CAD-BASED AUTOMATED CARCINOMA DETECTION AND CLASSIFICATION IN BREAST CANCER DIAGNOSIS

Mohamed Taoufik

M.Sc. Thesis

UNIVERSITY OF
EASTERN FINLAND

School of Computing

Kuopio Campus

August 2015

UNIVERSITY OF EASTERN FINLAND, Faculty of Science and Forestry, School of Computing, Kuopio Campus

Mohamed Taoufik: CAD Based Automated Carcinoma Detection and Classification in Breast Cancer Diagnosis
Master's Thesis, 64 p., 2 appendix (p.)
Supervisor of the Master's Thesis: Ph.D., M.Sc. (Tech.), Keijo Haataja
August 2015

Keywords: Breast Cancer, CAD, Feature Selection, Histopathology, LDA, Neural Network, PCA, SVM

The main objective of this M.Sc. Thesis is to study methods for automated carcinoma detection and classification. Computer Assisted Diagnosis (CAD) is a method designed to decrease the human intervention. It is a second reader that assist physician with interpretation of medical images. Everyday new CAD systems are developed towards histopathology in order to ameliorate diagnosis and/or prognosis.

Matlab software was used to pre-process histological breast images from which we have extracted first and second order statistical features, e.g., Standard Deviation, Mean, Variance, Skewness, Kurtosis, Smoothness, Range Filter, Entropy, Contrast, Correlation, Energy and Homogeneity. These feature descriptors will be transformed to feature vectors and then Principal Component Analysis (PCA) will be applied for feature selection, since in statistical learning feature selection or dimensionality reduction is an essential task when dealing with less observations but a large number of features. The input data will be classified with three different classifiers: SVM (Support Vector Machine), LDA (Linear Discriminant Analysing) and NN (Neural Network). The accuracy and performance will be measured for the three classifiers in order to show their importance in pattern classification.

Abbreviation

CAD	Computer Aided Diagnosis
CT	Computerized Tomography
DCIS	Ductal Carcinoma in Situ
DA	Discriminant Analysis
FCR	Finnish Cancer Registry
FOSF	First Order Static Feature
FN	False Negative
FNA	Fine Needle Aspiration
FP	False Positive
GMD	Gaussian Multivariate Distribution
HOG	Histogram Oriented Gradient
HTL	Histology Technologists
H&E	Hematoxylin and Eosin Staining Technique
GLCM	Gray-Level Co-occurrence Matrix
IDC	Invasive Ductal Carcinoma
IDC	Invasive Ductal Carcinoma
LBP	Local Binary Pattern
LDA	Linear Discriminant Analyzing
MST	Minimum Spanning Tree
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
NN	Neural Network

NNPR	Neural Network Pattern Recognition
PCs	Principal Components
PCA	Principal Component Analyzing
RGB	Red-Green-Blue
ROC	Receiver Operating Characteristics
ROI	Region-Of-Interest
SOSF	Second Order Static Feature
STD	Standard Deviation
SURF	Speed up Robust Features
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

Contents

1. Introduction

According to statistics breast cancer is the second disease causing mortality among women [WHO14]. The most common type is carcinoma, meanwhile it is the most curable tumor if it is diagnosed at premature stage (early detection). Based on the data provided by Finnish Cancer Registry (FCR) from 2008 to 2012 about 4397 new cases (4377 females and 20 males) diagnosed with breast cancer. The number of deaths is about 851 persons between 2008 and 2012 [FCR13]. The key to reduce the mortality rate is to inform the population to take an early diagnosis. According to FCR statistics, an early detection may help to cure the disease and reduce the mortality rate.

With rapid growth of computer power to analytical approaches, computer aided diagnosis (CAD) is becoming an essential tool to assist radiologists and pathologists with interpretation of diseases. CAD can be defined as secondary reader using different screening tools in detecting abnormalities, lesions and masses. Recently, there are different methods and tools for breast cancer detection: such tools are Mammogram and CT which both use *x-rays* with different wavelengths [CDT14, CDM14], while Ultrasound tools uses *sound waves* [CDU12] and MRI that uses *magnetic energy & radio waves*. Another technique used in diagnosis is biopsy (Fine Needle Aspiration) which is a bit different from the previous methods, since it is in a direct contact with tissue or fluid extracted from suspicious area. The tissue sample will be collected under local anesthesia using ultrasound or mammography guidance [CDB14]. Biopsy is a diagnosis procedure considered complementing the opinion of the radiologist.

In this topic, we review the state of the art CAD for histological images of breast cancer under Matlab software. The experimentation work consists of 7 sections. Starting by the introduction. In the second section we will describe the breast structures and breast cancer conditions. In section 3 we will go through data preprocessing techniques. In section 4 we will extract some statistical features. Through section 5 we will select the relevant feature by reducing the dimensionality using PCA method.

In section 6 we will build, train, and test three classifiers LDA, SVM and NN with input features. Finally we will measure the performance and accuracy for the three classifiers. Then we will select the one which present higher performance for testing other histological images [GRN13, Cra05]. In the last section we will discuss the results obtained from processing methods used for diagnosis and classification of carcinoma.

2. Breast Cancer Anatomy and Histology

In this section we will show the anatomy of the breast by citing different organs and their functions. Meanwhile we will discuss the cancer condition and the preparation of histopathological images. The last part will be a short discussion of the CAD.

2.1 Breast Anatomy

The breast is the organ which overlay the chest, generally the role of women's breast is to produce milk. Healthy breast is made up of fatty tissue which determines the size of the breast. It contains 12-20 sections called Lobes that are formed by smaller structures called lobules (glandular tissue) that produce milk. Each lobule is connected to the nipple by a thin tubes where the milk is drained [NCI14], (see Figure 1).

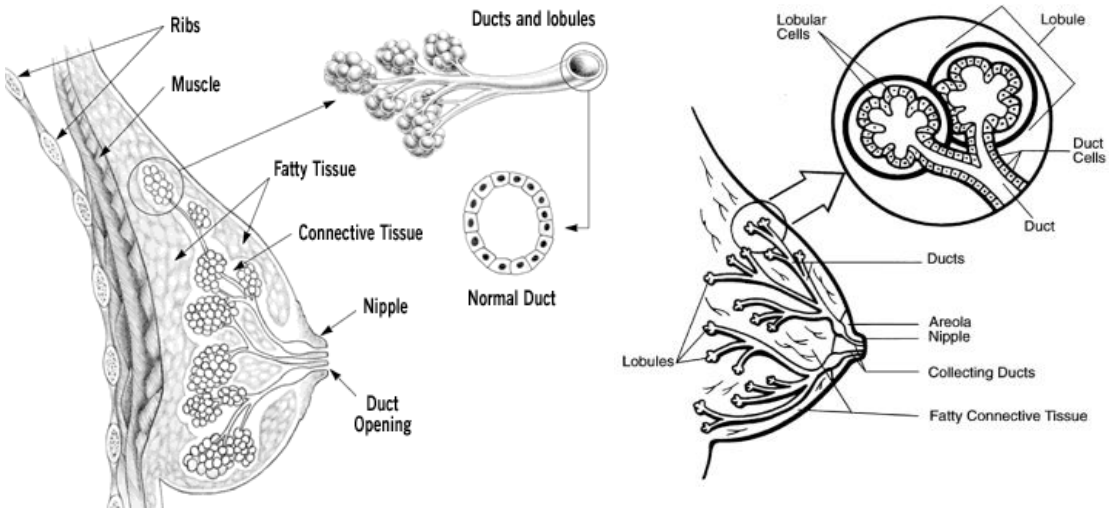

Figure 1. Anatomy of female breast. [CCV07]

2.2 Breast Cancer

The breast function and shape can be modified under certain conditions. Most women feel changes during their lifetime: this can be caused by changes in hormones. With menopause (advance in aging) breasts start developing some changes like reducing its size and feeling lumpy (presence of masses). This kind of changes are not cancerous and they do not make any risk to the breast called Cyst (Benign mass) (see Figure 2). [NCI14]

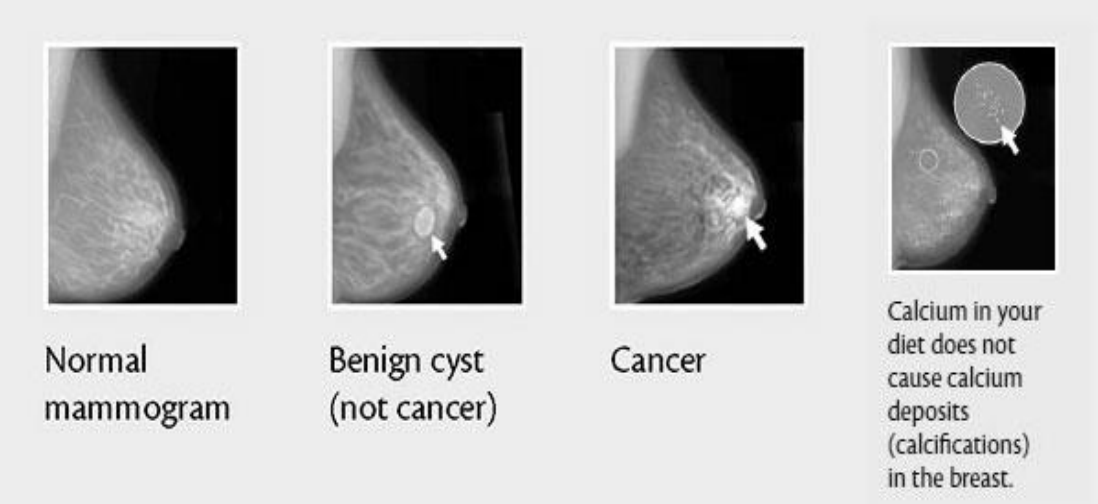

Figure 2. Mammogram shows breast mass (Cyst). [NCI14]

Cancer is a condition of abnormal cell growth in the body. Cells form tissues and tissues form the organ. Sometimes cells growth becomes unregulated (Mutation in genes). This can be occur under changes in some factors: physical, chemical or environmental. They start to replicate uncontrollably meaning more reproduced cells than died cells which results in masses of cells called Tumors. [VMC14]

Breast cancer is the most common cancer detected in women and rated as the second causing deaths in the world. There exists two invasive breast conditions in which cancerous cells start to invade the surrounding tissue and spread to other organs through a blood or lymphatic vessels. Non-invasive means that the cancerous cells are still contained in the duct [BCO13] (see Figure 3) [SLF14].

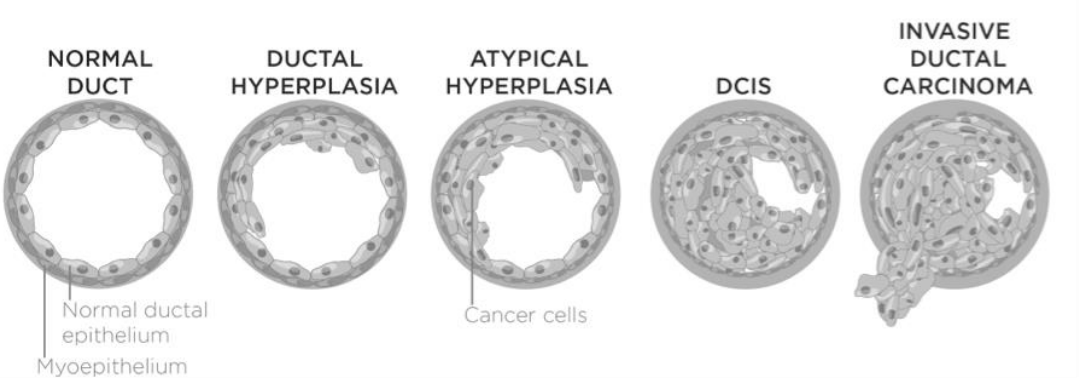

Figure 3. Progression from normal to invasive cancer. [WeJ73]

Sometimes breast can develop a mixture of tumors from ductal or lobular cells:

"*When the breast shows more than one tumor the breast cancer is described as either multifocal or multicentric. In multifocal breast cancer, all of the tumors arise from the original tumor, and they are usually in the same section of the breast. If the cancer is multicentric, it means that all of the tumors formed separately, and they are often in different areas of the breast*" (see Figure 4). [BCO14]

The most condition of breast cancer are ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC): [BCO14]

- *DCIS — Ductal Carcinoma In Situ*
- *IDC — Invasive Ductal Carcinoma*
    - *IDC Type: Tubular Carcinoma of the Breast*
    - *IDC Type: Medullary Carcinoma of the Breast*
    - *IDC Type: Mucinous Carcinoma of the Breast*
    - *IDC Type: Papillary Carcinoma of the Breast*
    - *IDC Type: Cribriform Carcinoma of the Breast*

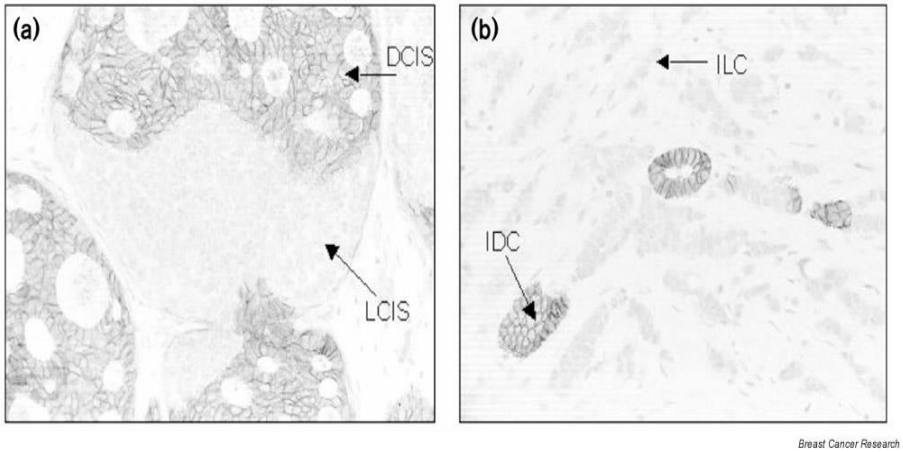

Figure 4. Difference between a) DCIS and b) IDC. [MaC02]

2.3 Preparation of Histological Breast Tissue

Histology is a branch of biology concerned with the study of microscopic anatomy of tissues by examining a group of cells (tissues) under light microscope or electron microscope. While in anatomical pathology Histopathology is the study of diseased tissue by trained physicians to provide a diagnostic of the tissue. The histological sections are prepared by histology technologists (HTL) who have been well trained in area of histotechnology. [HLA10, RoP06]

H&E (Hematoxylin and Eosin) stain or HE stain is staining technique in histology. It is an auxiliary method used in microscopy to enhance contrast in the histological images: sometimes combined stains are very important to reveal more details and features than a single stain.

The preparation of the histological tissue consists of two phases: tissue preparation and image production [HLA10, HLA12].

Tissue preparation is a method for processing tissues collected by FNA (Fine Needle Aspiration). The experimentation method include several steps: Fixation, tissue

processing sectioning and staining. An example of stained histological image is illustrated in Figure 5.

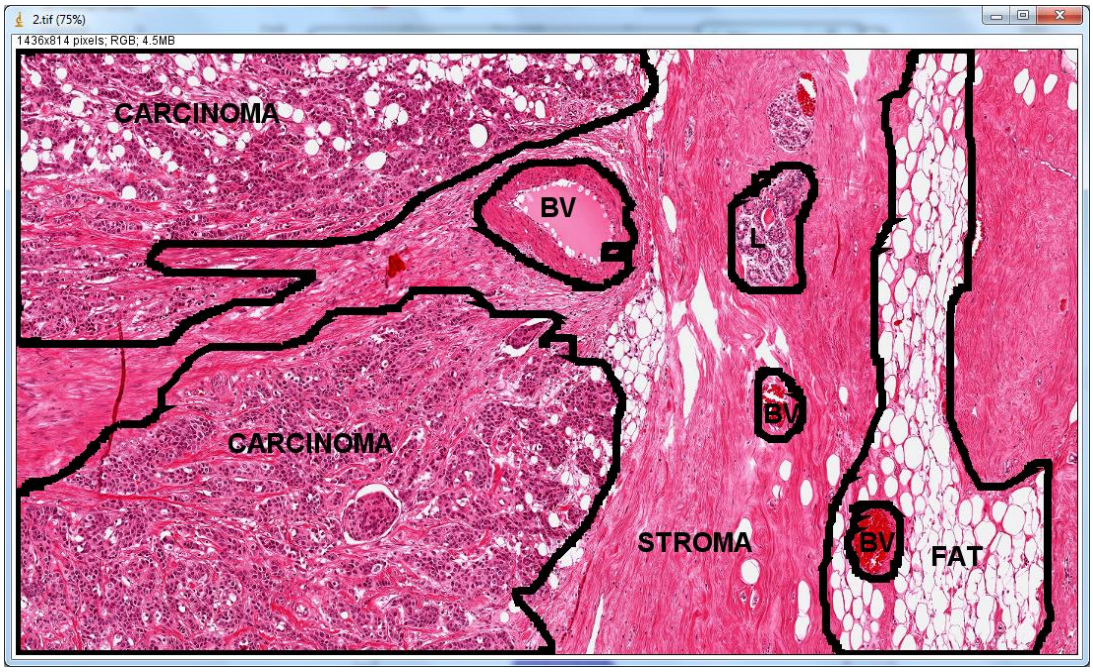

Figure 5. Stained histological image

"Even with careful operation by experienced technicians or clinicians, artifacts may still appear in the stained slides. These artifacts may result from improper fixation, wrong fixatives, poor dehydration and paraffin infiltration, improper reagents, or poor microtome sectioning. To reduce such artifacts, the tissue preparation procedures are usually implemented by automated systems." [HLA12] (see Figure 6).

