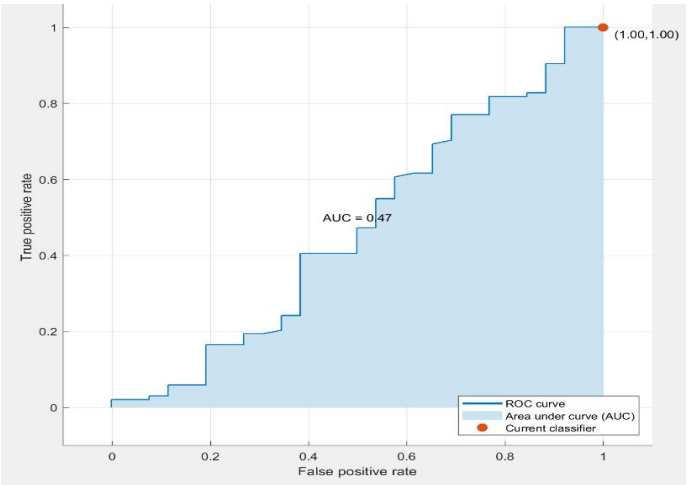


Figure 6.28: ROC curve for Aspect Ratio

Border Irregularity:

Accuracy: 87.7%

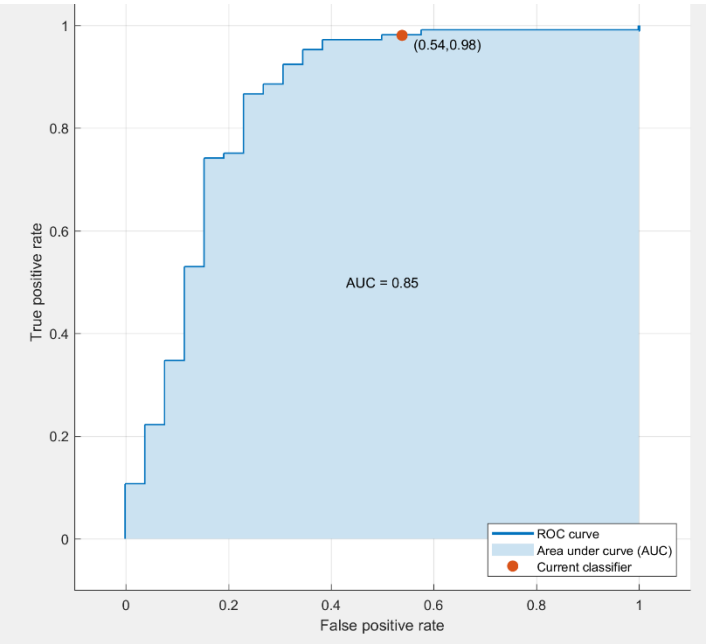


Figure 6.29: ROC curve for Border Irregularity

Absolute MG Area:

Accuracy: 80%

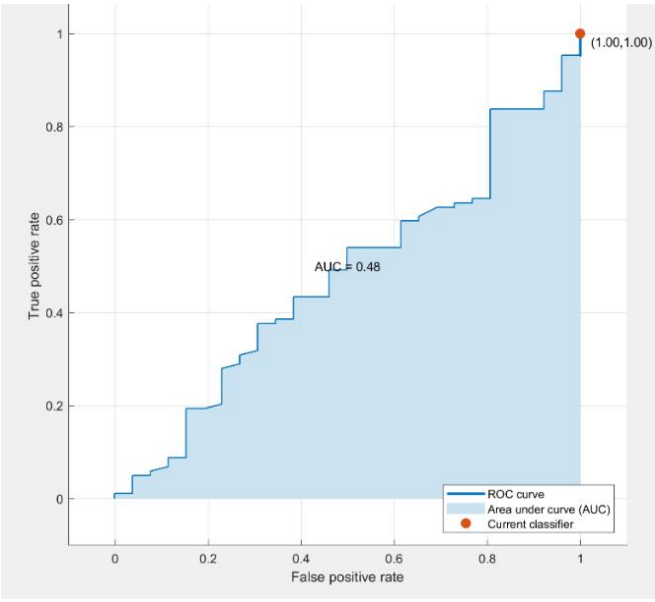


Figure 6.30: ROC curve for Absolute MG Area

Imaginary MG Gradient:

