T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

LUNG MASS CLASSIFICATION USING WAVELETS AND SUPPORT VECTOR MACHINES

Master Thesis

BAŞAK SARIKAYA

İSTANBUL, 2009

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

Institute of Sciences
Computer Engineering Graduate Program

LUNG MASS CLASSIFICATION USING WAVELETS AND SUPPORT VECTOR MACHINES

Master Thesis

Başak SARIKAYA

SUPERVISOR: ASSOC. PROF. DR. ADEM KARAHOCA

İSTANBUL, 2009

T.C

BAHÇEŞEHİR ÜNİVERSİTESİ

Institute of Sciences

Computer Engineering Graduate Program

Name of the thesis: **LUNG MASS CLASSIFICATION USING WAVELETS AND SUPPORT VECTOR MACHINES**

Name/Last Name of the Student: Başak SARIKAYA
Date of Thesis Defense: 05 June 2009

The thesis has been approved by the Institute of Sciences.

Director: Prof. Dr. Bülent ÖZGÜLER

Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

Head of Department:
Prof. Dr. Bülent ÖZGÜLER

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Co-Supervisor Supervisor

Examining Committee Members
Prof. Dr. Nizamettin AYDIN
Assoc. Prof. Dr. Adem KARAHOCA
Asst. Prof. Dr. Yalçın ÇEKİÇ

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Assoc. Prof. Dr. Adem Karahoca who was always very helpful about this thesis and my works, without his tolerance and support completion of this study wouldn't be possible. I can never repay him for the opportunity he gave me.

I would like to thank my cousin who has founded computerized lung tomographic data from different hospitals and research centers for me.

I would like to thank my co-workers who believed in me about my graduation and thesis.

Finally, I would like to express my gratitude to my family (Fuat, Gülsun and Mert Sarıkaya), my friends, especially Ezgi Erten and Nihan Akkuş, their love and support were essential for the completion of my thesis. It is to them I would like to dedicate this thesis.

ABSTRACT

LUNG MASS CLASSIFICATION USING WAVELETS AND SUPPORT VECTOR MACHINES

Başak SARIKAYA

Institute of Sciences, Computer Engineering Graduate Program

Supervisor: Assoc. Prof. Dr. Adem Karahoca

June 2009, 34 pages

This study deals with observation of an approach for classification the lung cancer masses as cancer or not. In this thesis, it is implemented a compound of Support Vector Machine (SVM) and wavelet based image decomposition. Decision making was performed with two partitions, feature determination by computing the wavelet coefficients and classification using the classifier trained on the feature determination. Support Vector Machine (SVM) is a learning machine which relies on statistical learning theory was trained in order to supervised learning to classify masses. The study implies 126 computerized lung tomography images. The masses were segmented by breast expert doctors manually at first step to the classification system. Test results demonstrate accuracy on lung cancer indicated over 76.74% classification accuracy by using the SVM with Radial Basis Function Kernel. Also confusion matrix, accuracy, sensitivity and specificity analysis with different kernel types were employed to demonstrate the classification performance of SVM.

Keywords: Lung Cancer Mass Detection, Lung Cancer Mass Classification, Support Vector Machine, Discrete Wavelet Transform, Image Processing.