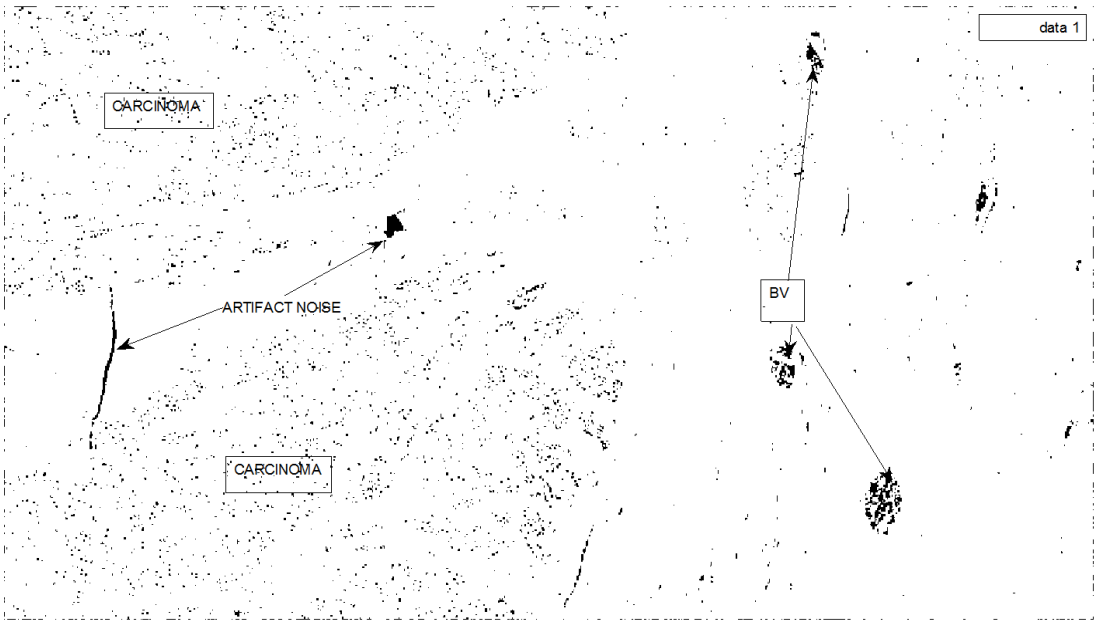

Figure 6. Artifacts in Negative Binary Image of Figure 5

Image production consists of taking digital histology images by a light or electron microscopes of the stained section (see Figure 7).

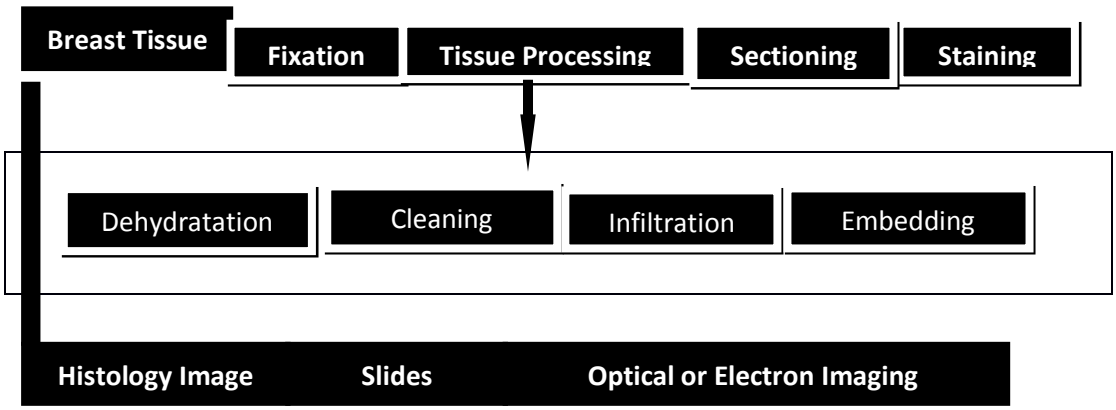

Figure 7. Tissue Preparation and Histological Image Production.

2.4 Computer Assisted Diagnosis

In the past decades, pathologists have examined histopathological images using manual methods for disease diagnosis. They visualize under microscope the regularities and distributions of the tissue. Based on their personal experiences in cell morphology, pathologists discriminate between different cells in tissue and then they determine the pattern of biopsy samples examined. The manual diagnostic preciseness and accuracy is a problematic task since the outputs will occur with considerable variability. For example, many pathologists will interpret the results differently depending on their expertise and their physical/emotional feeling (fatigue, happiness, sadness). Another fact is that, one pathologist cannot process thousands of images with same accuracy and preciseness in minimal time. However researches were focused on how to minimize errors-cost, standardizing the process, overcoming the stressful techniques and reducing the human intervention. It was important for them to develop computational tools towards automated cancer diagnosis. [HLA12]

CAD systems are becoming crucial to improve the reliability of cancer diagnosis: a tremendous amount of research papers were conducted for automated cancer detection/prognosis. This will help users and clinicians without computer training to interpret histological images and making decisions. [DAM06]

The automated cancer diagnosis consists of four main computational steps: preprocessing of the histological images, feature extraction, feature selection and diagnosis which is a classification (see Figure 8).

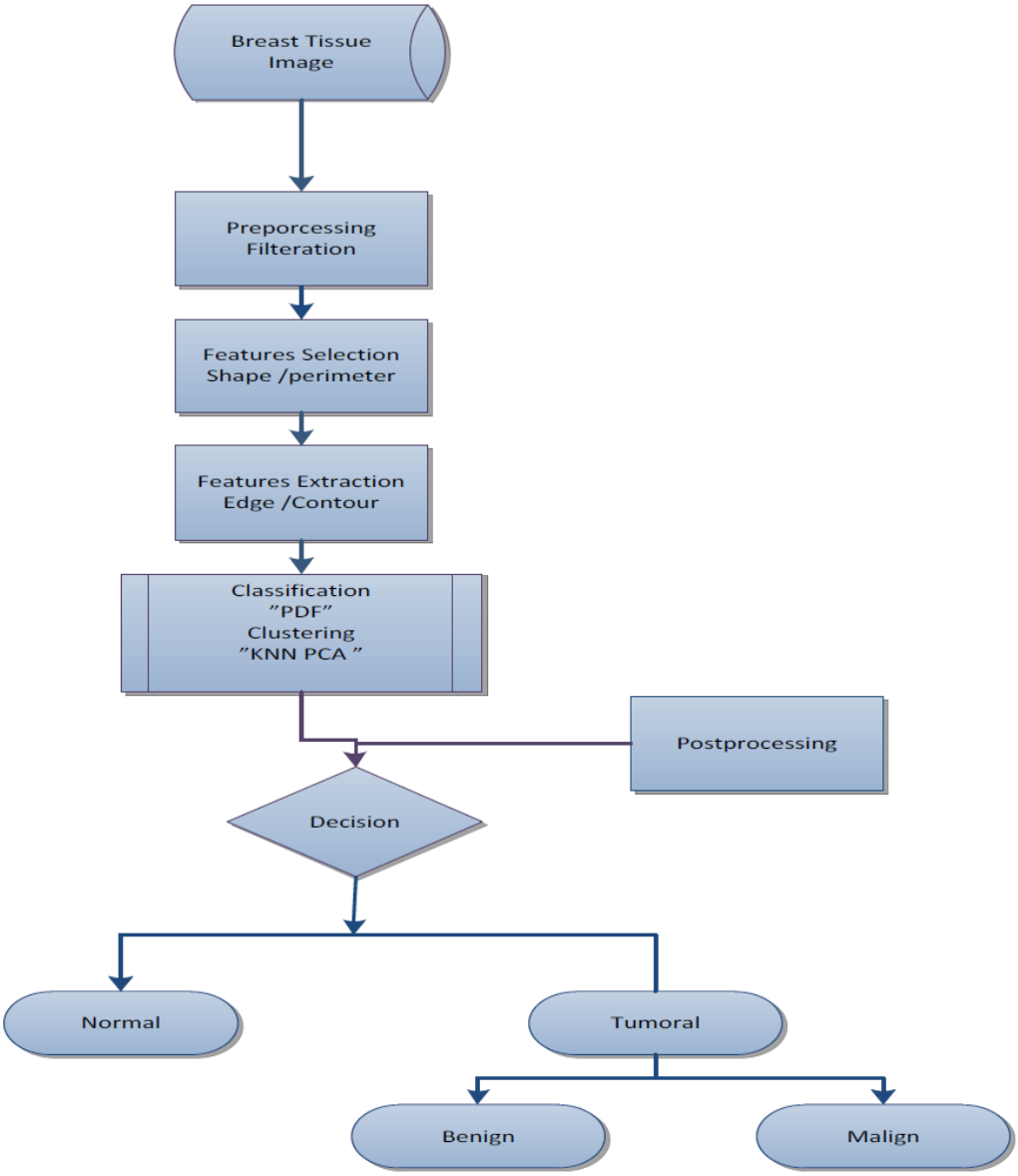

Figure 8. Automated cancer diagnosis flowchart.

3. Preprocessing

3.1 Grayscale Conversion

Conversion from 3-D RGB image to 2-D grayscale image by eliminating the hue and saturation information while retaining the luminance: The value of each pixel carries only the intensity information [McA04]. Since the three colors have different wavelengths, their contribution to grayscale level is then a result from luminosity method. Grayscale conversion (see Figure 9) is a weighted sum of the Red, Green, and Blue components deduced from Equation 1: [GWE03]

$$P_{intensity} = 0.2989 * R + 0.5870 * G + 0.1140 * B \quad (1)$$

Grayscale image called monochromatic is a type of digitized image in which every pixel is represented by its intensity information. To visualize the different intensities of a grey image, a histogram is used that provides a statistical representation of the intensities distribution in a discrete intervals called bins from black 0 to white 255 or vice versa. [GWE03]

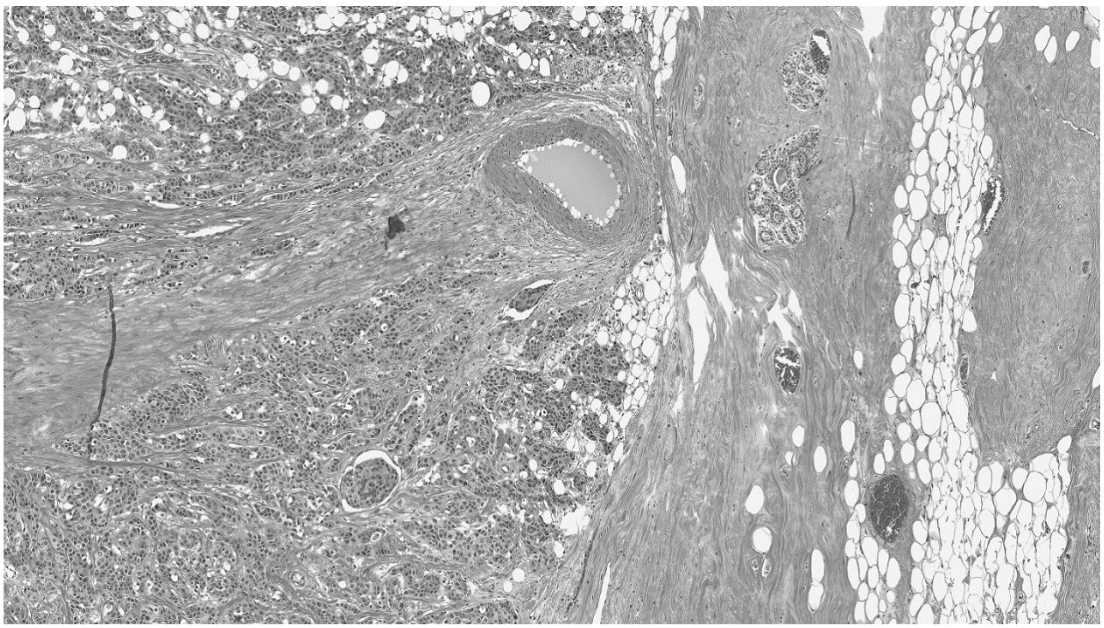

Figure 9. Grayscale conversion of a breast tissue.

3.2 Image Filtering

Image Filtering (and filters in general) is a processing method to suppress frequencies of images with preservation of image properties. In image processing, filtration is an important part of the image quality enhancement. It removes artifacts and cancels noises that may interfere with histology images [McA04]. Such method in a spatial domain for image filtering, the neighborhood pixels for any given pixel input contribute to assign a new output pixel in image. These pixel influences are performed by computing operations called Convolution and Correlation [McA04].

In order to enhance the quality of histology images a symmetric Gaussian filter is used [McA04] from Matlab with mask size of 3×3 and sigma value *"standard deviation 0.8"*. The result is an output image (see Figure 10).

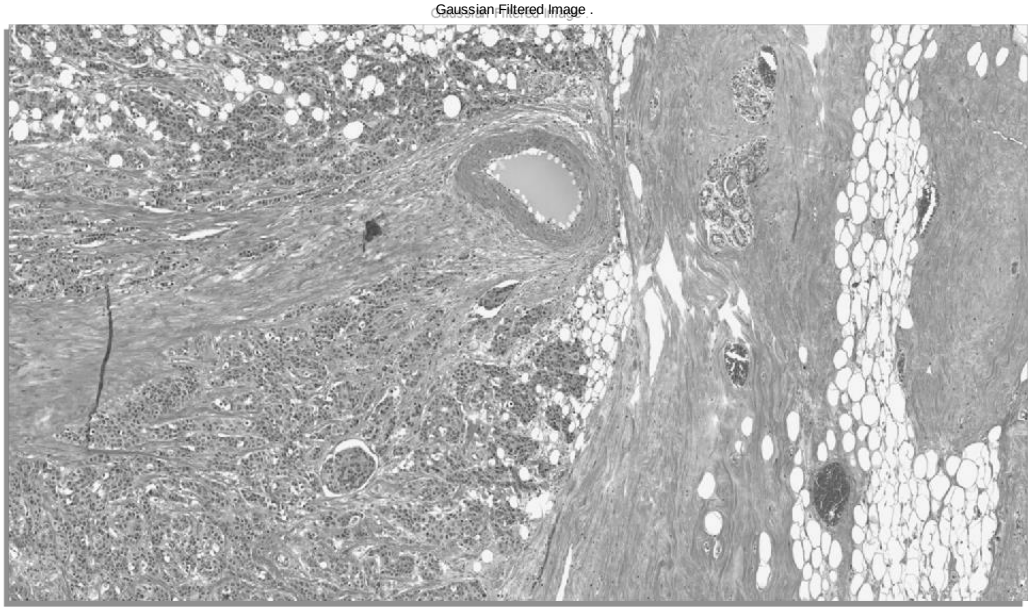

Figure 10. Filtration using Gaussian filter.

3.3 Contrast Enhancement

The contrast equalization consists of remapping the range of intensities to a new ones (contrast stretching). To enhance the contrast of histology images we have used histogram equalization. *The histogram of the output intensity image is flatter when number of discrete levels is much smaller than the number of discrete levels in input intensity image* [GWE03] (see Figure 11).

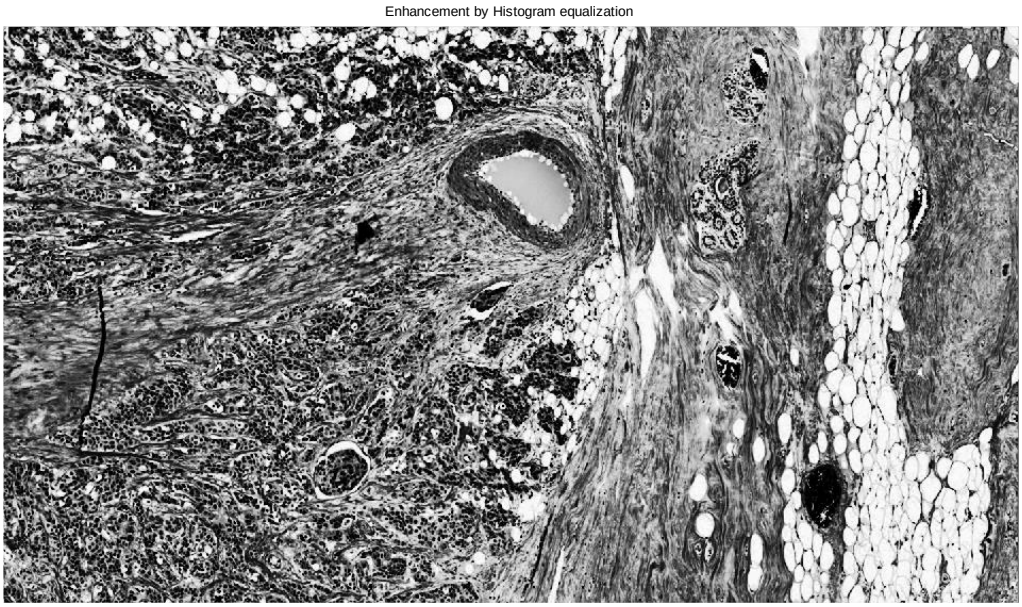

Figure 11. Contrast enhancement.

3.4 Unwanted Objects Removal

In order to remove artifacts, regions properties are used which consist of measuring the areas properties of each connected component. The objects having values greater than threshold are eliminated.

Region properties work only with binary images, thus we have converted the gray level into binary image (see Figure 12), then we have used a predefined function from Matlab with connected components to remove some unwanted areas and preserving regions of interest (ROI; see Figure 13).

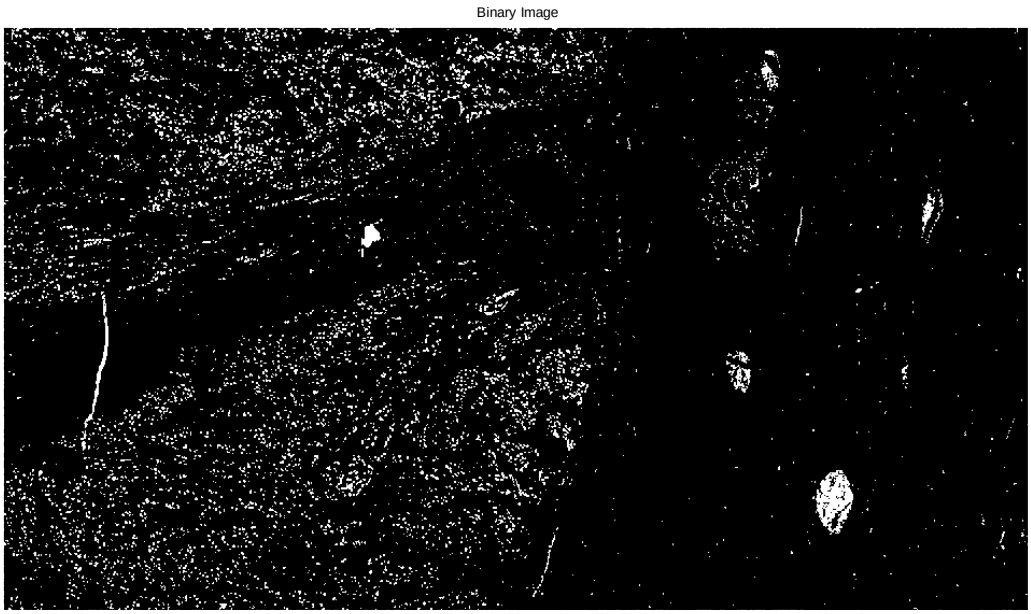

Figure 12. Binary Image.