Accuracy: 80%

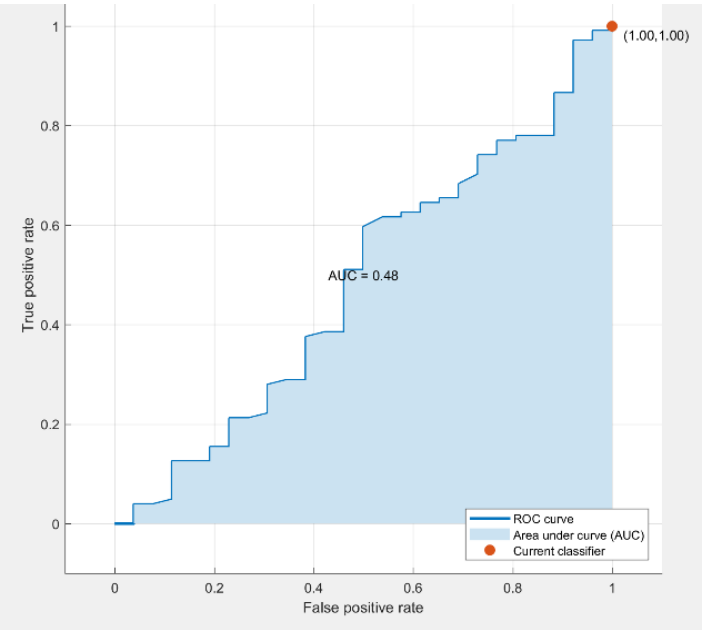


Figure 6.31: ROC curve for Imaginary MG Gradient

Phase MLP:

Accuracy: 80%

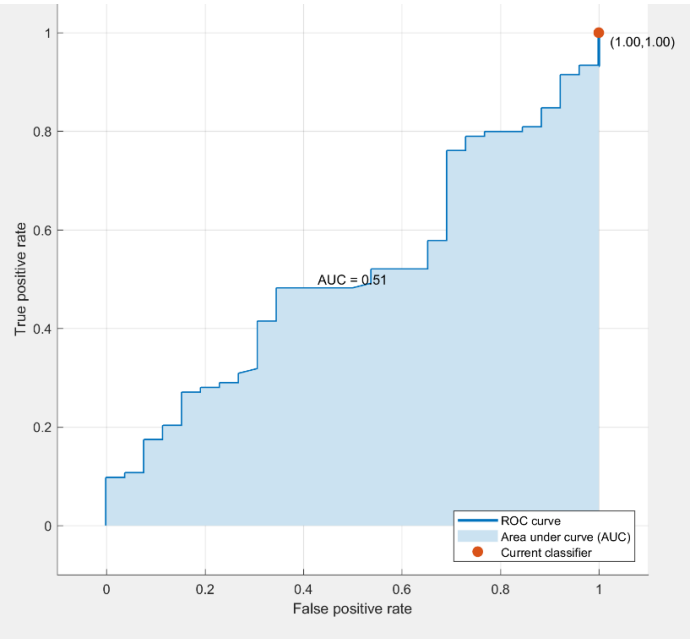


Figure 6.32: ROC curve for Phase MLP

Mu HTC:

Accuracy: 80%

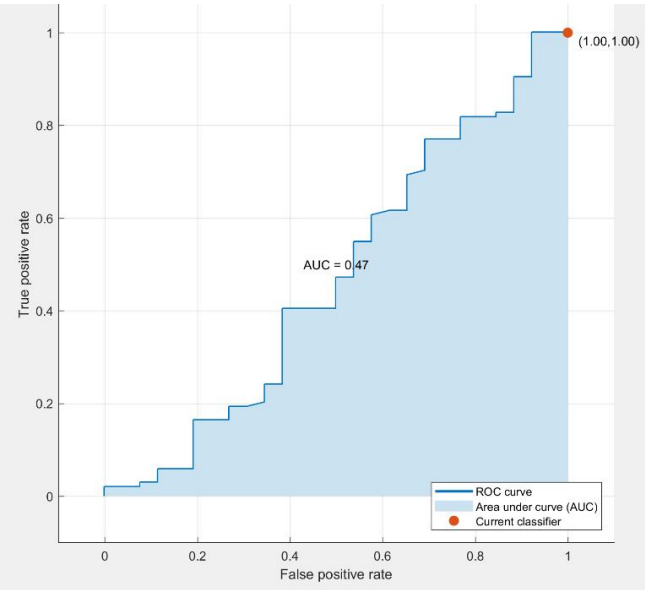


Figure 6.33: ROC curve for Mu HTC

6.7 Result summary

Performance summary of the individual features is tabulated below:

Table 6.1: Summary of the Performance Parameter Scores of the Selected Features

Features	**Percentage of Accuracy**	**Area under the curve**
Aspect Ratio	80%	0.47
Border Irregularity	87%	0.85
Absolute MG Area	80%	0.48
Imaginary Gradient	80%	0.48
Phase MLP	80%	0.51
Mu HTC	80%	0.47

From the above table it is clear that a feature alone cannot be a good classifier, rather the combination of all the six features gives a far better classification result.

Performance summary of the combination of six features is tabulated below:

Table 6.2: Comparison of performance parameter scores for three window sizes

Window Size	**Accuracy**	**Area under the curve**
5 (0.1875 mm)	87.7%	0.88
12 (0.4500 mm)	90.0%	0.93
20 (0.7500 mm)	92.3%	0.95

6.8 Empirical Analysis

An empirical analysis between different window sizes depicts an improvement of performance with increase in window size.

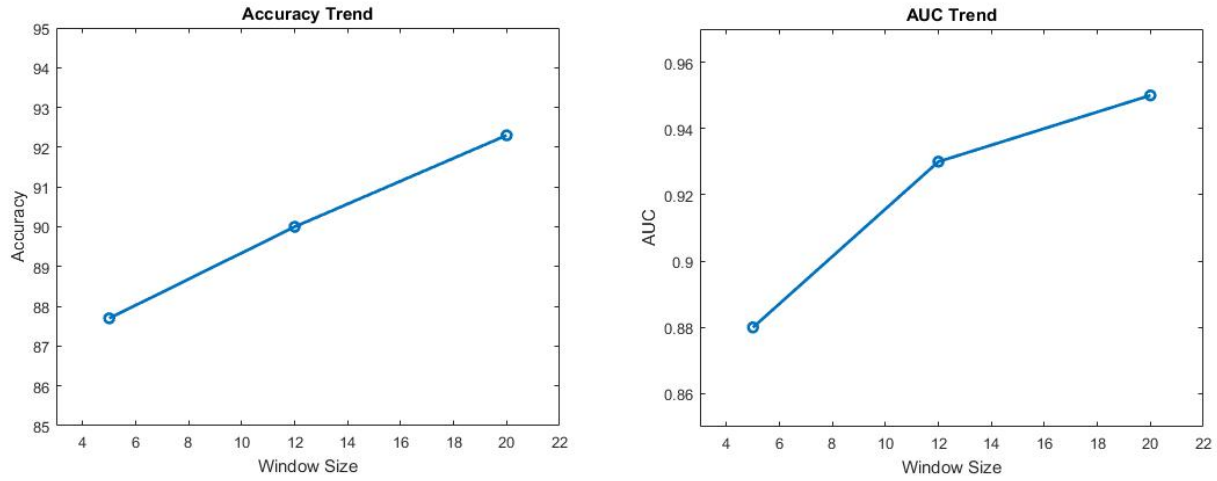


Figure 6.34: Empirical analysis of Accuracy and AUC trend vs window size

6.9 Conclusion:

All these results are obtained through a cross validation process, which is never the same with every run. So, the result of one run is not something constant and may change in another run, as cross validation sets are chosen at random. But with fixed initial seeds, this randomization can be stopped. A general analysis of this result depicts the viability of the generated features and due to the performance metrics being up to the mark these models along with these features can be applicable for classification of any new dataset.