ÖZET

DALGACIK DÖNÜŞÜMÜ VE DESTEK VEKTÖRÜ MAKİNELERİ KULLANILARAK AKCİĞER KÜTLESİNİN SINIFLANDIRILMASI

Başak SARIKAYA

Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Adem Karahoca

Haziran 2009, 34 Sayfa

Bu çalışmada, bilgisayarlı tomografi ile çekilmiş akciğer görüntülerindeki akciğer kütlelerinin kanserli olup olmadığını sınıflandırmak için kullanılabilecek yaklaşımlar incelenmektedir. Bu tez, Destek Vektörü Makineleri ve Dalgacık Dönüşümü tabanlı alt bant görüntü dönüşümü kombinasyonlarına dayanmaktadır. Karar verme mekanizması, dalgacık dönüşümü katsayıları ile desenden özellik vektörü çıkarım hesaplanması ve desenden çıkarılan vektörün üzerinde kullanılan eğitim sınıflandırıcısı vasıtasıyla sınıflandırılması olmak üzere iki bölümde gerçekleştirilmiştir. Destek Vektörü Makineleri, kütle sınıflandırması için öğreticili öğrenme eğitimi doğrultusunda istatistiksel öğrenme teorisine dayanan öğrenme makineleridir. Çalışmada 126 bilgisayarlı akciğer tomografi görüntüsü kullanılmıştır. Görüntüdeki kütleler sınıflandırma için ilk etapta başlangıç olarak göğüs hastalıkları uzman doktorları tarafından tek tek gözle ayrıştırılmıştır, daha sonra tezde uygulanan yöntemler ile sınıflandırma işlemi otomatize edilmiştir. Bilgisayarlı akciğer tomografileri üzerinde yapılan testlerden elde edilen sonuçlarda, dalgacık dönüşümü ile filtreleme yapıldıktan sonra, destek vektörü makineleri ve radyal tabanlı fonksiyon çekirdeği kullanımı, % 76.74 sınıflandırma doğruluğuna erişilmiştir. Destek Vektör Makineleri'nin sınıflandırmadaki performansını göstermek için, sonuç düzensizlik matrisi, doğruluk, hassasiyet ve kesinlik analizi değerleri, farklı çekirdek tipleri kullanılarak gösterilmiştir.

Anahtar Kelimeler: Akciğer Kütlesinin Belirlenmesi, Akciğer Kütlesinin Sınıflandırılması, Destek Vektörü Makineleri, Kısa Zamanlı Dalgacık Dönüşümü, Görüntü İşleme.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Artificial Neural Network	:	ANN
Support Vector Machine	:	SVM
Sequential Minimal Optimization	:	SMO
Computer Aided Detection	:	CAD
Computerized Tomography	:	CT
Receiver Operating Characteristic	:	ROC
Discrete Wavelet Transform	:	DWT
Continuous Wavelet Transform	:	CWT
Inverse Discrete Wavelet Transform	:	IDWT
Cellular Neural Network	:	CNN
Probabilistic Neural Network	:	PNN
Region Of Interest	:	ROI
Gaussian Radial Basis Function	:	RBF
High-High	:	HH
High-Low	:	HL
Low-High	:	LH
Low-Low	:	LL
Low-High-Low-High	:	LHLH
Area Under the Curve	:	AUC
Original Image	:	I
De-Noised Image	:	DI
Root Mean Squared Error	:	RMSE

1. INTRODUCTION

Lung cancer is the most prevalent cancer among people, which is in the second order of the world's cancer statistics after skin cancer. And it is in the first order of Turkey's cancer statistics is published by Turkish Republic Ministry of Health in National Cancer Week in Turkey. Lung cancer is leading cause of death from cancer among people of ages between 45 and 70 in Turkey as reported in Turkish Republic Ministry of Health Reports 2009 (www.saglik.gov.tr, 2009).

1.1 PROBLEM DEFINITION

In the Past 20 years the incidence and death rates of lung cancer have been taking the lead in all malignancies, and the incidence rate is as high as 29,51 per 100 thousand people in our country. Moreover, the incidence and death rates are still increasing continuously. The first reason that people caught lung cancer is smoking 90% men and 70% women directly and because of the other reasons second-hand smoking radon, asbestos and other toxic products come in second order (Cancer Care Ontario, 1964-2002).

The two most prevalent forms of lung cancer are non-small cell lung cancer and small cell lung cancer. Non-small lung cancer is more wide spread than small cell lung cancer and accounts for 85 to 87 percent of all lung cancers. Small cell is very aggressive and spreads quickly. By the time that most people are diagnosed, the cancer has metastasized to other parts of the body.