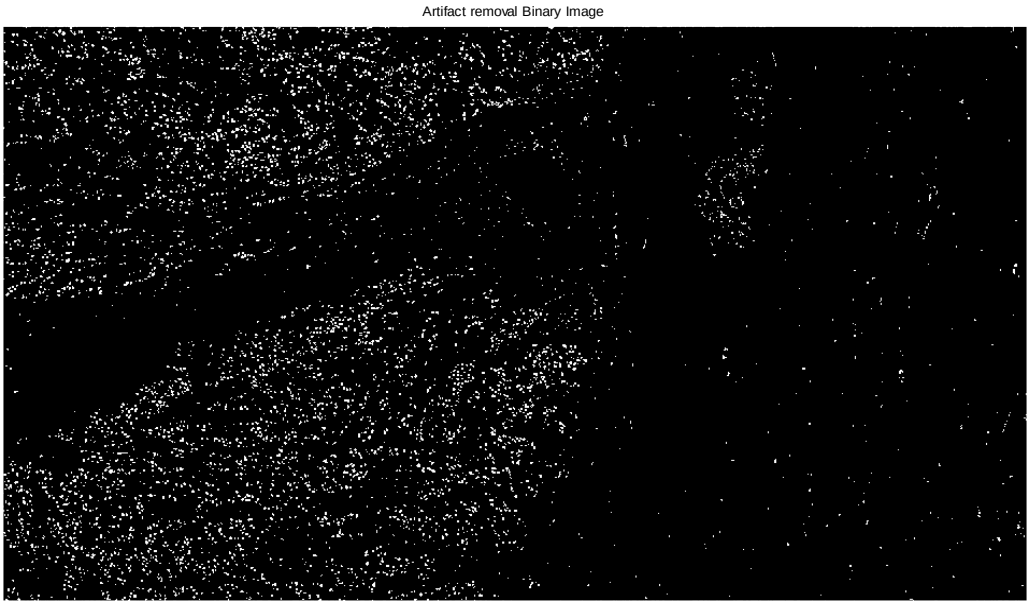

Figure 13. Artifact Removal.

4. Feature Extraction

When histologists need to diagnose a tissue, they start look out for some meaningful feature that discriminate between normal and malicious tissues. This section is the most important part of the work. Since it is used to extract those significant feature for classification. In this section we will give a short description of distance transform and how it is used with histological images. In the second subsection we will discuss some methods used in feature extraction. The last part consists of extracting some statistical and textural feature descriptors.

4.1 Distance Transform

Binary Distance transform

How to get back to grayscale image from binary image?

Borgefors in 1986: "*A distance transform converts a binary digital image, consisting of feature and non-feature pixels, into an image where all non-feature pixels have a value corresponding to the distance the nearest feature pixel*". [Bor86]

By applying Distance Transform (DT) which is a morphological transformation used for measuring the separation of pixels in binary image regions. The Euclidean distance (8 connections) was chosen to calculate the distance map between any pixels to the nearest boundary pixel (non-zero pixel) [RoP86]. The resulted image is an intensity image (grayscale) shown in Figure 14.

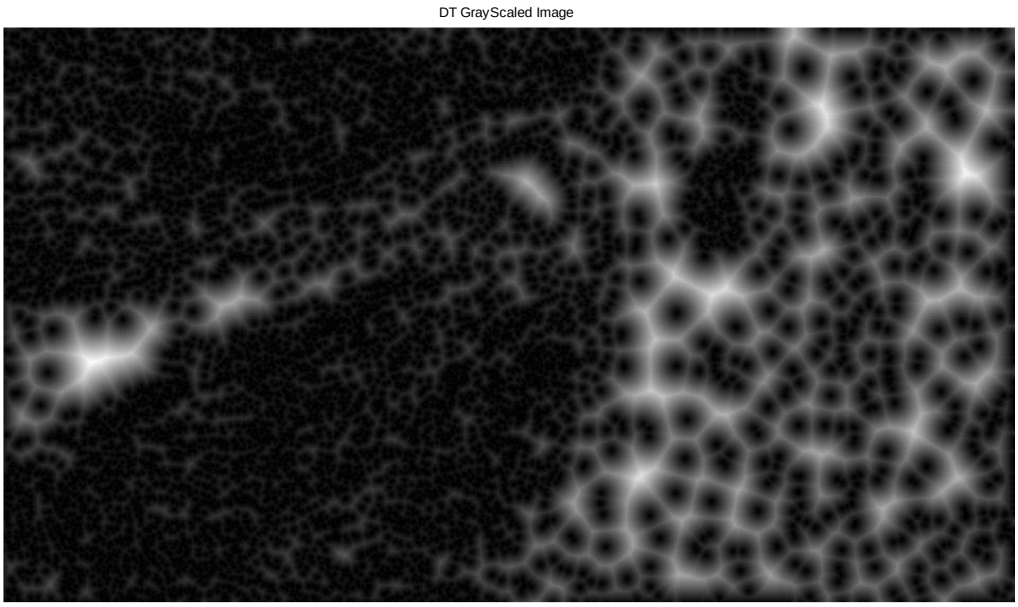

Figure 14. Distance Transform.

Distance transform on curved space DTOCS

DTOCS is a distance transform on curved space which is a geodesic distance transform method.

DTOCS as defined by Ikonen and Toivanen: "*The images are treated as height maps, where low-gray values (black and dark pixels) indicate low areas, and high gray-values (white and light pixels) indicate high areas*." [IkT05, ToI05]. (See Figure 15)

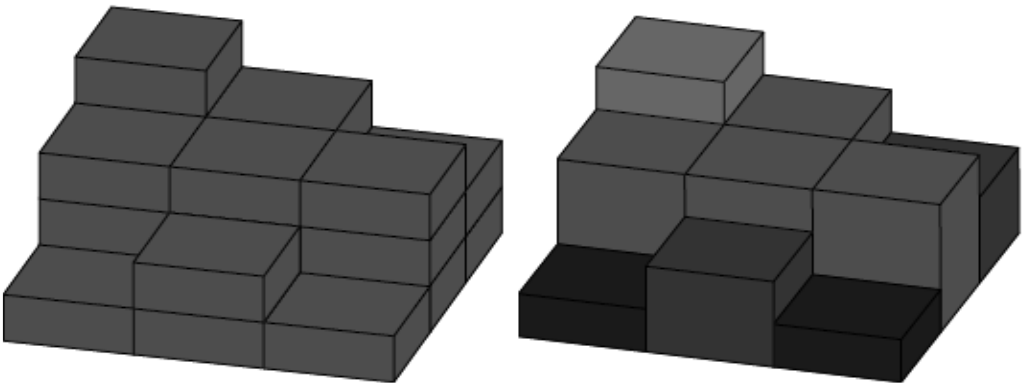

Figure 15. 3D Object (left) and Visualization of Gray-level surface (Right).

DTOCS was applied to image in Figure 14 which is a gray-level image, transforming it into a distance image, where the value of a pixel indicates its distance to the nearest reference pixel (feature pixels) [ToI05]. The result is an image distance in which clearly we see the higher values as brighter pixels and low values are darker pixels (see Figure 16).

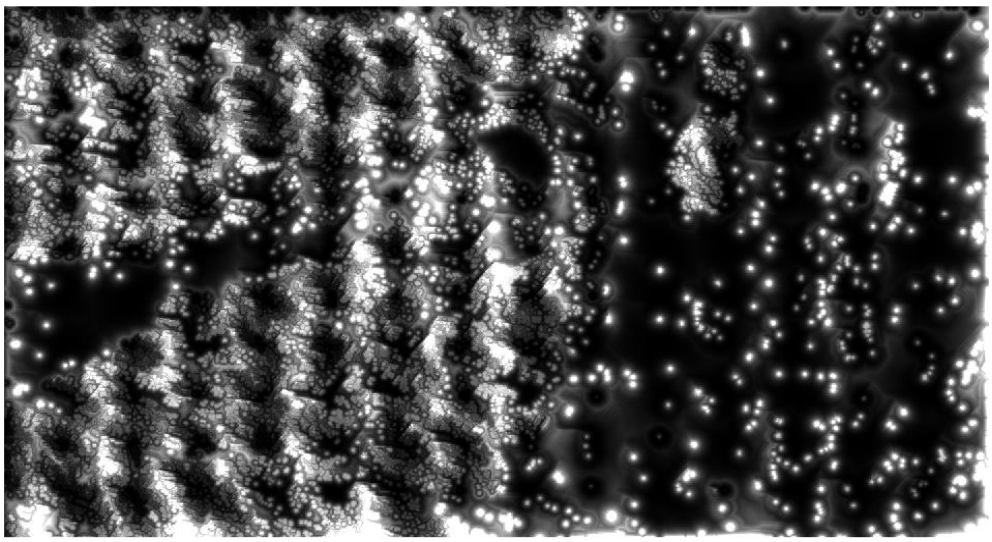

Figure 16. DTOCS, Visualization of Gray-level Surface (Block 11x15).

4.2 Feature Extraction Methods

The feature extraction can occur at the cell level or the tissue level in order to measure the properties of image abnormality or to assign the histological image to its pattern [BeM12].

When we analyze an individual cell we do not consider a spatial dependency: we only focus on different elements in the cell based on their morphology and texture. Otherwise at the tissue level feature, we take in consideration the distribution and the spatial dependency of the cells across the whole tissue [BeM12].

The objective of feature extraction is to reveal all possible features from input data that are expected to be effective in diagnosis with no dimensionality concerning. Such common feature extraction techniques are histogram of oriented gradients (HOG),

speed up robust features (SURF), local binary patterns (LBP), Haar wavelet and color histogram [BET08].

Scot Doyel had included textural and nuclear architectural features for analysis of breast cancer based on histological images: in his paper a set of features were extracted: Textural Features, Haralick Features, Graphical Features, Gabour Filter Features, Graph Features using Voronoi Diagram, Delauny Triangulation, Minimum Spanning Tree, Nuclear Features, Morphological Features and Topological Features [DAM08]. Table 1 shows feature types and their descriptors.

Table 1. Different features and their descriptors.

Features:	**Descriptors:**
Textural Features: First order statistics, GLCM, Run length Matrix	For Tissue Classification: Mean, standard deviation, variance, skewness, kurtosis contrast, correlation, smoothness, coarseness, regularity, energy, homogeneity, range filter, entropy.
Graph Features: Voronoi Diagram, Delauny Triangulation, MST	For Cell Detection: Edges, area, perimeter, roundness factor, number of nodes, spectral radius.
Morphological features of cell	For Abnormality Detection: Radius, area, perimeter, size, shape, clumb thickness, nucleoli, bare nuclei.

4.3 Statistical and Textural Features

Since we were interested in tissue classification, we have used some statistical feature extracted from gray-level intensity images. Such textural features are first order and second order statistics determined from the distribution of the gray-level pixels. In the first subsection we will extract some descriptors based on first order histogram and in the second subsection we will extract some textural descriptors based on gray level co-occurrence matrix which is a second order histogram.

4.3.1 First Order Statistics Features (FOSF)

The histogram is a key tool in collecting information about images. It is useful when working with contrast: if the gray level are concentrated near a certain level the image is interpreted as low contrast, meanwhile if it is spread out over the entire image then higher contrast [GWE03, Mat14a]. An example of histological breast image foreground histogram is depicted in Figure 17.

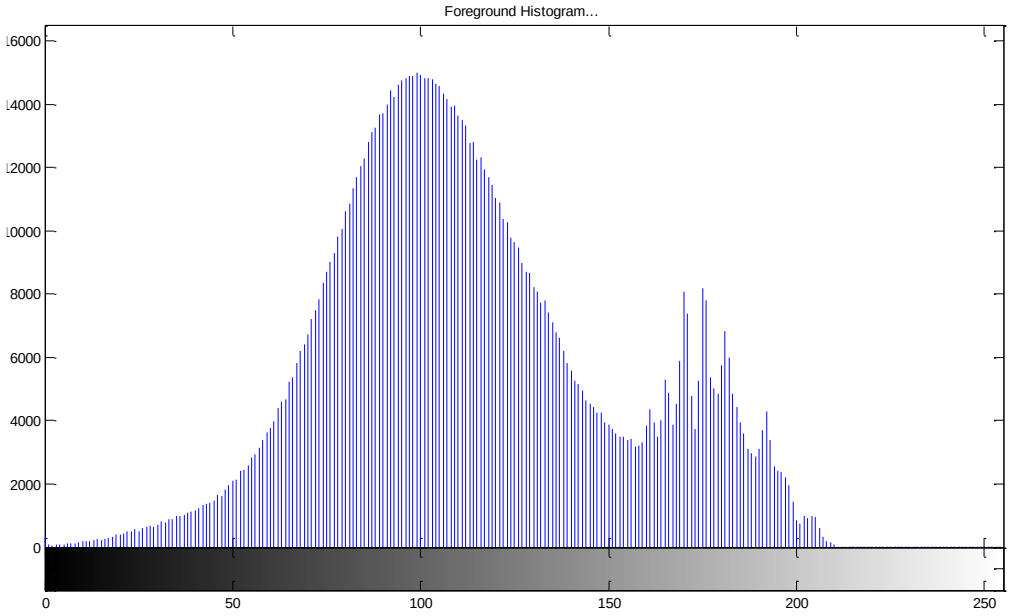

Figure 17. Foreground Histogram.

The useful approach for texture measurement is based on statistical properties of grey level histogram. Many textural descriptors can be derived from statistics of grey level histogram based on their number n[th] order moments from Equation 2. [ShS07]

$$\mu_n = \sum_{k=0}^{x-1}(z_k - m)^n P(z_k) \qquad (2)$$

$z_k: random\ intensity \quad P(z_k): Histogram\ intensity \quad x: number\ of\ intensity\ level$

Some FOSFs, their mathematical expressions, and measure of textures are explained in Table 2.

Table 2. Statistical descriptors. [NiS11]

Descriptor	**Mathematical Expression**	**Measure of Texture**
First moment	$m = \sum_{k=0}^{x-1} (z_k)^1 P(z_k)$	Mean: Measure of average intensity
Second moment	$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}$	Standard Deviation: Average contrast
Variance	$\sigma^2 = (standard\ deviation)^2$	Variability of intensity
Third moment	$\mu_3 = \sum_{k=0}^{x-1} (z_k - m)^3 P(z_k)$	Skewness: Asymmetry of intensity around the mean
Fourth moment	$\mu_4 = \sum_{k=0}^{x-1} (z_k - m)^4 P(z_k)$	Kurtosis: Peaked or Flat of intensity distribution

4.3.2 Second Order Statistical Features (SOSF)

In pattern recognition, texture features are useful features in classification tasks. Generally, texture is a fluctuation in surface described by the sense of touch such as smoothness, coarseness and regularity. There is no mathematical definition of texture: it only provides information about the variation in intensity of surface.

Gray level co-occurrence matrix (GLCM) is the model that quantifies the various textural features defined based on spatial dependencies by Matlab such as contrast, correlation, energy and homogeneity. Haralick 1973 in his paper have suggested a few more parameters for GLCM such as: f1. Uniformity / Energy / Angular Second Moment, f2. Entropy, f3. Dissimilarity, f4. Contrast / Inertia, f5. Inverse difference,

f6. Correlation, f7. Homogeneity / Inverse difference moment, f8. Autocorrelation, f9. Cluster Shade, f10. Cluster Prominence, f11. Maximum probability, f12. Sum of Squares, f13. Sum Average, f14. Sum Variance, f15. Sum Entropy, f16. Difference variance, f17. Difference entropy, f18. Information measures of correlation, f19. Maximal correlation coefficient, f20. Inverse difference normalized (INN), f21. Inverse difference moment normalized (IDN) [HSD73].

14 statistical feature descriptors computed from 32 samples of histopathology images are illustrated in Table 3. These, are unstandardized numerical results values computed by the use of Matlab code in my project: there are 448 statistical values given by 32 histological images represented as rows and 14 feature descriptors as column.

Table 3. Computation of FOSF and SOSF feature descriptors.