Chapter 7
Conclusion

7.1 Introduction

The scope for using ultrasound image in cancer classification is although a very necessary demand but due to the image properties it is a real challenge. Our research finds an way of using statistical model to find a better way to process the RF signal of an ultrasound image and this process has been validated throughout several the performance metrics. Our research not only provides the way of image processing but also incorporate works of using advanced machine learning algorithm to for classification and feature selection. Combining the statistical methods with modern machine learning techniques this research provides a practical implementation of automated classification techniques in the field of cancer detection.

7.2 Future Work

Though the results of this thesis are quite convincing to be used for further research in the field of medical diagnosis, still there are numerous areas where much more enhancements can be brought. The scopes of our research can be divided into short term and long-term prospects.

7.2.1 Short Term Goals

The next attempt for research in the foreseeable future would be to increase the window size and generate the selected six parametric features for the increased window size. The performance parameters like, accuracy, area under the ROC curve would be observed again for these six features for increased window sizes. Another attempt would be to include other distributions like, K distribution, Homodyne K distribution for the statistical analysis of the backscattered envelope, as these statistical distributions exhibited high potential for breast lesion classification in the previous research works. An attempt would be to take nonlinear approaches like neural networks into account as classifier and observe the difference with linear approaches through observing performance parameters scores. A research scope would be to consider the error factors in studies for accurate estimation of Nakagami parameters.

7.2.2 Long Term Goals

An obvious attempt for research in the distant future would be to increase the number of samples by collecting new datasets. This would help us to verify the results we obtained from our study and determine whether similar results are obtained for the new patients'

datasets. Another major attempt would be to make the segmentation procedure of the region of interest fully automated. Till now the segmentation procedure used in our analysis is semi-automated as the tracing of the ROI was done manually. We can try to detect the boundary using more sophisticated image segmentation techniques. Another attempt would be to develop a better algorithm to compute parametric images, including the structures and the context around the lesion. There is also scope for applying convolutional neural networks to classify the parametric images we generated for classifying malignant and benign lesions.

7.3 A comparative analysis

A comparative analysis of significant findings in the field of ultrasound imaging in cancer classification has been listed in Table 7.1 below. From this table, we can observe that

Comparison	**Number of Patients**	**Parameter Types**	**ROC Area**
Ultrasonic Multi-Feature Analysis Procedure for Computer-Aided Diagnosis of Solid Breast Lesions	130 (26 malignant,104 benign)	Spectral and morphometric	0.947
Classification of Ultrasonic B-Mode Images of Breast Masses Using Nakagami Distribution	52 (14 malignant,38 benign)	M (effective number) and α (effective cross section)	0.79 0.828
Classification of breast masses in ultrasonic B scans using Nakagami and K distributions	99 (29 malignant,70 benign)	Nakagami and K distribution parameters	0.94
Ultrasonic Nakagami Imaging: a strategy to visualize the scatterer properties of benign and malignant breast tumors	100 (50 malignant,50 benign)	Nakagami image (m parameter)	0.81
Our research	130 (26 malignant,104 benign)	Nakagami and derived Nakagami parameters	0.95

Table 7.1: Comparison of result with other papers

Nakagami parameters and derived Nakagami parameters along with morphometric features which gives an AUC of 0.95. Comparatively, the achieved AUC is highest among all the studies using Nakagami distribution only. Therefore, we can conclude that Nakagami parameters and derived Nakagami parameters improves classification between benign and malignant breast tumors.

References

[1] American Cancer Society. *Breast Cancer Facts & Figures 2019-2020*. Atlanta: American Cancer Society, Inc. 2019.