The survival time is very short once cancer is diagnosed as being in advanced stage and surgically unresectable and a deadly disease in the world. Early detection of this disease is very important to prevent this disease. Therefore, a good model of prediction of disease outcome is important for a treatment plan.

Every year 20.000 new lung cancer diagnosis occurs in Turkey from the public speech of Özdemir (2009) in 2009 National Cancer Week in Turkey. And Özdemir (2009) indicates this illness could be caught in earlier stages 15% in our country and 30% in USA.

For this early detection that reduce the death rate or increase the death ages of the most trustable method for the determination of early lung cancer of all determination methods currently available.

But, there are many difficulties in detecting early pathological changes and evaluating oncology parameters in treating because of the difficulty that to date the pathogeny of lung cancer is not clear yet. By this side, various methods and criteria of evaluating pathological diagnosis are being improved day by day. In order to increase the speed of detecting lung nodules, it is using artificial neural networks (ANN) methods to determine the target position in the observed image and to select an adequate template image from several reference patterns.

Detecting nodules is such a complicated task. Nodules show up as relatively low-contrast white circular objects within the lung fields. By this side, CT (Computerized Tomography) provides extra peculiarities not available with standard film-screen tomographic images in past years such as computer aided diagnosis, contrast enhancement and digital archiving. It could be missed cancers which are visible on computerized tomographies by radiologists in retrospect, but studies before have demonstrated (on computerized tomographies) that computer aided detection (CAD) and determination can meaningfully correct radiologists' truth ratio in intuition set microcalcifications. CAD systems are planned to procure a second idea, to help not to put any other radiologist. Cancer lung masses often filter the besieging tissue as they have been widen. They separate cancer masses from non cancer masses in shape and density (Adhami and Bruce, 1999, pp. 1170-1177).

In this thesis, classification of uncertain masses that includes cancer or non cancer masses is comprehended to treat in the newest image processing and artificial neural networks techniques. At first, it has derived genuine peculiarities from computerized tomographic images using discrete wavelet transform (Mallat, 1980, pp. 674-693). Wavelet Transform is a tool for time-space frequency analysis against Fourier Transform (Bracewell, 1999, pp. 15-35), which procures only frequency analysis of signals. Wavelet Transform procures time-frequency analysis, which is especially an advantage for pattern recognition. With wavelet transform it is able to divide the signal as much as it can, at first it has taken the signal through with low pass and then high pass filters. With this sequence, it could be able to separate two signals with sub band divisions (This technique is using especially compression applications, for example, jpeg2000 and mp4 formats uses wavelet technique and it is used often in digital signal processing area).

The prediction of the cancer region was made by comparing real data obtained from follow-up periods with data generated by Wavelet Approach. And beside this, Wavelet image processing technique provides good prediction results when it is used Support Vector Machine for classification in the model.

So, after feature extraction, it has been used Support Vector Machine (SVM) machine learning algorithm to classify of images with one of two categories, that cancer or not cancer. As showed in Figure 1.1, the segmented computerized tomographic images are wavelet separated into multi-level low pass and high pass subbands, which will be followed by as an input of SVMs for training and testing goals. SVM decreases constructed risk in learning level (Vapnik, 1998, pp. 28-46). In here the purpose is to minimize generalization error, not to minimize learning error towardly.

In conclusion, SVM is capable of processing pleasantly when developed to data which external images from the training set. Recently, SVM learning has been applied lots of applications in the world where it has been thought to present the best performance for computation of methods.