Sample	MeanGL	VarianceGl	Standard devi	Skewness	Kurtosis	Contrast	Correlation	Energy	Homogeneit	Range	Local Entropy	Entropy	STD Filter	Smoothness
B1	121,405079	3413,526	58,425388	0,349125	2,50611	0,510517	0,903538	0,080616	0,800744	18,964983	4,934877	7,738637	15,440038	0,999707
B2	132,737447	3780,6849	61,487275	0,204631	2,270267	0,676713	0,884166	0,066301	0,769259	21,881255	4,937739	7,739541	17,944171	0,999736
B3	140,501891	4191,2286	64,739699	0,041081	2,11999	0,682617	0,894715	0,061593	0,770768	22,166405	4,885105	7,731646	18,188778	0,999761
B4	133,79056	3929,4624	62,685424	0,169859	2,204125	0,64811	0,893716	0,067012	0,78075	21,644703	4,877622	7,748251	17,652136	0,999746
B5	124,425684	3446,9182	58,710461	0,305984	2,506979	0,458387	0,914119	0,081028	0,807268	16,545884	4,904007	7,728085	13,992763	0,99971
DCIS2	133,127409	3966,2189	62,977924	0,150759	2,182535	0,719411	0,882867	0,063432	0,769083	23,813141	4,855935	7,774587	18,472129	0,999748
DCIS3	166,574821	4685,3455	68,449583	-0,341639	2,068785	1,000253	0,861619	0,063363	0,739946	29,05457	4,475256	7,223761	22,383713	0,999787
DCIS4	146,623727	4176,8697	64,628707	0,017294	2,124943	0,803927	0,875937	0,05887	0,755529	25,883037	4,767432	7,611404	19,969966	0,999761
DCIS5	132,895242	3740,588	61,160346	0,158705	2,304057	0,562306	0,902948	0,070822	0,78919	19,951641	4,935107	7,746671	16,46406	0,999733
GR1	117,063624	3206,9025	56,62952	0,366364	2,49402	0,602751	0,878868	0,081794	0,7921	21,420603	4,849185	7,751444	16,245772	0,999688
GR10	179,787041	4537,6601	67,362156	-0,575064	2,286186	1,022257	0,853916	0,084306	0,745451	29,294303	4,188647	6,805504	22,546659	0,99978
GR11	148,519352	4497,4882	67,063315	-0,066356	2,025735	0,813385	0,883217	0,057379	0,759323	25,643204	4,695126	7,618501	19,808416	0,999778
GR16	153,498903	4377,1098	66,159729	-0,09462	2,019714	0,955306	0,859038	0,055476	0,740051	28,110699	4,708893	7,548883	21,927685	0,999772
GR17	172,09006	4628,833	68,035527	-0,451282	2,16331	1,028373	0,856202	0,06908	0,739745	30,181367	4,385948	7,087506	22,991665	0,999784
GR18	173,869173	4686,2795	68,456406	-0,473621	2,13696	1,022777	0,858766	0,072671	0,740308	27,897322	4,345669	7,030826	22,021382	0,999787
GR2	158,796294	4673,5499	68,363367	-0,237134	2,040548	0,914674	0,873176	0,05753	0,742212	27,035762	4,679989	7,445285	21,864815	0,999786
GR23	142,942392	5060,0313	71,133897	-0,007525	1,849579	0,89781	0,884925	0,055905	0,757476	26,900604	4,608435	7,675145	20,601833	0,999802
GR24	164,511428	4725,5147	68,74238	-0,341919	2,099938	1,022158	0,859494	0,060062	0,735865	29,529417	4,507037	7,291956	22,837677	0,999788
GR26	161,115661	4877,4906	69,839034	-0,242418	1,971347	1,038554	0,862159	0,059494	0,741006	31,238342	4,439438	7,294003	22,886667	0,999795
GR29	160,471981	4431,8265	66,571965	-0,227081	2,076552	0,967797	0,858656	0,057701	0,74148	30,506426	4,607515	7,400598	22,6327	0,999774
GR3	143,203668	4324,467	65,76068	0,037023	2,063377	0,779907	0,883459	0,058718	0,761555	24,812525	4,806015	7,680379	19,597835	0,999769
GR30	136,591855	3924,6019	62,646643	0,134473	2,217931	0,710642	0,882966	0,064204	0,770519	23,71657	4,876767	7,729637	18,745148	0,999745
GR31	171,505509	4680,9725	68,417633	-0,43585	2,137622	1,085157	0,849551	0,067376	0,734175	31,02242	4,380856	7,098644	23,550598	0,999786
GR32	157,913585	4515,1403	67,194794	-0,17311	2,017699	0,925853	0,867612	0,057546	0,744624	28,243273	4,628484	7,44695	21,596834	0,999779
GR33	162,977283	5057,7544	71,117891	-0,303569	1,961917	1,05776	0,864339	0,060817	0,740664	30,524352	4,409106	7,252295	23,077646	0,999802
GR34	119,654434	3640,5235	60,336751	0,373921	2,387147	0,686935	0,878129	0,074545	0,781856	22,554571	4,816645	7,77162	17,394214	0,999725
GR4	138,131587	4332,6424	65,822811	0,126397	2,071044	0,715837	0,893154	0,062371	0,774748	23,371095	4,736587	7,693155	18,317451	0,999769
GR5	136,355579	4032,2196	63,49976	0,129881	2,207022	0,667735	0,892918	0,063765	0,771514	21,988428	4,849519	7,714689	17,700757	0,999752
GR6	185,776083	4244,5216	65,149993	-0,647814	2,40466	1,256644	0,808112	0,090299	0,725002	35,282353	4,062318	6,586317	25,272789	0,999764
GR8	158,459986	5254,2864	72,486457	-0,215916	1,845403	0,939601	0,884094	0,061981	0,755465	27,586451	4,395389	7,323673	21,001593	0,99981
GR9	117,300788	2900,7302	53,858427	0,314869	2,697101	0,456485	0,898663	0,091071	0,812272	17,19786	4,930241	7,689234	14,159432	0,999655

From Image analysis toolbox we have used other texture analysis functions that filter an image using standard statistical measures. These descriptor provide qualitative information described as smooth, rough, silky, or bumpy (see Table 4). "*In areas with smooth texture, the range of values in the neighborhood around a pixel will be a small value; in areas of rough texture, the range will be larger*. *Similarly, calculating the standard deviation of pixels in a neighborhood can indicate the degree of variability of pixel values in that region.*" [GWE03, Mat14a].

Table 4. Textural Filter Function and their description [Mat14a]

Function	**Description**
rangefilt	Calculates the local range of an image.
Stdfilt	Calculates the local standard deviation of an image.
entropyfilt	Calculates the local entropy of a grayscale image. Entropy is a statistical measure of randomness.

5. Feature Selection Dimensionality Reduction (PCA)

Feature selection have become an essential task in various areas of research: this is due to the large number of features and their correlation [GuE03]. Feature selection seeks to eliminate those redundant information, thus preserving the uncorrelated ones. Another method by PCA was used to reduce dimensionality by converting uncorrelated attributes into correlated variables.

5.1 Feature Selection

In machine learning and statistics, data can contain a lot of redundant and irrelevant information. To overcome the problem of using feature which reduce the performance of classification, we have used a feature selection technique. It is the process of selecting a subset of relevant and uncorrelated features from the original set [GuE03]. The feature selection process by Mark A.Hall (1997) is represented in Figure 18. [HaS97]

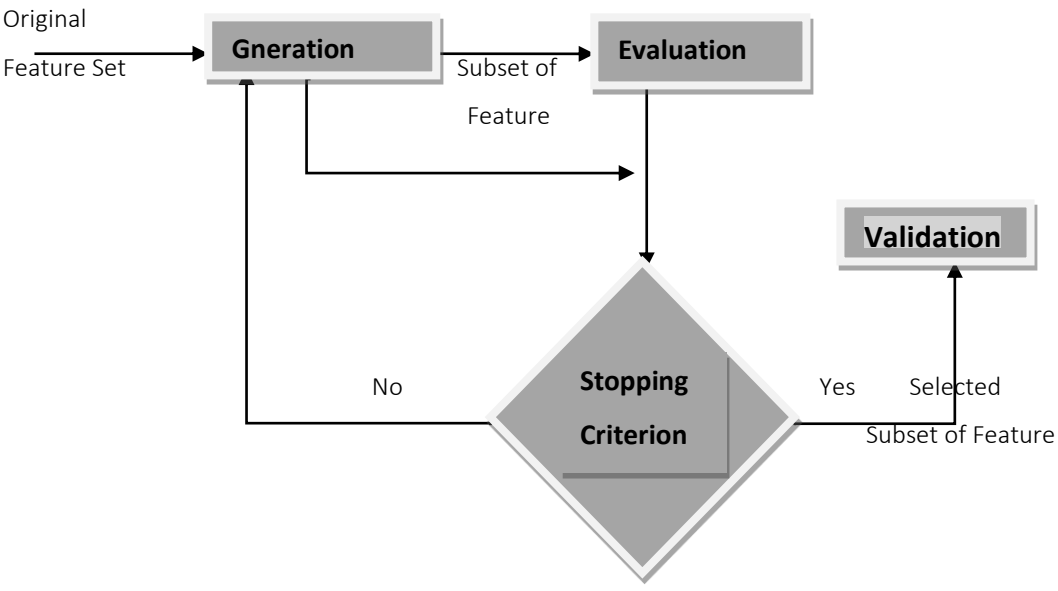

Figure 18. Feature Selection Process.

The large amount of features due to wide diversity of the normal tissues and the variety of the abnormalities can reduce the performance of diagnosis and prognosis. To increase the performance and reduce the redundancy. There are many algorithms used in bioinformatics and statistical data for selecting relevant attributes, e.g., biological datasets of gene expressions containing hundreds of thousands variables. Some of these algorithms for feature selection are summarized in Table 5 [ZhL08].

Table 5. Feature Selection Algorithms. [FSA14]

Algorithms	**Reference**
BlogReg	Gene selection in cancer classification using sparse logistic regression with Bayesian regularization
CFS	Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper
Chi Square	Chi2: Feature Selection and Discretization of Numeric Attributes
FCBF	Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution
Fisher Score	R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification
Information Gain	Cover, T. M. & Thomas, J. A. Elements of Information Theory Wiley, 1991
Relief-F	Computational Methods of Feature Selection
SPEC	Spectral Feature Selection for Supervised and Unsupervised Learning Zheng Zhao & Huan Liu

Another method implemented within Matlab consists of sequential feature selection (SFS) which consists of selecting a subset of features from the data matrix X that best predict the data in y by sequentially selecting features until there is no improvement in

prediction [Mat14c]. The main objective of feature selection is to reduce the dimensionality of the data in order to improve the accuracy and reduce the computational time of the classifier. The reduction can be applied by Principal component analysis (PCA) which is widely applied on datasets. It is a linear dimensionality reduction technique from which we determine a minimal feature subset from the entire set of features.

5.2 Dimensionality Reduction

To map data from high dimension into lower dimension we used principal component analysis (PCA). Hotelling (1933) PCA seeks to lower the dimensionality space by preserving the linear structure of relevant feature intact [Hot97]. It is a statistical procedure that uses an orthogonal transformation to convert a set of correlated attributes into a set of uncorrelated principal components ("PCs") [Mat14c].

Algorithm of PCA is the following:

Input Data Matrix
Output Reduced
Step1: X ← Create N x d data matrix with 1 row vector x_n data point
Step2: X subtract mean x from each row x_n in X
Step3: $\sum$ ← Covariance matrix of X
Step4: Find Eigen vectors and Eigen values of $\sum$
Step5: PCs The N Eigen vectors with largest Eigen values
Step6: Output PCs

Matlab commands of PCA

PCA is simple, non-parametric method for extracting relevant information. It is a data reduction technique that creates components [Hot97]. The main idea behind using the PCA for feature selection is to select components from their largest to their smallest

values of variability [Mat14c]. PCA replaces "*j"* more or less correlated variables by "*k < j"* uncorrelated linear combinations (projections) of the original variables (see Figure 19).

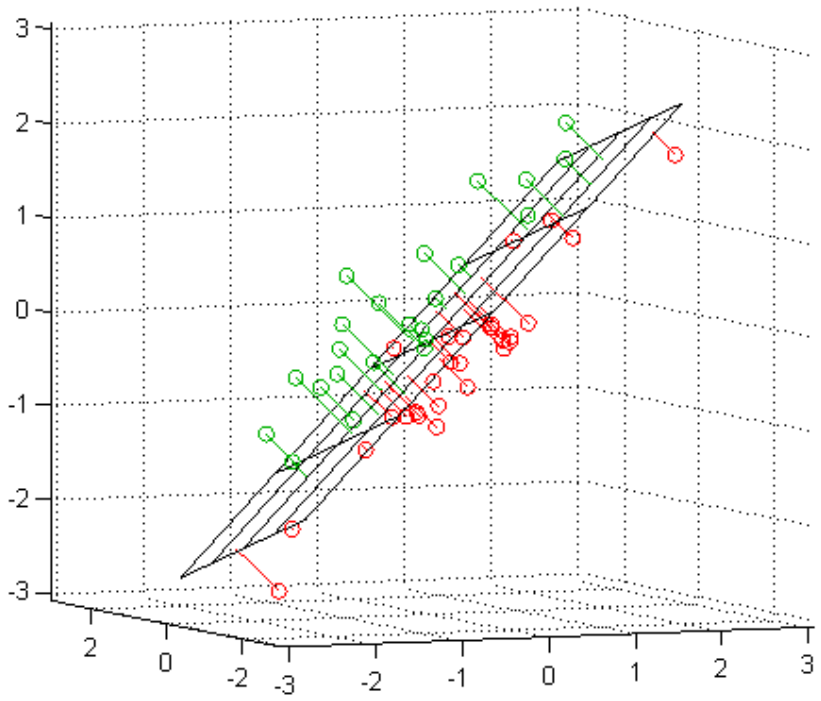

Figure 19. Data projection (PCA).

From Matlab statistical toolbox we have used a predefined functions *princomp, pca, pcacov* and *pcares*. They assume rows as observations and column as attributes for the input data. The outputs consist of Coefficient, Score, Variances, Hotelling's T2 and Explained [Mat14c]:

[coeff score latent tsquared] = ***princomp*** *(zscore(inputdata))*
[coeff score latent tsquared explained] = ***pca*** *(inputdata)*
[residuals reconstructed] = ***pcares*** *(inputdata, dimension)*
[coeff score latent tsquared explained] = ***pcacov*** *(cov(inputdata)).*

1st Output (**Loading):** *Contains the coefficients of the linear combinations of the original variables that generate the principal components.*

2nd Output **(Scores):** *Contains the coordinates of the original data in the new coordinate system defined by the principal components.*

3th Output **(Latent):** *PC Columns Variances.*

4th Output **(Hotelling's T2):** *Statistical measure of the multivariate distance of each observation from the center of the data set. This is an analytical way to find the most extreme points in the data.*

5th Output (**Explained):** *Percentage of the total variance explained by each principal component.*

Interpretation of PCA computed results

Data consists of 32 samples of histopathological images as observations (rows) and their 14 descriptor values as attributes (columns) (see Table 3). Data was submitted to *zscore* (Standardized *z*-scores) command which lead to center the data around the mean: each element of data columns are centered to have mean 0 and scaled to have standard deviation 1.The outputs by PCA are the loading (coefficients) consist of principal components with size of 14x14 sorted from highest to lowest pcs. The plot of second output (scores) shows a new coordinate system defined by the 2 first principal components (see Figure 20).

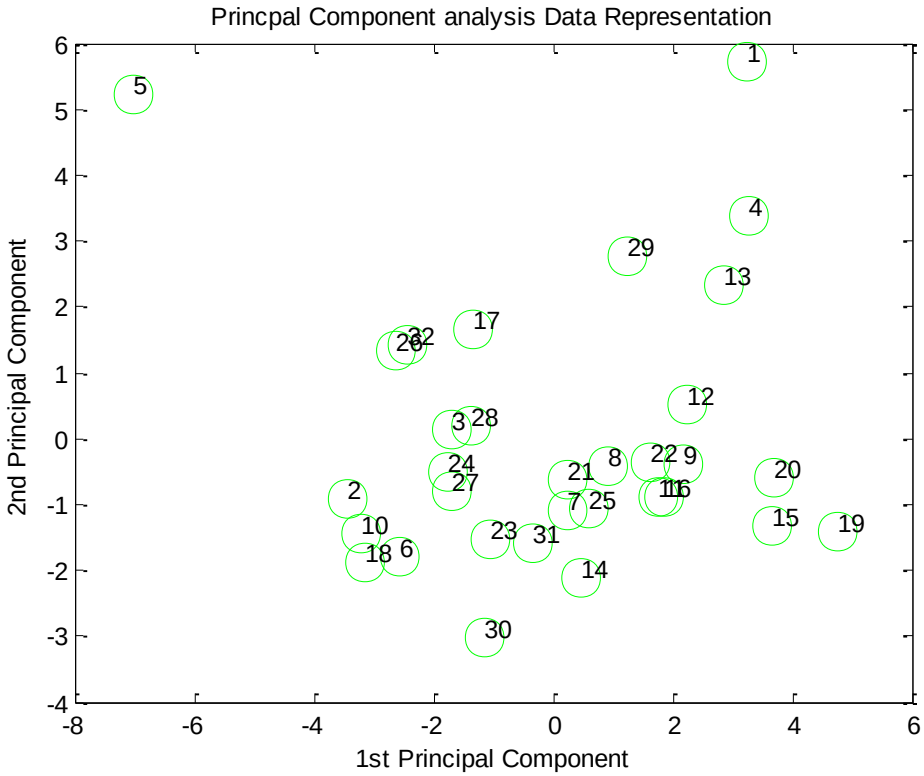

Figure 20. Data Representation in Principal Component Analysis Coordinate.

The variability (latent) containing the percentage variance by the corresponding principal component obtained by PCA using Matlab is shown in Table 6 and the total variance explained by SPSS software is illustrated in Table 7. The table shows that the first component presents the highest variance followed by the second and the third component (values in Bold). While from the fourth component to the last one, the variances are not meaningful.

Table 6. Total variance explained by Matlab.

Latent	Cumulative
49.8977	**49.8977**
29.7717	**79.6694**
13.3286	**92.9980**
3.6161	96.6141
1.0719	97.6860
1.0041	98.6901
0.5283	99.2185
0.3132	99.5317
0.2336	99.7653
0.1042	99.8695
0.0750	99.9445
0.0355	99.9800
0.0192	99.9992
0.0008	100.0000

Table 7. Total variance explained by SPSS.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**6,986**	**49,898**	**49,898**	**6,986**	**49,898**	**49,898**
2	**4,168**	**29,772**	**79,669**	**4,168**	**29,772**	**79,669**
3	**1,866**	**13,329**	**92,998**	**1,866**	**13,329**	**92,998**
4	,506	3,616	96,614			
5	,150	1,072	97,686			
6	,141	1,004	98,690			
7	,000	,001	100,000			
.						
.						
14						

Cumulative variances shows the first four principal components (see Figure 21) that retain the relevant information which will be used to classify data. The Cumulative from Table 6 shows that 93% of the total variances was reached by 3 first components. The first component explains about 50%, added to the second component gives 80% which is more than two third of the total variances. Thus it is efficient to reduce the

original variables into a lower number of orthogonal non correlated components (factors).

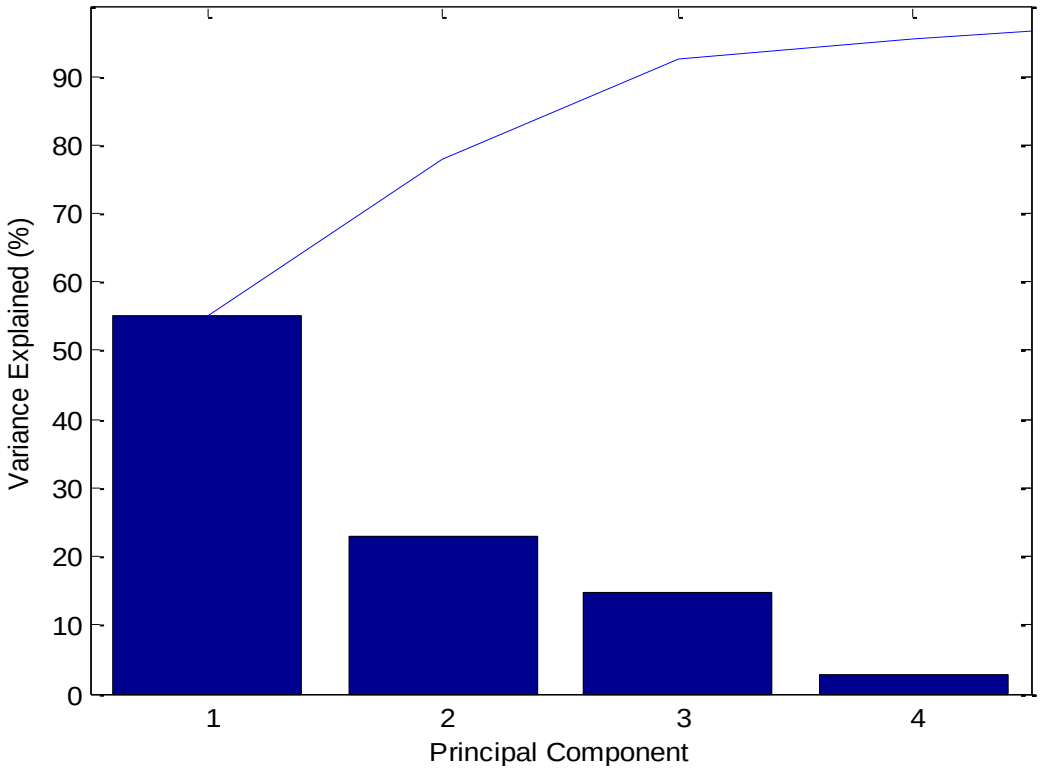

Figure 21. Cumulative variances.