[2] Cancer today, International Agency for Research on Cancer. Available: https://gco.iarc.fr/today/

[3] J.Mamou, Michael L. Oelze,W.D. O'Brien, et al.,"Perspective on Biomedical Quantitative Ultrasound Imaging,"*Applications Corner, IEEE Signal Processing Magazine*, pp. 112-116, May 2006.

[4] D'Astous FT and Foster FS, "Frequency Dependance of Ultrasound Attenuation and Backscatter in Breast Tissue," *Ultrasound in Medicine and Biology*, vol. 12, issue 10, pp. 795-808, 1986.

[5] Nam K, Zagzebski JA and Hall TJ, "Quantitative Assessment of In Vivo Breast Masses Using Ultrasound Attenuation and Backscatter," *Ultrasonic Imaging*, vol. 35, issue 2, pp. 146–161, April 2013.

[6] Michael L.Oelze,"Quantitative Ultrasound Techniques and Improvements to Diagnostic Ultrasonic Imaging," in *IEEE International Ultrasonics Symposium Proceedings*, Dresden,Germany, 2012, pp.232-239.

[7] Lizzi F.L., Astor M., Liu T., Deng C., Coleman D.J. and Silverman R.H., "Ultrasonic Spectrum Analysis for Tissue Assays and Therapy Evaluation," *International Journal of Imaging Systems and Technology*, vol. 8, issue 1, pp. 3-10, 1997.

[8] Lizzi F.L., Astor M., Feleppa E.J., Shao M. and Kalisz A., "Statistical Framework for Ultrasonic Spectral Parameter Imaging," *Ultrasound in Medicine and Biology*, vol. 23, issue 9, pp. 1371-1382, 1997.

[9] Alam S.K., Feleppa E.J., Rondeau M., Kalisz A. and Garra B.S., "Ultrasonic Multi-Feature Analysis Procedure for Computer-Aided Diagnosis of Solid Breast Lesions," *Ultrasonic Imaging*, vol. 33, issue 1, pp. 17-38, 2011.

[10] "Potential of Ultrasound Nakagami Imaging in Clinical Tissue Characterization," *Journal of Medical Ultrasound,Elsevier,* pp. 51-53, 2013. Available: www.sciencedirect.com

[11] P. M. Shankar,"A general statistical model for ultrasonic scattering from tissues," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 47, issue 3, pp. 727-736, May 2000.

[12] P. M.Shankar, V. A. Dumane, John M. Reid, Vladimir Genis, Flemming Forsberg,Catherine W. Piccoli, Barry B. Goldberg, "Classification of Ultrasonic B-Mode Images of Breast Masses Using Nakagami Distribution," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 48, issue 2, March 2001.

[13] P. M.Shankar, "A Compounding Scattering Pdf for the Ultrasonic Echo Envelope and its Relationship to K and Nakagami Distributions," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 50, issue 3, pp. 339-343, March 2003.

[14] P. M.Shankar, "The Use of the Compounding Probability Density Function in Ultrasonic Tissue Characterization," *Physics in Medicine and Biology* ,vol. 49, issue 6, pp. 1007-1015, April 2004

[15] Karmeshu, Agrawal R., "Study of Ultrasonic Echo Envelope Based on Nakagami-Inverse Gaussian Distribution,''*Ultrasound in Medicine and Biology* , vol. 32 issue 3, pp.371-376, March 2006.

[16] Agrawal R., Karmeshu,"Ultrasonic Backscattering in Tissue:Characterization Through Nakagami-Generalized Inverse Gaussian Distribution," *Computers in Biology and Medicine,Elsevier,* vol.37, issue 2, pp.166-172, Feb 2007.

[17] Bouhlel N., Sevestre-Ghalila S., "Nakagami Markov random field as Texture Model for Ultrasound RF Envelope Model," *Computers in Biology and Medicine,Elsevier,* vol.39, issue 6, pp. 535-544, June 2009.