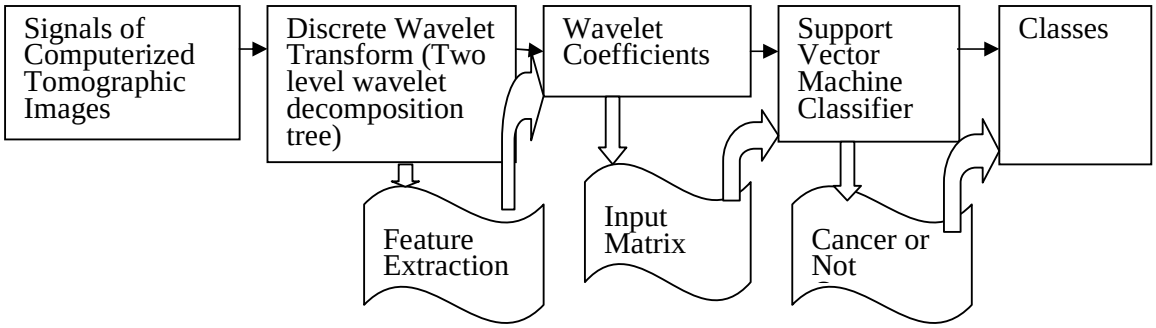

Figure 1.1: Structure of the SVM Classifier

1.2 BACKGROUND

The researches related to this thesis are followed in this background section. Early prostate cancer diagnosis by using artificial neural networks and support vector machines has been searched which is designing a classifier based expert system for early diagnosis of the organ in constraint phase to reach informed decision making without biopsy by using some selected. The other purpose is to investigate a relationship between BMI (body mass index), smoking factor, and prostate cancer. The data used in this study were collected from 300 men (100: prostate adenocarcinoma, 200: chronic prostatism or benign prostatic hyperplasia). Weight,

height, BMI, PSA (prostate specific antigen), Free PSA, age, prostate volume, density, smoking, systolic, diastolic, pulse, and Gleason score features were used and independent sample t-test was applied for feature selection. In order to classify related data, it is used following classifiers; scaled conjugate gradient (SCG), Broyden–Fletcher–Goldfarb–Shanno (BFGS), and Levenberg–Marquardt (LM) training algorithms of artificial neural networks (ANN) and linear, polynomial, and radial based kernel functions of support vector machine (SVM). It was determined that smoking is a factor increases the prostate cancer risk whereas BMI is not affected the prostate cancer. Since PSA, volume, density, and smoking features were to be statistically significant, they were chosen for classification. The proposed system was designed with polynomial based kernel function, which had the best performance (accuracy: 79%). In Turkish Family Health System, family physician to whom patients are applied firstly, would contribute to extract the risk map of illness and direct patients to correct treatments by using expert system such proposed (Çınar et al., 2009, pp. 6357-6361).

Support vector machines combined with feature selection for breast cancer diagnosis is investigated and which insists that Breast cancer is the second largest cause of cancer deaths among women. At the same time, it is also among the most curable cancer types if it can be diagnosed early. Research efforts have reported with increasing confirmation that the support vector machines (SVM) have greater accurate diagnosis ability. In this paper, breast cancer diagnosis based on a SVM-based method combined with feature selection has been proposed. Experiments have been conducted on different training-test partitions of the Wisconsin breast cancer dataset (WBCD), which is commonly used among researchers who use machine learning methods for breast cancer diagnosis. The performance of the method is evaluated using classification accuracy, sensitivity, specificity, positive and negative predictive values, receiver operating characteristic (ROC) curves and confusion matrix. The results show that the highest classification accuracy (99.51%) is obtained for the SVM model that contains five features, and this is very promising compared to the previously reported results (Akay, 2009, pp. 3240-3247).

Mining lung cancer patient data to assess healthcare resource utilization is inquired which objects in this study is to assess the utilization of healthcare resources by lung cancer patients related to their demographic characteristics, socioeconomic markers, ethnic backgrounds, medical histories, and access to healthcare resources in order to guide medical decision making and public policy. The study compares alternative data mining techniques in combination with traditional regression methods and uses propensity scoring to differentiate the predictive power of various models. The study demonstrates that data mining methods can

be applied to large, complex, public-use Medicare insurance claims files to reveal insights such as geographic variation in healthcare delivery practice patterns for lung cancer. The results indicate that decision trees and artificial neural networks, particularly when used in combination, can produce better predictive and descriptive models than regression alone to guide healthcare decisions (Dy, Phillips-Wren, and Sharkey, 2008, pp. 1611-1619).