To visualize two and three principal component coefficients for variables and principal component scores for observation in a single plot we have used the plots in 2D and 3D respectively (see Figure 22).

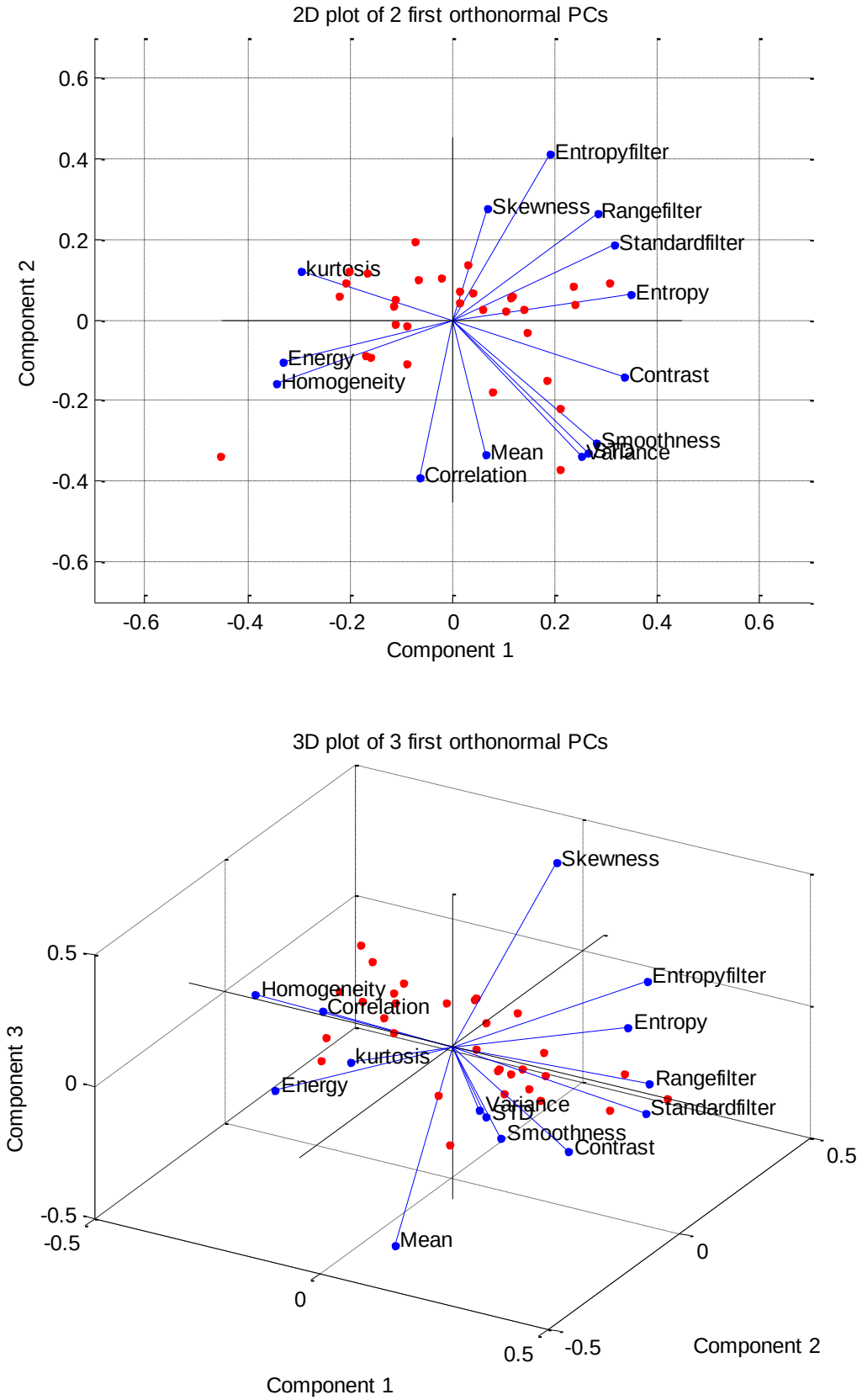

Figure 22. Orthonormal pcs coefficients of variable and the pcs scores.

2D bi-plot in Figure 22 represents observations (red dots) and variables (blue vectors). Direction and length for each variable vector explain how they contribute to the principal components. The first component has the largest coefficients for contrast, entropy, standard filter, range filter, entropy filter, smoothness, variance, standard deviation, skewness and mean. For the second component, the largest coefficient are kurtosis, homogeneity energy and correlation. SPSS software was used in order to simulate the PCA and extract the interesting table values. Skewness, mean and correlation seems not loading for the first pcs (see Figure 23). Thus we have to reduce the original variables to 11 instead of 14 variables which was held by a rotation (technique used by the PCA that projects the remaining data after a rotation). The result in Table 8 shows the importance of data reduction technique (PCA) since the total explained variance increase with 11 variables. Two principal components were extracted which means that the information are revealed in two first components (2nd component of cumulative % increased from 79% to 89%).

Components	1	2	3
Mean	0,1764	,682	,659
Variance	,672	,687	-,153
STD	,705	,676	-,113
Skewness	,186	-,563	-,721
Kurtosis	-,779	-,247	,368
Contrast	,895	,290	,212
Correlation	-,163	,796	-,532
Energy	-,873	,209	,311

Figure 23. Some variables and their loading value after 1 rotation.

Table 8. Total Variance Explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6,899	62,715	62,715	6,899	62,715	**62,715**
2	2,882	26,196	88,911	2,882	26,196	**88,911**
3	,836	7,597	96,509			
4	,152	1,379	97,888			
5	,098	,892	98,780			
6	,083	,755	99,535			
7	,030	,273	99,808			
8	,012	,109	99,917			
9	,005	,050	99,966			
10	,004	,032	99,999			
11	,000	,001	100,000			

PCA seeks to select those uncorrelated features that could distinguish between different classes. For the next work (classification) we will use some of the features extracted from the 2D bi-plot such as kurtosis and contrast, since variable contrast contributes positively to the first pc and negatively to the second pc and vice versa for variable kurtosis. Another method which is quite efficient is to use the reconstructed data by the predefined function in Matlab *pcares*, since the number of principal components which retain the two third of information was determined (extracted from Table 8).

6. Classification Techniques of Histopathological Images (SVM, LDA, and NN)

In machine learning that focuses on classifying objects /patterns, the separation of images is a vital challenge that face medical image analyzers. Pattern recognition is a scientific discipline in machine intelligence systems aimed to make decision based on recognized patterns. Computer-aided diagnosis where pattern recognition is used can assists doctors in making diagnostic decisions [TPK10]. Classification methods are procedures from which we assign the object to its specific category. The classifier is a decision boundary (linear or non-linear): its main role is to divide the feature space into regions that distinguish between different classes. There are two types of classification techniques in machine learning (see Figure 24): supervised and unsupervised. [Mat14c, TPK10]

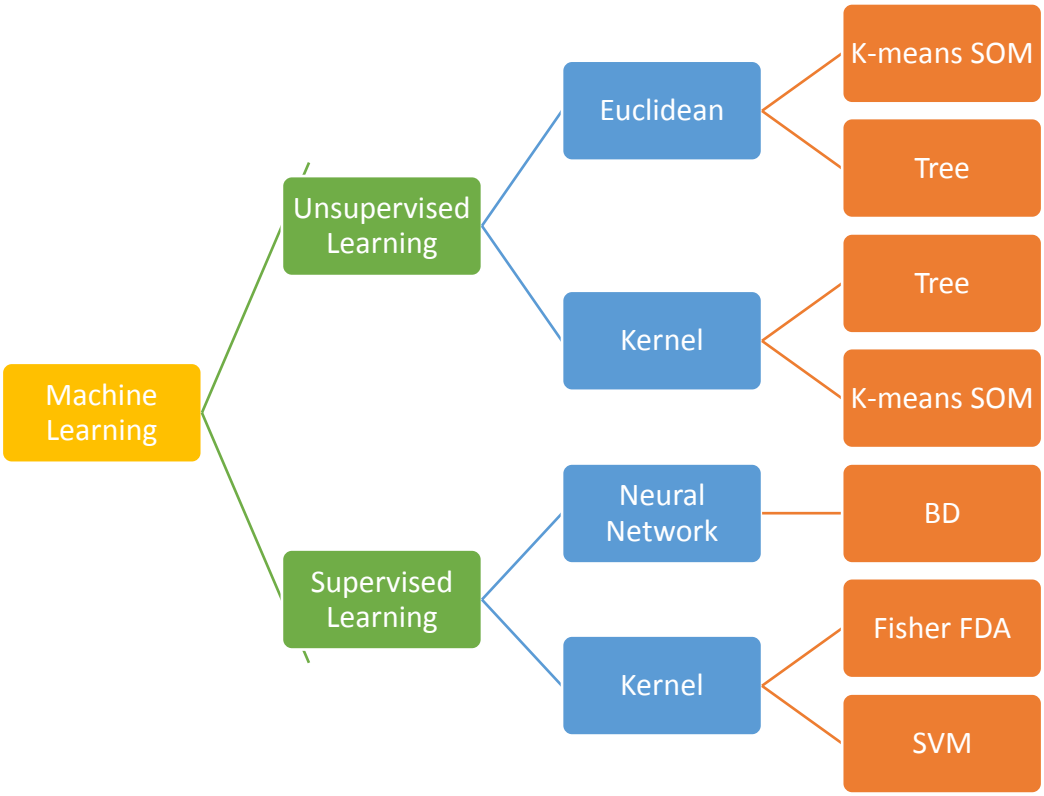

Figure 24. Machine Learning Algorithms. [Ksy90]

Supervised Learning is a heuristic in which the input and output data are known a priori and the training data is provided to guide the classification (see Figure 25). Unsupervised Learning is a heuristic that tries to find hidden structure of unlabeled data and thus it is a clustering method of data into groups. There are no information about input-output a priori (no training data).

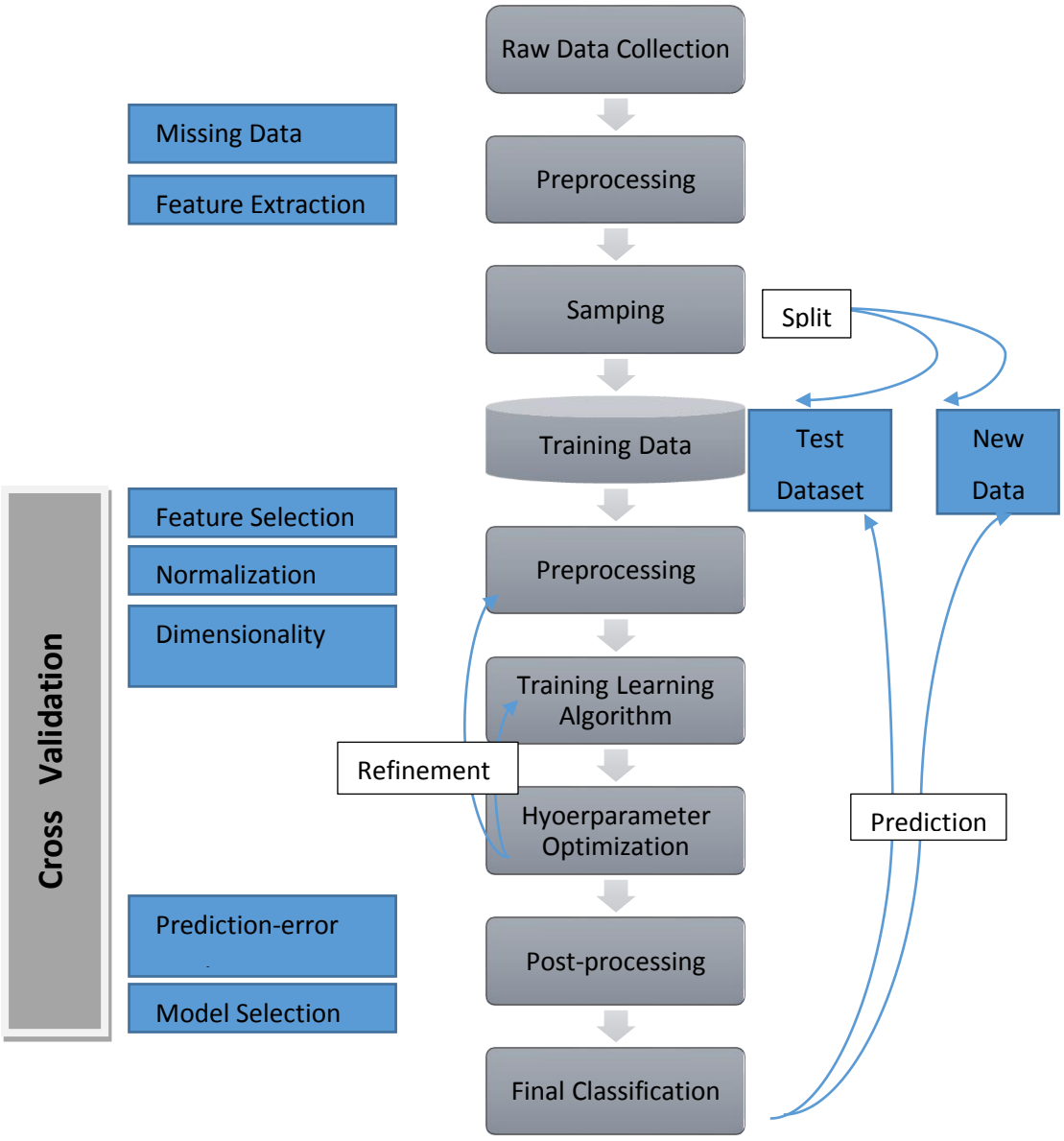

Figure 25. Flow diagram of supervised learning. [Rse14]

Table 9 is showing characteristic of different supervised learning algorithms, in some tasks the table can be inaccurate: [Mat14c]

> *"* — SVM prediction speed and memory usage are good if there are few support vectors, but can be poor if there are many support vectors. When you use a kernel function, it can be difficult to interpret how SVM classifies data, though the default linear scheme is easy to interpret.*
>
> *** — Naive Bayes speed and memory usage are good for simple distributions, but can be poor for kernel distributions and large data sets.*
>
> **** — Nearest Neighbor usually has good predictions in low dimensions, but can have poor predictions in high dimensions. For linear search, Nearest Neighbor does not perform any fitting. For kd-trees, Nearest Neighbor does perform fitting. Nearest Neighbor can have either continuous or categorical predictors, but not both.*
>
> ***** — Discriminant Analysis is accurate when the modeling assumptions are satisfied (multivariate normal by class). Otherwise, the predictive accuracy varies."*

Table 9. Characteristics of Supervised Learning Algorithms. [Mat14c]

Algorithm	Predictive Accuracy	Fitting Speed	Prediction Speed	Memory Usage	Easy to Interpret	Handles Categorical Predictors
Trees	Medium	Fast	Fast	Low	Yes	Yes
SVM	High	Medium	*	*	*	No
Naive Bayes	Medium	**	**	**	Yes	Yes
Nearest Neighbor	***	Fast***	Medium	High	No	Yes***
Discriminant Analysis	*****	Fast	Fast	Low	Yes	No

In this part of the work, we will focus on supervised learning algorithms such as discriminant analysis (DA) with its two methods linear and quadratic to predict the right pattern, support vector machine (SVM) for linearity and non-linearity to find a line in 2D space for separation between classes. In the last part of this section a graphical user interface of supervised neural network for pattern recognition is used to classify the histological images.

6.1 Discriminant Analysis

In statistic and machine learning, Discriminant Function Analysis (DA) or Fisher's linear discriminant predicts an outcome by undertaking the multiple linear regression tasks as PCA (see Equation 3). It involves the linear combination of variables which explain the data to predict the category of pattern: [Fra36]

$$D = v1X1 + v2X2 + v3X3 = \cdots viXi + a \quad (3)$$

Where D = discriminate function
v = the discriminant coefficient or weight for that variable
X = respondent's score for that variable
a = a constant
i = the number of predictor variables

Discriminant Analysis is characterized by some properties such as: [Mat14c]

- Good for simple problems and few training samples.
- Each class (Y) generates data (X) using GMD (Gaussian Multivariate Distribution) (see Figure 26).
- Linear discriminant analysis: same covariance matrix, only the means vary.
- Quadratic discriminant analysis: both means and covariance of each class vary.

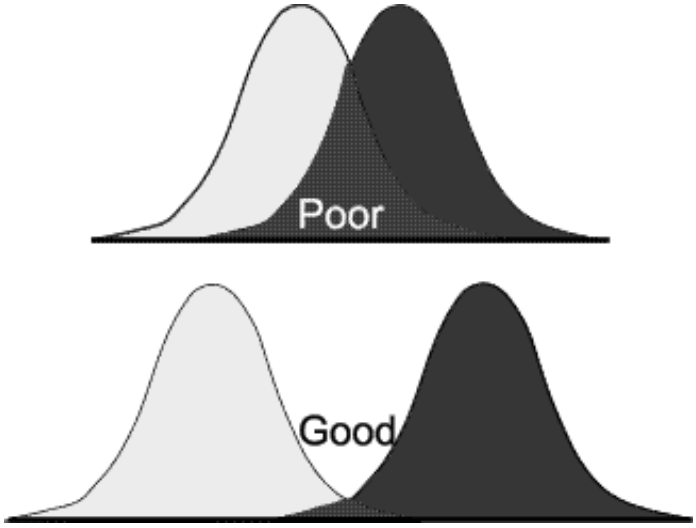

Figure 26. Gaussian Multivariate Distribution GMD of Scores "DA"-

The classification consists of applying two methods: Linear Discriminant Analysis (LDA) and Quadratic Discriminant analysis (QDA) with the selected features for differentiating normal and cancerous tissues. Mardia Kurtosis Test for Linear and Quadratic Discriminants was used in order to determine the consistency of data with the multivariate normal distribution, thus to decide if suitable to use discriminant analysis. [Mat14c]

Linear Discriminant Analysis method

Data consists of two features (variables) selected from the original features and 32 observations. *gscatter* function was used to create a scatter plot of feature contrast and feature kurtosis (see Figure 27).