[18] P. M. Shankar, V. A. Dumane, John M. Reid, Vladimir Genis, Flemming Forsberg,Catherine W. Piccoli, Barry B. Goldberg,"Classification of Breast Masses in Ultrasonic B-Mode Images Using a Compounding Technique in the Nakagami Distribution Domain," *Ultrasound in Medicine and Biology*, Vol. 28, issue 10, pp. 1295–1300, Oct 2002.

[19] P.M. Shankar, V. A. Dumane1, Thomas George,Catherine W Piccoli, John M. Reid,Flemming Forsberg,Barry B Goldberg, "Classification of Breast Masses in Ultrasonic B Scans Using Nakagami and K Distributions", *Physics in Medicine and Biology*, vol. 48, issue 14, pp. 2229–2240, Jul 2003.

[20] Po-Hsiang Tsui, Chih-Chung Huang, Shyh-Hau Wang, "Use of Nakagami Distribution and Logarithmic Compression in Ultrasonic Tissue Characterization," *Journal of Medical and Biological Engineering*, vol.26, issue 2, pp.69-73, June 2006.

[21] Po-Hsiang Tsui,Chien-Cheng Chang , "Imaging Local Scatterer Concentrations by the Nakagami Statistical Model,''*Ultrasound in Medicine and Biology*, vol. 33, issue 4, pp. 608–619, Apr 2007.

[22] Po-Hsiang Tsui ,Chih-Kuang Yeh,Chien-Cheng Chang and Wen-Shiang Chen, "Performance Evaluation of Ultrasonic Nakagami Image in Tissue Characterization,'' *Ultrasonic Imaging*, vol. 30, issue 2, Apr 2008.

[23] Po-Hsiang Tsui , Chih-Kuang Yeh, Yin-Yin Liao , "Ultrasonic Nakagami Imaging: A Strategy To Visualize The Scatterer Properties of Benign and Malignant Breast Tumors," *Ultrasound in Medicine and Biology*, vol. 36, issue. 2, pp. 209–217, Feb 2010.

[24] Aymeric Larrue, J. Alison Noble, "Modeling of Errors in Nakagami Imaging: Illustration on Breast Mass Characterization," *Ultrasound in Medicine and Biology*, vol. 40, no. 5, pp. 917–930, May 2014.

[25] Dutt V & Greenleaf JF, "Ultrasound Echo Envelope Analysis Using a Homodyned K Distribution Signal Model," *Ultrasonic Imaging*, vol. 16, issue 4, pp. 265-287, Oct 1994.

[26] Hruska, David P (2009), "Improved Techniques for Statistical Analysis of the Envelope of Backscattered Ultrasound Using the Homodyned K Distribution," University of Illinois at Urbana-Champaign, Champaign, IL, USA.

[27] Hruska D.P. and Oelze M.L., "Improved Parameter Estimates Based on the Homodyned K Distribution," *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, vol. 56, issue 11, pp. 2471-2481, Nov 2009.

[28] Trop I., Destrempes F., Khoury M.E., Robidoux A., Gaboury L., Allard L., Chayer B. and Cloutier G., "The Added Value of Statistical Modelling of Backscatter Properties in the Management of Breast Lesion at US," *Radiology*, vol. 275, issue 3, June 2015.