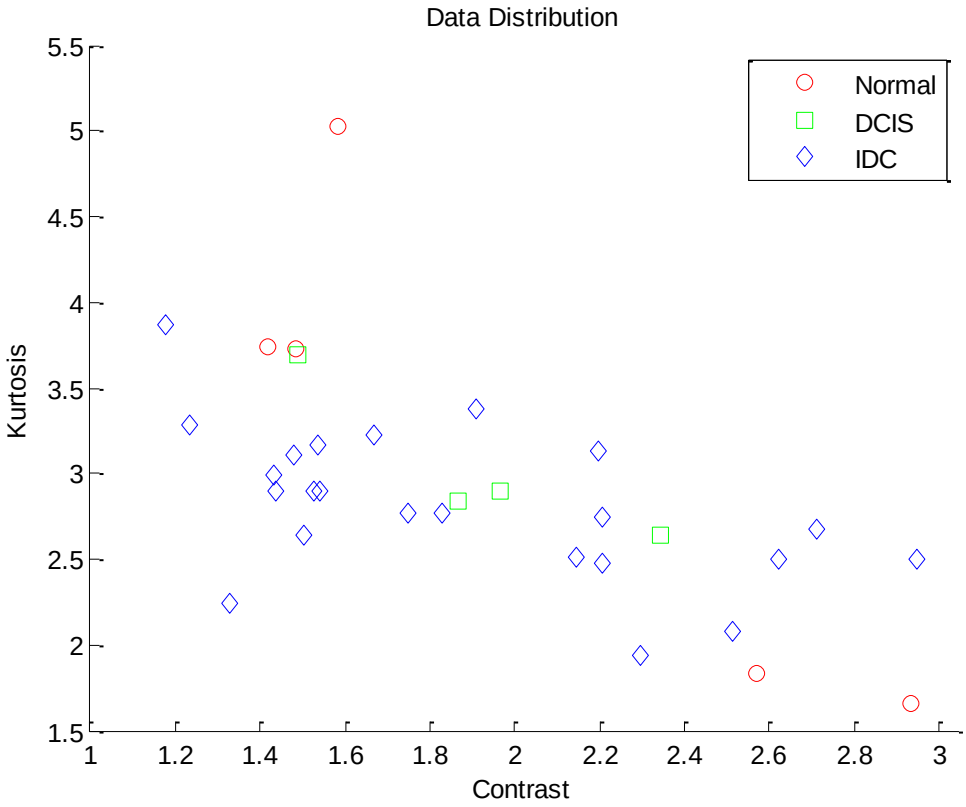

Figure 27. Scatter plot of data on 2D dimensional feature space.

The plot of LDA classifier (black line) in Figure 28 shows the distinction between the two classes red and blue circles with some misclassified data.

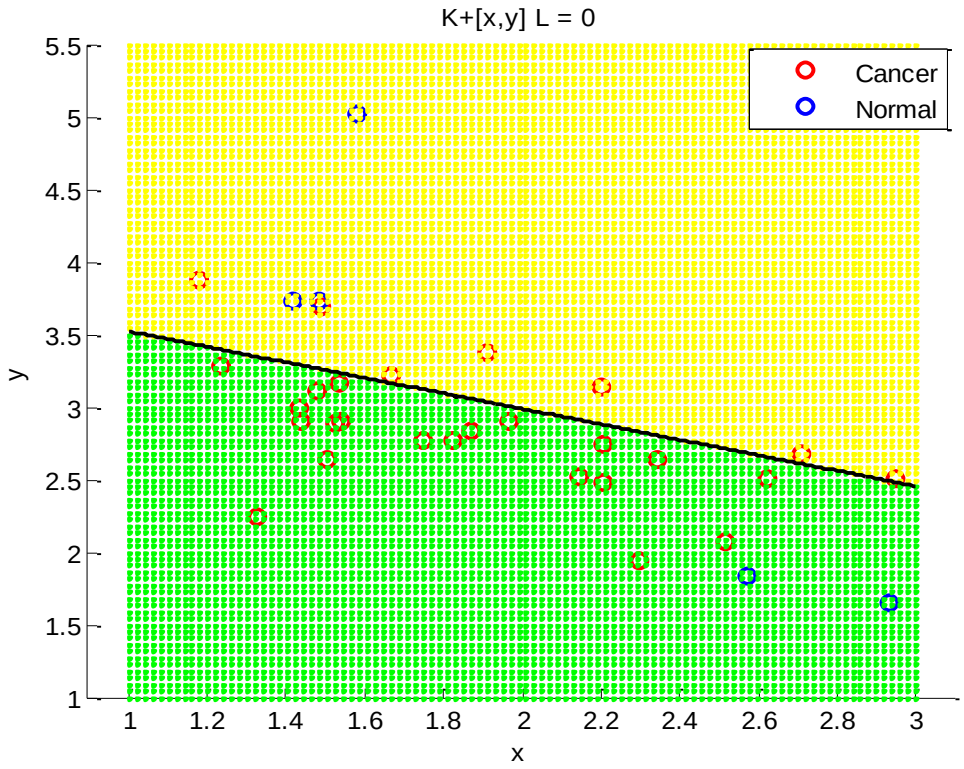

Figure 28. Linear Discriminant Analysis Classifier.

The idea from classification is to measure the performance of the classifier that is, to determine the classification errors, the number of correct classified data and the number of misclassified data (see Table 10).

Table 10. Confusion Matrix of LDA with 2 discriminant functions type.

Type of discriminant function	**diaglinear**				**Linear**			
	Class 1		**Class 0**		**Class 1**		**Class 0**	
Class 1	20	75%	7	25%	18	66%	9	34%
Class 0	2	40%	3	60%	2	40%	3	60%

	Total %	
	diaglinear	**linear**
Correct classified	72%	65%
Misclassified	28%	35%

Interpretation of Table 10:

We have used two types of discriminant functions for LDA linear and diaglinear: [Mat14c]

- *"linear: Fits a multivariate normal density to each group, with a pooled estimate of covariance. This is the default.*
- *diaglinear: Similar to linear, but with a diagonal covariance matrix estimate (naive Bayes classifiers)".*

From Table 10 above we see that the correct classification for diaglinear type for class 1 is 75% and for class 0 is 60%. The misclassified data is 25% for class 1 and 40% for class 0. The same interpretation for the linear type 66% and 60% were in correct classification, while for the misclassification 34% and 40% respectively for class1 and class 0.

The total correct classification of data by diaglinear type is about 72% which seems to be a good discriminant function unless linear discriminant function which falls by 6% in assigning the right pattern.

Quadratic Discriminant Analysis method

Quadratic discriminant function does not assume homogeneity of variance-covariance matrices. The means and covariance of each class vary: QDA discriminates two or more classes by a defined quadric surface (see Figure 29). [Fra36]

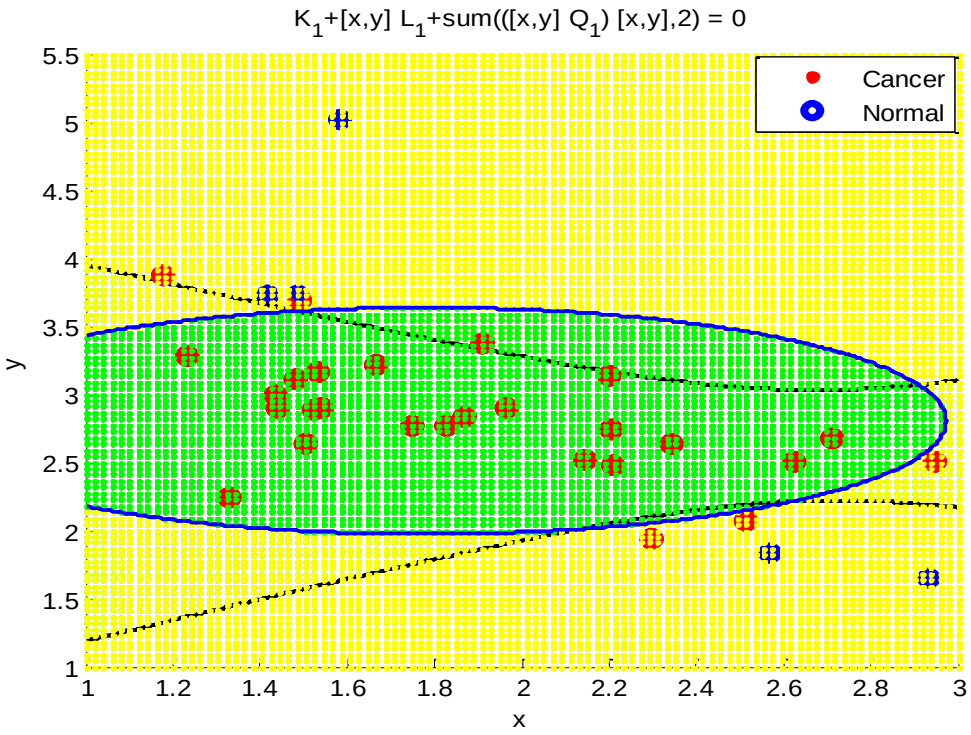

Figure 29. Quadratic Discriminant Analysis classifier type diaglinear.

The discriminant function type quadratic separates two classes normal and cancerous using quadratic surface shown in Figure 29 with blue color of spherical shape. While the classifier type quadratic shows different shape in black with the percent classification error is still the same for both types.

Table 11 shows the classifier (quadratic discriminant) of linear and diaglinear type. The classification percentage remains the same in both types. While the performance of the classifier has increment by 9% from the LDA towards QDA for diaglinear type and by 19% for linear type.

Table 11. Confusion Matrix of QDA with two discriminant function types.

Type of discriminant function	**diaglinear**				**Linear**			
	Class 1		**Class 0**		**Class 1**		**Class 0**	
Class 1	22	81%	5	19%	22	81%	5	19%
Class 0	0	0%	5	100%	0	0%	5	100%

	Total %	
	diaglinear	**Linear**
Correct classified	84%	84%
Misclassified	16%	16%

6.2 Support Vector Machine

In machine learning, support vector machine (SVM) is a model listed as supervised learning. SVM training algorithm builds a classifier that assign data to the right category. The aim of SVM is finding a line in 2D space, plan in 3D space and hyperplane in higher dimension space that separates two classes with maximum margin (maximal width of the slab parallel to the hyperplane that has no interior data points) (see Figure 30). [Mat14c]

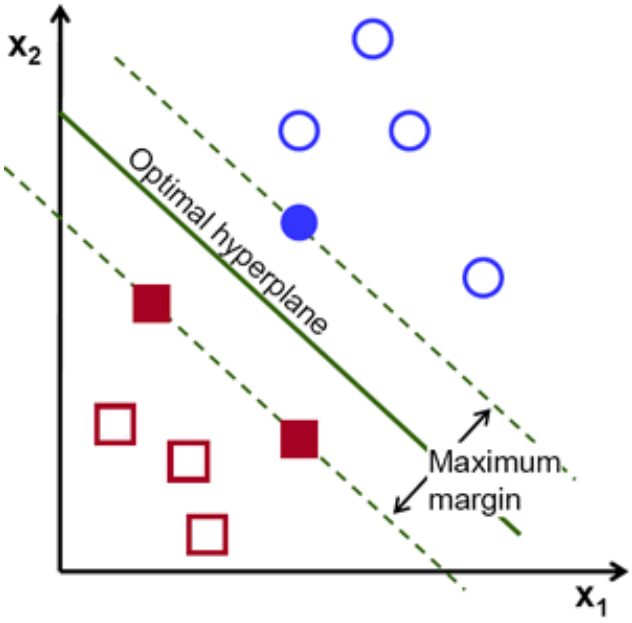

Figure 30. Support Vector Machine.

The filled blue circles and filled brown squares form the support vector that separate the two classes shown in Figure 30. Support vector machine is composed of **N** given support vectors **z** and a set of weights **w**. The computation of the output is given by Equation 4: [CoV95]

$$F(x) = \sum_{i=1}^{N} w_i \ < z_i, x > + \mathrm{b} \qquad (4)$$

SVM uses a set of support vectors for discrimination with efficient memory use. It uses Kernel functions for the decision function. The disadvantages of support vector machines include no probability estimates: they can be computed by using a k-fold-cross-validation and when there are more features (dimensions) than the number of observation. [LWe04]

We have used two SVM methods, linear and non-linear, in order to check which one performs well with kurtosis and contrast for class distinctions. Sometimes data are not linearly separable which m not allow for a separating hyperplane. In that case, SVM can use a soft margin, meaning a hyperplane that separates many, but not all data points. [Mat14c]

Linear SVM

We assume that data is separable in two classes. The cross validation was used in order to generate indices. The training is about 70% of data and 30% of data was reserved for test. Matlab predefined function was c*rossvalind* that returns logical index vectors for cross-validation of N observations by randomly selecting P*N (approximately) observations to hold out for the evaluation set. P must be a scalar between 0 and 1 (P value taken is 0.6) [Mat14c]. The resulting plot of linear SVM is depicted in Figure 31.

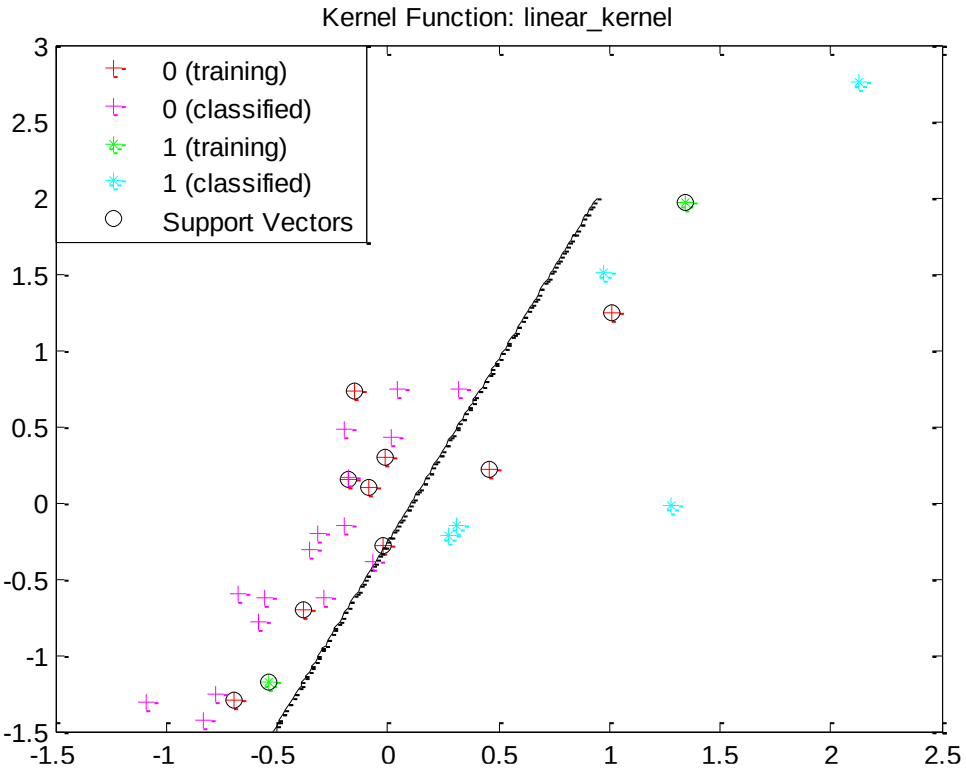

Figure 31. Linear SVM separation.

The result of different classifiers were compared using ROC Parameters which is an effective of evaluating the performance of the classifier in diagnostic tests such as accuracy percentage (correct classified and misclassified), sensitivity percentage and specificity percentage. The expressions used are TP, TN, FP, and FN. [GRN13, Ona03]

ACCURACY = (TP+TN)/(TP+TN+FP+FN)

SENSITIVITY= (TP)/(TP+FN)

SPECIFICITY= (TN)/(TN+FP)

The different fractions (TP, FP, TN, FN) are represented in Table 12.

Table 12. Schematic outcomes of a test. [Ona03]

	Disease	
Test	**Present**	**Absent**
Positive	*True Positive (TP)*	*False Positive (FP)*
Negative	*False Negative (FN)*	*True Negative (TN)*

The result of Linear SVM classifier in Table 13 shows that 80% of the data were correct classified (only 20% are misclassified). The sensitivity shows 84% which means that there is a presence of false negative data FN. While the specificity is 67% that means presence of false positive data FP.

Table 13. ROC parameters Result by linear SVM.

CorrectRate	**ErrorRate**	**Sensitivity**	**Specificity**
80%	**20%**	**84%**	**67%**

Non Linear SVM

The algorithm of SVM can be mapped into a higher dimension feature space for non-linearly separable features. The visualization of data in higher-dimensional feature space is infeasible. It is straightforward to apply kernel functions aiming to train SVM classifiers with feature Kernel which are easier to compute (See Figure 32). [Mat14c, Scd15]

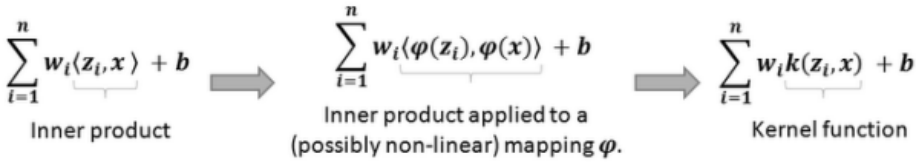

Figure 32. Kernel applied to the original Support Vector Machine.

Assuming data non-separable, we used the RBF kernel (Radial Basis Function Kernel) meaning that non-linear function is learned by linear function. Kernel transforms the data into higher dimensional space in order to apply a linear separation.

The predefined functions *svmtrain* was used with the sigma value 0.8. The penalty parameter C is called BoxConstraintValue in MATLAB. *svmtrain* uses a default value of 1: we chose the value of $2e^{-1}$. To find the optimal values for (sigma, C) we used cross validation with 70% for data training. The plot of non-linear SVM is illustrated in Figure 33.

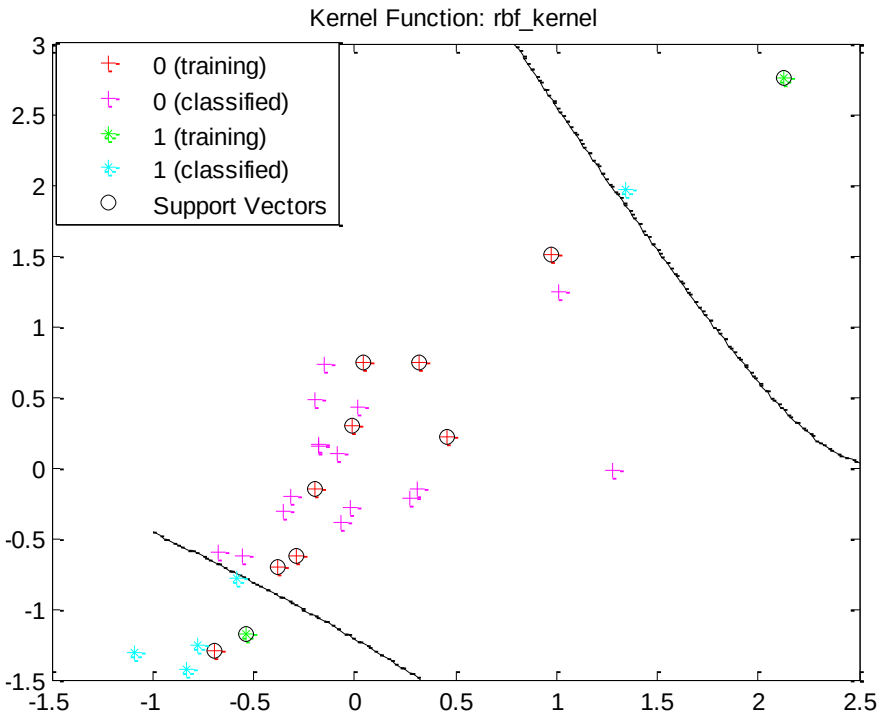

Figure 33. Non-Linear SVM separation.

The result from non-linear SVM classifier in Table 14 shows that 75% of the data were correctly classified (25% are misclassified). Compared to linear classifier there is an increase of errors by 5%. The sensitivity shows 80% which means that there is a presence of false negative data FN. Mathematically we can interpret the decrease of sensitivity by an increase of FN or decrease of TP. While the specificity is 41% that means presence of false positive data FP. Thus linear SVM classifier performs better than non-linear which can be explained in such a way that the data is separable in two classes.

Table 14. ROC parameters Result by non-linear SVM.

CorrectRate	ErrorRate	Sensitivity	Specificity
75%	**25%**	**80%**	**41%**

6.3 Neural Network

Currently NNs (Neural Networks) provide a good solutions to various problems in image recognition, character recognition, speech recognition and in medical image analysis for diagnosis / prognosis of certain diseases. Neural networks use learning algorithms that are inspired from biological nervous system by how neurons learn (see Figure 34) [TPK10]. Their evaluation depend on their efficiency with applications. NN Toolbox supports supervised learning with feedforward, radial basis, and dynamic networks. It also supports unsupervised learning with self-organizing maps and competitive layers. [Mat14c].

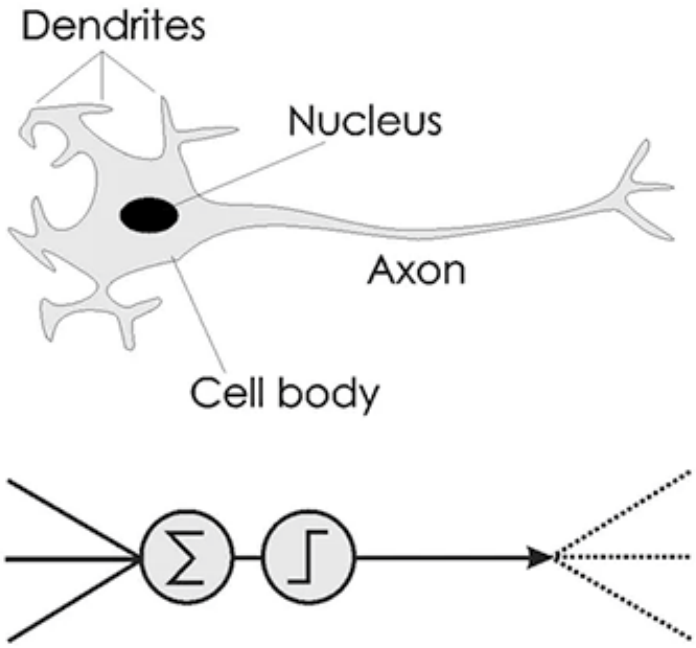

Figure 34. Biological Neuron.

Similarly perceptron organized in NNs provide an algorithm able to learn to distinguish between classes such as cancerous and non-cancerous histological images. The features form the inputs in input layer (see Figure 35). [Arb13]

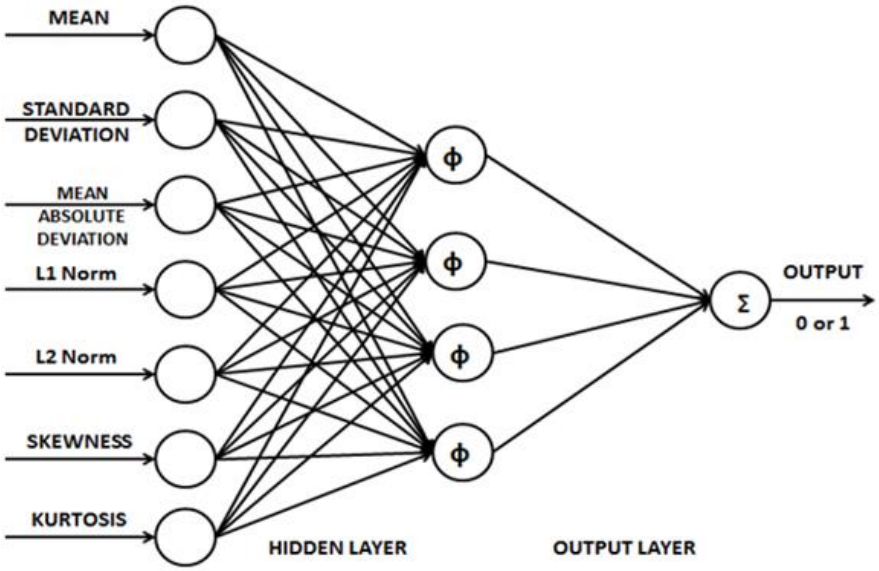

Figure 35. Feedforward Neural Network.

The classifier of NNs considered as supervised will separate the normal and tumoral histopathological images. The input layer consists of original 14 features and the 11 features extracted by the PCA and the reconstructed data by the PCA for the two first components. The performance is measured and compared. The hidden layer taken is about the half of the sum of inputs and outputs and the output consists of one neuron. The output outcome is 1 or 0 respectively, i.e., cancerous or normal. In the command window the predefined function *nprtool* launches the graphical user interface NN pattern recognition (see Figure 36). It is trained using algorithm by choosing the training, target and testing sets for evaluation. The weights are initialized randomly at each execution. The network continue updating during the epochs until the validation reaches the minimum error (MSE between desired and actual outputs) which stops the training "Performance plots" (see Figure 37).

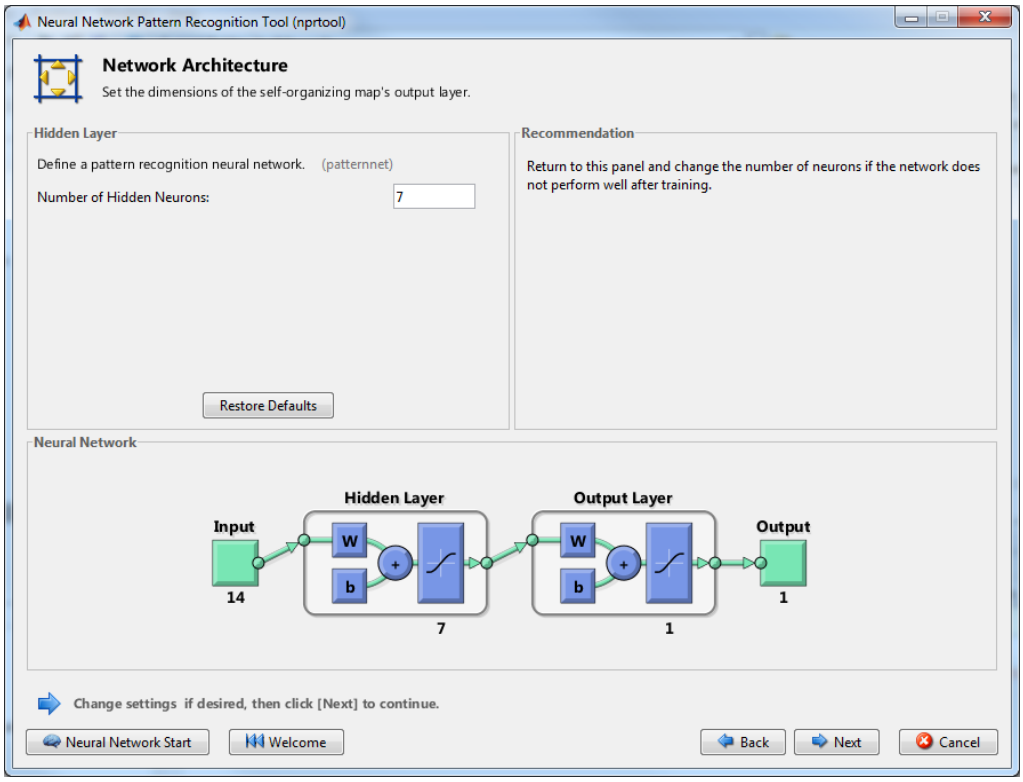

Figure 36. Neural Network Pattern Recognition Tool.

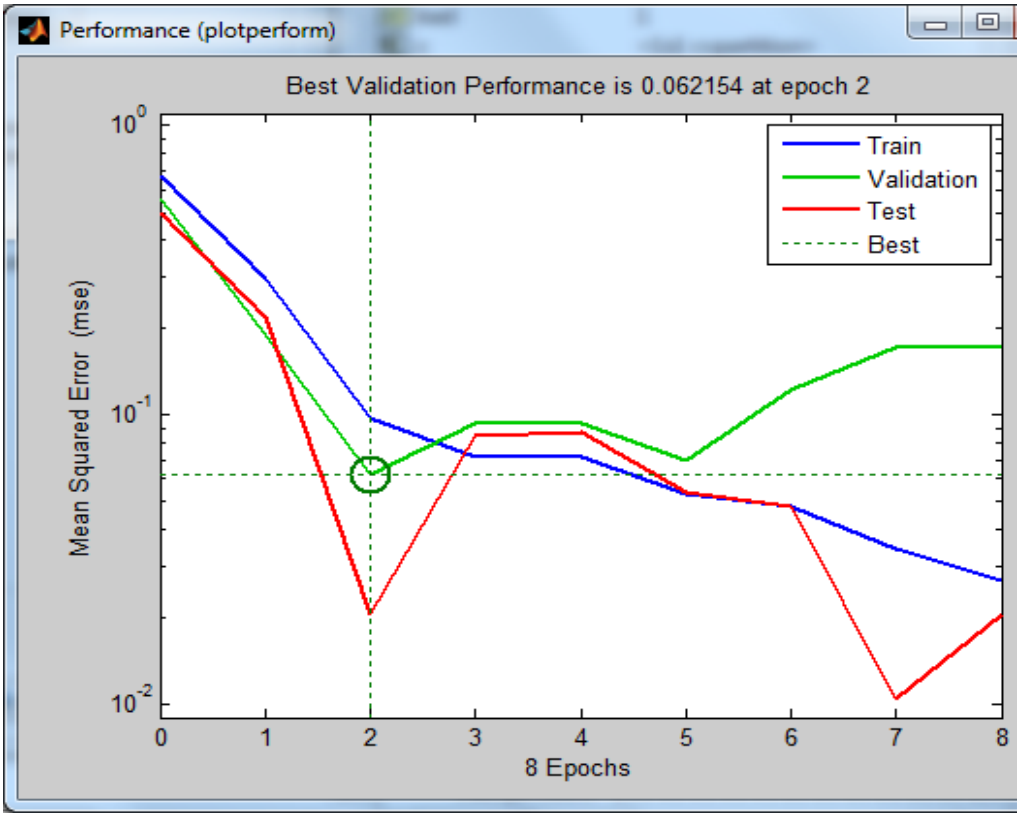

Figure 37. Performance plot of NNPR.

NNPR with 14 Inputs

The inputs consist of 14 extracted statistical features standardized to mean 0 and standard the deviation 1 by the use of predefined Matlab function *zscore*. Since we have two classes we gave the value of 1 to cancerous and 0 to normal, indicating the class of the corresponding input. The hidden layer consists of seven neurons. Data was divided on 70% for training, 15% for validation and 15% for testing. The results are presented in Figure 38 for the performance and Figure 39 for the confusion matrix.

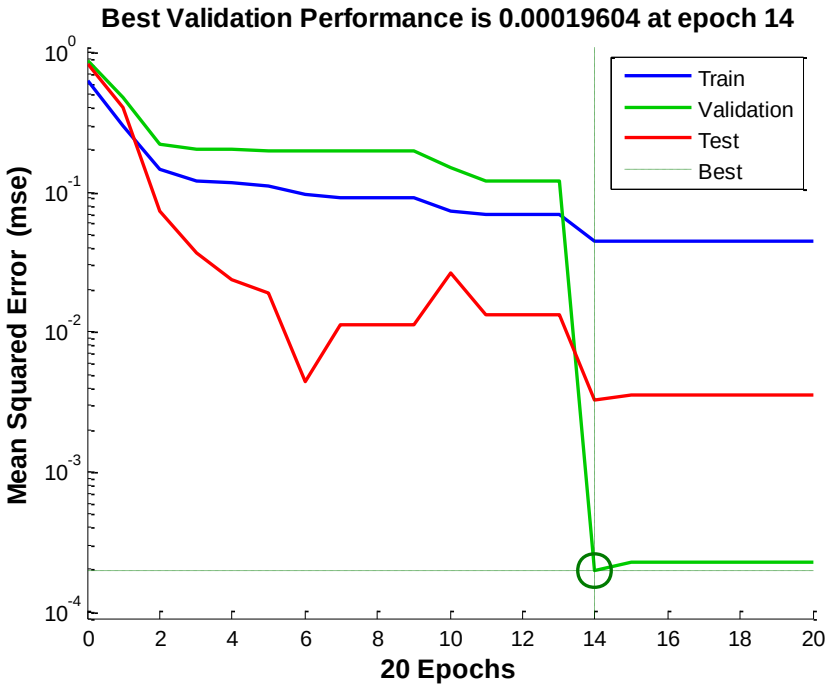

Figure 38. Performance of NNPR with 14 inputs.

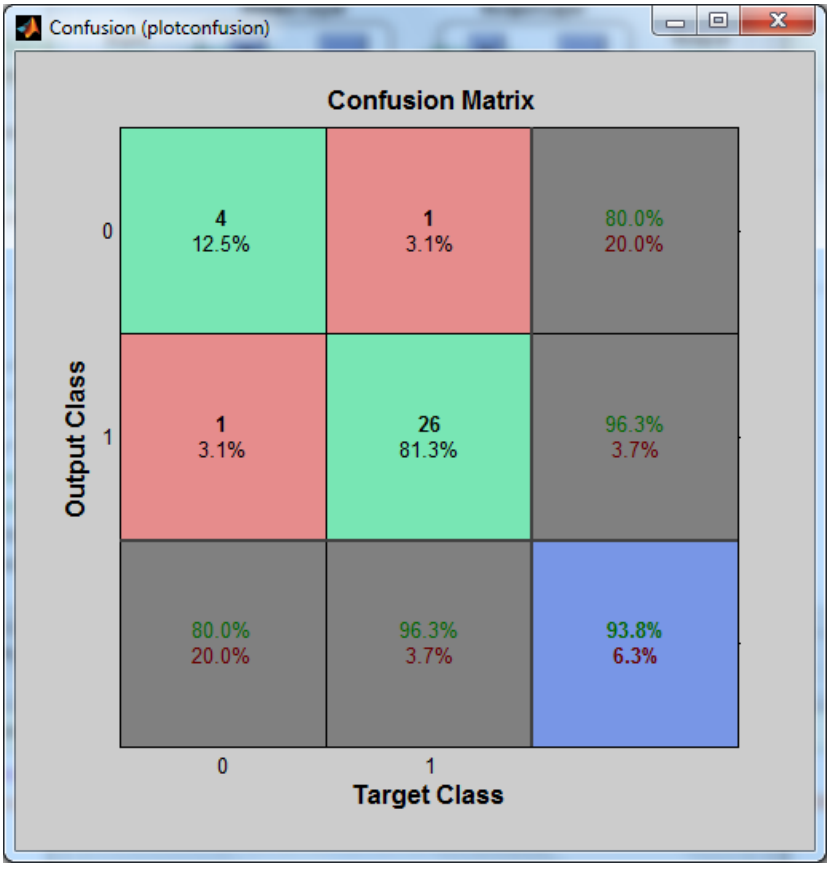

Figure 39. Confusion Matrix of NNPR with 14 inputs.

Figure 39 shows the confusion matrix of 14 inputs with neural network pattern recognition. The results in diagonal green squares show the correct classification and red color represent those incorrect responses. The accuracy of the networks shows 94% in the blue square in green text color. The result was reached after 20 epochs and the execution time was a bit slower.

NNPR with 11 Inputs Obtained by Dimensionality Reduction

Same network was used except for the inputs which has been reduced to 11 features extracted by the PCA. The confusion matrix in Figure 40 shows the total accuracy of 97%. The execution time was faster than 14 features with only 8 epochs.

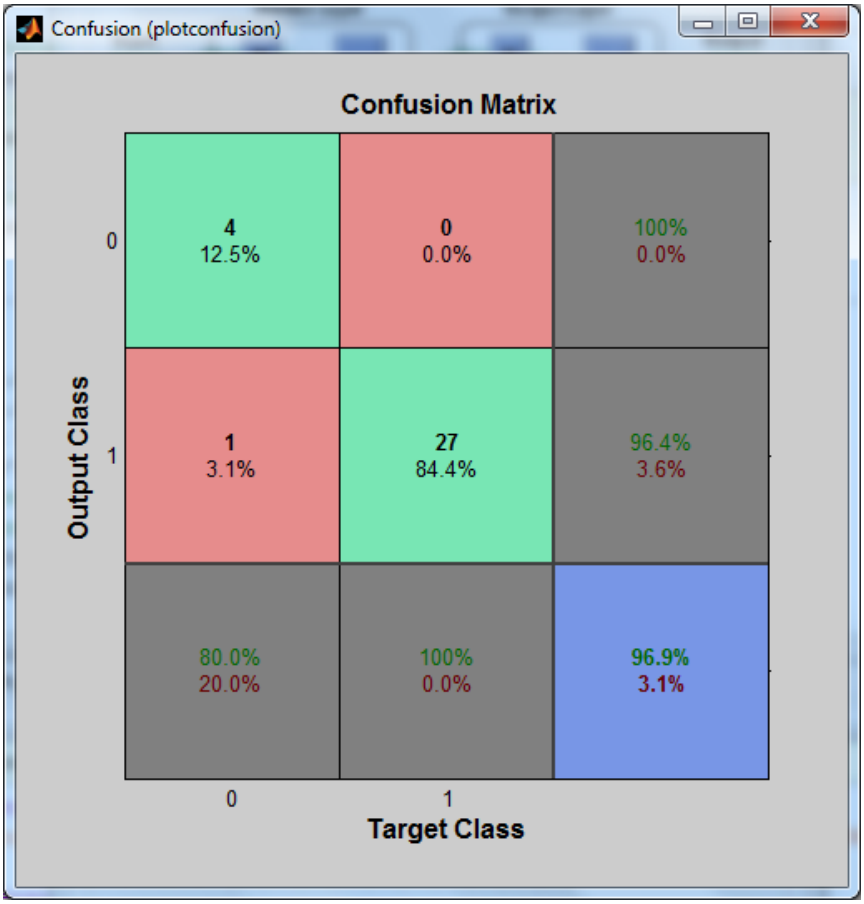

Figure 40. Confusion Matrix of 11 features (Dimensionality Reduction).