[29] H.D. Cheng, JuanShan, WenJu, YanhuiGuo, LingZhang , "Automated Breast Cancer Detection and Classification using Ultrasound Images: A Survey, " *Pattern Recognition,Elsevier*, vol. 43, issue 1 pp. 299-317, Jan 2010.
[30] F.Destrempes and G.Cloutier , "A Critical Review and Uniformized Representation of Statistical Distributions Modeling the Ultrasound Echo Envelope," *Ultrasound in Medicine and Biology*, vol. 36, issue 7, pp. 1037–1051, Jul 2010.
[31] P.Sprawls. *Ultrasound Production and Interactions.* Availbale:http://www.sprawls.org/ppmi2/USPRO/#THE%20ULTRASOUND%20IMAGING%20SYSTEM
[32] Anon. Available: https://www.google.com/search?q=logistic+regression&hl=en&sxsrf=ALeKk03uCh5auWYVxzznGNVfcG1CNLsbQA:1615536878282&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiA_4Dup6rvAhVh4nMBHeceC4kQ_AUoAXoECBgQAw&biw=189 7&bih=909
[33] Anon. *The Ultimate Guide to Decision Trees for Machine Learning*. Available: https://www.keboola.com/blog/decision-trees-machine-learning
[34] A.Navlani. *Decision Tree Classification in Python*. Available: https://www.datacamp.com/community/tutorials/decision-tree-classification-python
[35] Anon. Available: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DgoPiwckWE9M&psig=AOvVaw1EjsFD94j3UPTuDaGn9Prd&ust=1615387887307000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPiz8Py6o-8CFQAAAAAdAAAAABAD
[36] A.Navlani. *Understanding Random Forests Classifiers in Python.* Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python
[37] M.Deshpande.(2020, Sept.29). *Python Machine learning, Zenva*. Available: https://pythonmachinelearning.pro/perceptrons-the-first-neural-networks/
[38] T.Vasiloudis.(2019, Aug.26). *Block-distributed Gradient Boosted Trees*. Available: http://tvas.me/articles/2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html
[39] Anon. *Gradient Boosting, AI Wiki*. Available: https://docs.paperspace.com/machine-learning/wiki/gradient-boosting
[40] Anon. Available: https://www.google.com/search?q=support+vector+machine&sxsrf=ALeKk02F5b8RSW4pqIoR_gMGkxZYKTDNEQ:1615301800492&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjSnpeQvKPvAhUIlEsFHWBGDIQQ_AUoAnoECBUQBA&biw=1897&bih=909
[41] A.Navlani. *Support Vector Machines with Scikit-learn*. Available: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python
[42] American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) Atlas (American College of Radiology, Reston, VA, 2003).
[43] Mendelson EB, Böhm-Vélez M, Berg WA, et al., ACR BI-RADS® Ultrasound, In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, Reston, VA, American College of Radiology, 2013

[44] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, special Issue on variable and Feature Selection 3, pp.1439-1461, 200

[45] Xue-wen Chen, Jong Cheol Jeong, "Enhanced Recursive Feature Elimination", Sixth International Conference on Machine Learning and Applications, DOI: 10.1109/ICMLA.2007.35
[46] X. Chen, "Margin based wrapper methods for gene identification using microarray," *Neurocomputing*, vol. 69(16-18): 2236-2243, 2006.
[47] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, special Issue on variable and Feature Selection 3, pp.1439-1461, 2003
[48] M. Ashraf, G. Chetty, and D. Tran, "Feature selection techniques on thyroid,hepatitis, and breast cancer datasets," *International Journal on Data Mining and Intelligent Information Technology Applications(IJMIA)*,vol. 3, no. 1, pp. 1-8, 2013.
[49] M. Leach, "Parallelising feature selection algorithms," University of Manchester, Manchester, 2012.
[50] Pinar Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease", *International Journal of Machine Learning and Computing,* Vol. 5, No. 4, August 2015
[51] Mouhamadou Lamine Samb, Fodé Camara, Samba Ndiaye, Yahya Slimani and Mohamed Amir Esseghir, "A Novel RFE-SVM-based Feature Selection Approach for Classification", International Journal of Advanced Science and Technology Vol. 43, June, 2012
[52] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machining Learning*, vol. 46, no. 1-3, pp. 389-422, 2002.
[53] Naeem Seliya, Taghi M. Khoshgoftaar, Jason Van Hulse, "A Study on the Relationships of Classifier Performance Metrics", 2009 21st IEEE International Conference on Tools with Artificial Intelligence
[54] N. Japkowicz. Classifier evaluation: A need for better education and restructuring. In Proceedings of the 3rd Workshop on Evaluation Methods for Machine Learning, ICML 2008, Helsinki, Finland, 2008.
[55] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In In the Australian Conference on Artificial Intelligence, pages 1015–1021, 2006.