NNPR with Reconstructed Data 14 (PCs)

In the last method of NNPR we used the reconstructed data by *pcares* function for the 2 first components as inputs (same dimension as original data = 14). The total accuracy shows 97% of correct responses (see Figure 41).

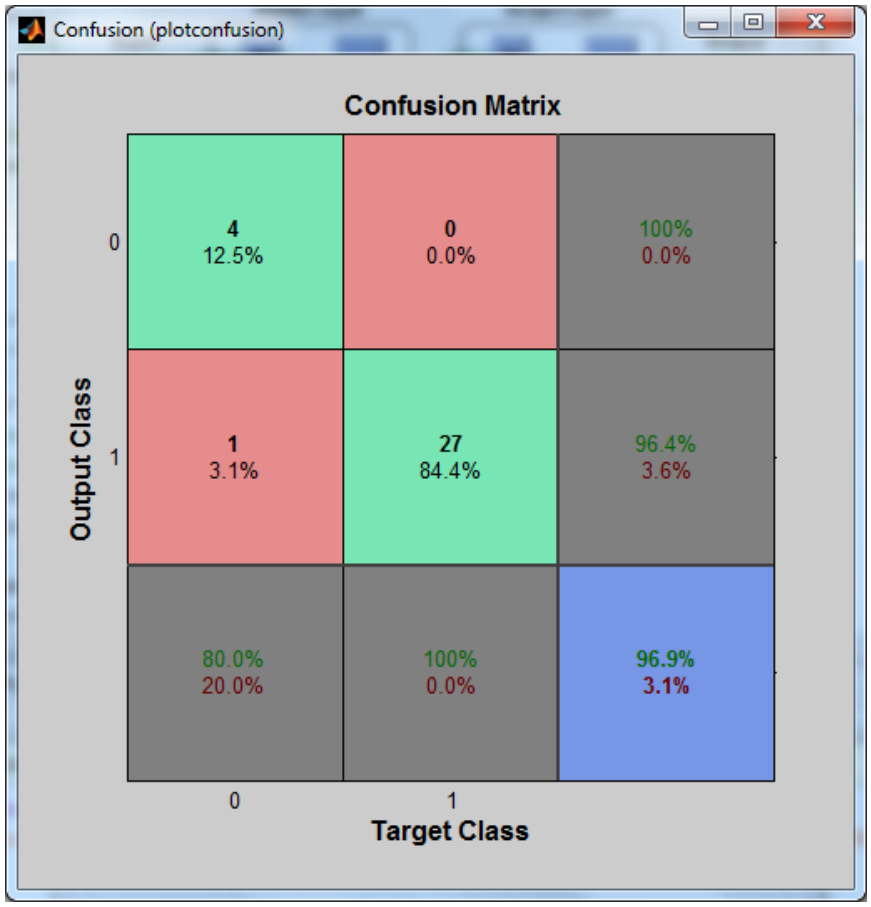

Figure 41. Confusion Matrix of 14 PCs (PCA).

The result for the reconstructed data and 11 features by PCA looks the same. Unless the number of epochs was different: more epochs for the reconstructed data than the features obtained by the PCA (less number of iteration). Also the network execution time of the PCA was too fast compared to the reconstructed data network.

The *ROC curve* is a plot of the true positive rate versus the false positive rate. Y coordinate correspond to sensitivity and x coordinate corresponds to 1-specificity. A perfect test would show points in the upper-left corner, with 100% sensitivity and 100% specificity [Mat14c] (see Figure 42 for reconstructed data).

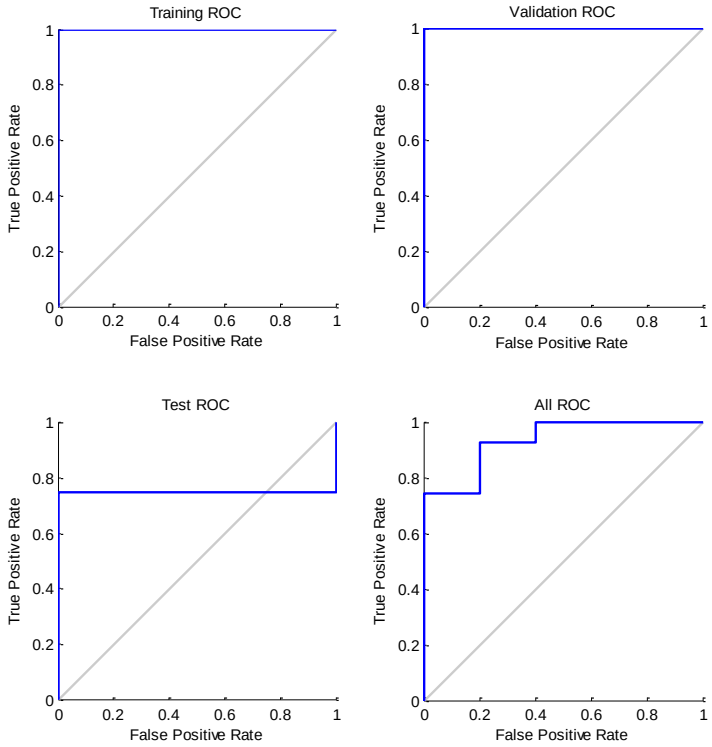

Figure 42. ROC curve plot for reconstructed data.

The total accuracy increases slightly by feature reduction or by using the reconstructed data from the PCA. It reaches 97% of the correct responses. The results of different inputs classification with neural network is summarized in Table 15.

Table 15. Accuracy by NNPR for Inputs.

Total Inputs	**Original Features (14)**	**Dimensionality Reduction (11) Features**	**Reconstructed PCA (14) Features**
Accuracy	94%	97%	97%

The comparison of the accuracy shows that there was not a vast distinction between all classifier algorithms. Since we have two classes to differentiate, it is useful to work with linear classifier (see Table 16).

Table 16. Accuracy of different classifiers.

Classifier Type	**LDA**	**QDA**	**LSVM**	**NLSVM**	**NNPR**
Accuracy	72%	84%	80%	75%	97%

7. Conclusion and Future Work

Automated carcinoma detection and classification in breast cancer may serve as an effective way to classify histological images: Histological images were prepared by HTL followed by preprocessing technique in order to improve the quality of images (noise elimination and contrast enhancement). Fourteen features first and second order are extracted and then treated with the PCA in order to select those with relevant information. The accuracy of the classification depends on the features used and the type of the classifier. The accuracy of the classifiers were very interesting. Based on our research work the QDA, non-linear SVM and NNPR with dimensionality reduction seem to perform best in terms of accuracy. We can conclude that CAD in histological imaging is an effective means of early detection and classification of breast cancer.

Some remarks and conclusions from different part of this Thesis can be summarized as follows:

- Supervised Learning Classification LDA, SVM and NN leads to assign the tissue to the right pattern.
- LDA is a good classifier when selecting the relevant features.
- PCA performs well with LDA classifier (Feature Selection).
- Neural Network can achieve a good results as well as LDA without Dimensionality reduction.
- LDA and NN method demonstrated equal diagnostic power in Breast Cancer detection.
- The execution time is faster for LDA than NN.
- NN is useful to justify the LDA when needed.
- It is not possible to know in advance the ideal network for application.
- Most NNs requires a long training period and many iterations of the same pattern set.
- Most NNs include complex computation.

The aim of the future research work will be based on finding out how to distinguish and classify between DCIS and IDC using some of the pre-extracted feature. The idea is to process only those cancerous images and trying to detect which type of carcinoma they present.

References

[Arb13] R.B.Aswin, Implementation of ANN Classifier using Matlab for Skin Cancer Detection. *International journal of Computer Science and Mobile Computing*, *ICMIC13*. pp. 87-94, 2013.

[BCO13] Breastcancer.org, Types and symptoms of breast cancer, Breastcancerorg_Pathology_Report_Guide_2014. Available online http://www.breastcancer.org/symptoms/diagnosis/invasive. Retrieved 2013.

[BCO14] Breastcancer.org Non-invasive or Invasive breast cancer. Available online http://www.breastcancer.org/symptoms/diagnosis/invasive. Modified 22.1.2014.

[BeM12] A.Belsare, M.Mushrif: Histological image analysis using image processing techniques: An overview. *Signal &Image processing: An international journal (SIPIJ)*, Vol.3, 2012.

[BET08] Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool. "SURF:Speeded Up Robust Features." *Computer Vision and Image Understanding (CVIU)*.Vol. 110, No. 3, pp. 346–359, 2008.

[Bor86] G.Borgefors: Distance Transformations in Digital Images. *Computer Vision, Graphics, and Image Processing*, Vol. 34, No. 3, pp. 344–371, 1986.

[CCV07] Cancer Council Victoria, Ductal carcinoma in situ Available online http://www.cancervic.org.au/preventing-cancer/attendscreening/breasts-health/ductal-carcinoma-in-situ. Retrieved 28.6.2007.

[CDB14] Cancer Diagnosis based on thorough examination, Available online http://www.hus.fi/en/medical-care/medical-imaging-and physiology/Mammografia%20%20englanti/Breast%20biopsy.pdf. Retrieved 15.10.2014.

[CDT14] Cancer Diagnosis based on thorough examination, Available online http://www.hus.fi/en/medical-care/medical-imaging-and-physiology/TT%20englanti/A%20plain%20CT%20examination%20of%20the%20urinary%20tract%20due%20to%20pain%20in%20the%20flank.pdf. Retrieved 21.8.2014.

[CDM14] Cancer Diagnosis based on thorough examination, Available online http://www.hus.fi/en/medical-care/medical-imaging-and-physiology/Mammografia%20%20englanti/Mammography.pdf. Retrieved 15.10.2014.

[CDU12] Cancer Diagnosis based on thorough examination, Available online http://www.hus.fi/en/medical-care/medical-imaging-and-physiology/U%20potilasohjeet%20englanti/Abdominal%20ultrasound%20examination.pdf. Retrieved 7.12.2012.

[CoV95] C.Cortes, V.Vapnik Support-vector networks". *Machine Learning* 20 Vol 3, pp.273, 1995

[Cra05] R.A.Castellino. Computer aided detection (CAD). *Journal PMC BioMedCentral Cancer Imaging*. Vol.5, pp.17-19, 2005.

[DAM06] S.Doyle, S.Agner, A.Madabhushi, M.Feldman, J.Tomaszewski. : A boosting cascade for automated detection of prostate cancer from digitized histology. *13th International Conference on Medical Image Computing and Computer Assisted Intervention* pp. 504–511, 2006.

[DAM07] S.Doyle, S.Agner, A.Madabhushi, M.Feldman, J.Tomaszewski. Automated Grading of Prostate Cancer Using Architectural and Textural Image Features, *IEEE International Symposium on Biomedical Imaging*, pp. 1284-1287, 2007.

[DAM08] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated Grading of Breast Cancer Histopathology Using Spectral Clustering with Textural and Architectural Image Features, *Proc. IEEE International Symposium on Biomedical Imaging IEEE Press,* pp. 496-499. 2008.

[FCR13] Finnish Cancer Registry Mass Screening Registry available online http://wwwdep.iarc.fr/NORDCAN/English/StatsFact.asp?cancer=200 &country=246. NORDCAN, Association of Nordic cancer Registries. Retrieved 16.01.2015.

[Fra36] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* **7** Vol 2, pp. 179–188, 1936.

[GBC09] M.N.Gurcan, L.E.Boucheron, A.Can, A.Madabhushi, M.N.Rajpoot. B.Yener: Histopathological Image Analysis. *IEEE Re-views in Biomedical Engineering*. Vol. 2, pp.147-171, 2009.

[GWE03] R.C.Gonzalez, R.E. Woods, and S.L Eddins.: *Digital Image Processing Using MATLAB*. New Jersey, Prentice Hall, 2003.

[GRN13] N.Gupta, A.Rawal, V.L. Narasimhan, S.Shiwani. Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data. *Journal of Computer Engineering (IOSR-JCE)*. Vol.11, pp. 70-73, 2013.

[HaS97] M. A. Hall, L. A. Smith, Feature Subset Selection: A correlation based filter approach. *International Conference on Neural Information Systems*, pp.855-858, 1997.

[HLA10] L.He, R.Long, S.Antani, G.R.Thoma. Local and Global Gaussian mixture models for hematoxylin and eosin stained histology image segmentation, *IEEE International Symposium on Biomedical* Imaging. pp. 223-228, 2008.

[HLA12] L.He, R.Long, S.Antani, G.R.Thoma. Histology image analysis for carcinoma detection and grading, *Journal computer Methods and programs in Biomedicine*, vol. 107, pp. 538-556, 2012.

[Hot97] H. Hotteling 1933. Analysis of a Complex of Statistical Variables Into Principal Components, *Journal of Educational Psychology*, Vol 24, pp. 417-441 and 498-520, 1997.

[HSD73] R. M. Haralick, K. Shanmugam, and I. Dinstein, Textural Features of Image Classification, *IEEE Transactions on Systems, Man and Cybernetics*, vol.6, pp. 610-621, 1973.

[IkT05] L. Ikonen., P. Toivanen.: Shortest Routes Between Sets on Gray-Level Surfaces. *Pattern Recognition and Image Analysis*. Vol. 15, pp. 195–198, 2005.

[KFH00] A.Krzyzak, T.Fevens, M.Habibzadeh, L Jelen.: Application of Pattern Recognition Techniques for the Analysis of Histopathological Images.

[Ksy90] S. Y. Kung, Kernel Approaches to Unsupervised and Supervised Machine Learning. *Advances in Multimedia Information Processing – PCM Lecture Notes in Computer Science*. Volume 5879, pp 1-32 2009.

[LWe04] W.Lin, Weng, Probability estimates for multi-class classification by pairwise coupling. *JMLR*. Vol 5, pp. 975-1005, 2004.

[MaC02] A.Marie, C.Jansen. E-cadherin and loss of heterozygosity at chromosome 16 in breast carcinogenesis: different genetic pathways in ductal and lobular breast cancer? .*US National Library of Medicine National Institutes of Health.* Vol. 4, pp.5-8, 2002.

[Mar70] K. V. Mardia, Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57 Vol.3, pp. 519–530, 1970.

[Mat14a] MathWorks Inc.: *MATLAB and Image processing Toolbox Release 2014b*. MathWorks Inc., 2014. http://www.mathworks.se/products/image/whatsnew.html (8.10.2014).

[Mat14b] MathWorks Inc.: *MATLAB and Computer Vision Toolbox Release 2014a*. MathWorks Inc., 2014. http://www.mathworks.se/products/computer-vision/ (8.10.2014).

[Mat14c] MathWorks Inc.: *MATLAB and Statistics Toolbox Release 2014b*. MathWorks Inc., 2014. http://se.mathworks.com/products/statistics/ (8.10.2014).

[McA04] McAndrew A.: *An Introduction to Digital Image Processing with MATLAB.* Notes for SCM2511 School of computer Science and Mathematics, Victoria University of Technology. Available online http://visl.technion.ac.il/labs/anat/An%20Introduction%20To%20Digital%20Image%20Processing%20With%20Matlab.pdf (8.10.2014).

[NCI14] National Cancer Institute at the National Institutes of Health. Understanding Breast Changes: A Health Guide for Women.

http://www.cancer.gov/cancertopics/screening/understanding-breast-changes/understanding-breast-changes.pdf. Revised 2.2014.

[NiS11] R.Nithya, B.Santhi, Comparative Study on Feature Extraction Method for breast cancer classification. *Journal of Theoretical and Applied Information Technology JATIT*, Vol 33, No.2, 2011

[Ona03] N. A. Obuchowski. Receiver operating characteristic curves and their use in radiology. *US National Library of Medicine National Institutes of Health PubMed*. 229(1), pp. 3–8, 2003.

[RoP06] M.H.Ross, W.Pawlina, *Histology: A Text and Atlas*. Baltimore Lippincott Wiliams &Wilkins, 2006.

[RoP86] A. Rosenfeld, J. Pfaltz. *Distance Functions in Digital Pictures*, Pattern Recognition, Vol. 1, pp.33 – 61, 1986.

[Rse14] S. Raschka, Predictive modeling, supervised machine learning, and pattern classification available online http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html. 24 August, 2014.

[Scd15] C.D.Souza. Handwriting Recognition Revisited: Kernel Support Vector Machines. Available online at: http://www.codeproject.com/Articles/106583/Handwriting-Recognition-Revisited-Kernel-Support-V. 31 Dec, 1999 and Update: 20 Jan, 2015.

[SiS07] H.Shimazaki, S.Shinomoto, A method for selecting the bin size of a time histogram *Neural Computation* Vol 6, pp. 1503–1527, 2007.

[SLF14] Susan Love Research Foundation Available online http://www.dslrf.org/breastcancer/content.asp?CATID=4&L2=1&L3=6&L4=0&PID=&sid=132&cid=440. Reviewed 17.11.2014.

[ToI05] P. Toivanen, L. Ikonen.: Shortest Routes on Varying Height Surfaces Using Gray-Level Distance Transforms. *Journal of Image and Vision Computing*, Vol. 23, pp. 134–141, 2005.

[TPK10] S.Theodoridis, A.Pikrakis, K.Koutroumba, D.Cavouras. *Introduction to Pattern Recognition: A Matlab Approach.* Academic Press. U.S., 3.2010.

[WeJ73] S.Wellings, H.Jensen. On the origin and progression of ductal carcinoma in the human breast. *J Natl Cancer Inst*, Vol 5, pp.1111-1118, 1973.

[VMC14] MyVMC Virtual Medical Center. Referenced Health & Medical Information Endorsed by Doctors Available online, http://www.myvmc.com/anatomy/what-is-cancer/.Modified 25.8.2014.

[WHO14] World Health Organization Media center Cancer Key facts N 297 available online http://www.who.int/mediacentre/factsheets/fs297/en/ Updated 2014.