Cancer informatics by prototype networks in mass spectrometry has become a standard technique to analyze clinical samples in cancer research. The obtained spectrometric measurements reveal a lot of information of the clinical sample at the peptide and protein level. The spectra are hi Summary in and methods are wavelet-based techniques for the efficient processing and encoding of mass spectrometric measurements from clinical samples are presented. A prototype-based classifier is extended by a functional metric and combined with the concept of conformal prediction to classify the clinical proteomic spectra and to evaluate the results. And as a result, Clinical proteomic data of a colorectal cancer and a lung cancer study are used to test the performance of the proposed algorithm. The prototype classifiers are evaluated with respect to prediction accuracy and the confidence of the classification decisions. The adapted metric parameters are analyzed and interpreted to find potential biomarker candidates. In conclusion, it is shown that the proposed algorithm can be used to analyze functional data as obtained from clinical mass spectrometry, to find discriminating mass positions and to judge the confidence of the obtained classifications, providing robust and interpretable classification models (Schleif et al., 2008).

Evolutionary ensemble of diverse artificial neural networks using speciation claims that recently many researchers have designed neural network architectures with evolutionary algorithms but most of them have used only the fittest solution of the last generation. To better exploit information, an ensemble of individuals is a more promising choice because information that is derived from combining a set of classifiers might produce higher accuracy than merely using the information from the best classifier among them. One of the major factors for optimum accuracy is the diversity of the classifier set. In this paper, it is presented a method of generating diverse evolutionary neural networks through fitness sharing and then combining these networks by the behavior knowledge space method. Fitness sharing that shares resources if the distance between the individuals is smaller than the sharing radius is a representative speciation method, which produces diverse results than standard evolutionary algorithms that converge to only one solution. Especially, the proposed method calculates the distance between the individuals using average output, Pearson correlation and modified Kullback–Leibler entropy to enhance fitness sharing performance. In experiments with

Australian credit card assessment, breast cancer, and diabetes in the UCI database, the proposed method performed better than not only the non-speciation method but also better than previously published methods (Kim and Cho, 2009, pp. 1604-1618).

Collection of Cancer Stage Data by Classifying Free-text Medical Reports implicates that Cancer staging provides a basis for planning clinical management, but also allows for meaningful analysis of cancer outcomes and evaluation of cancer care services. Despite this, stage data in cancer registries is often incomplete, inaccurate, or simply not collected. This article describes a prototype software system (Cancer Stage Interpretation System, CSIS) that automatically extracts cancer staging information from medical reports. The system uses text classification techniques to train support vector machines (SVMs) to extract elements of stage listed in cancer staging guidelines. When processing new reports, CSIS identifies sentences relevant to the staging decision, and subsequently assigns the most likely stage. The system was developed using a database of staging data and pathology reports for 710 lung cancer patients, then validated in an independent set of 179 patients against pathologic stage assigned by two independent pathologists. CSIS achieved overall accuracy of 74% for tumor (T) staging and 87% for node (N) staging, and errors were observed to mirror disagreements between human experts (Bowman et al., 2007).

Artificial neural networks and decision tree model analysis of liver cancer insists that Hepatocellular carcinoma (HCC) is a heterogeneous cancer and usually diagnosed at late advanced tumor stages of high lethality. The present study attempted to obtain a proteome-wide analysis of HCC in comparison with adjacent non-tumor liver tissues, in order to facilitate biomarkers' discovery and to investigate the mechanisms of HCC development. A cohort of 66 Chinese patients with HCC was included for proteomic profiling study by two-dimensional gel electrophoresis (2-DE) analysis. Artificial neural network (ANN) and decision tree (CART) data-mining methods were employed to analyze the profiling data and to delineate significant patterns and trends for discriminating HCC from non-malignant liver tissues. Protein markers were identified by tandem MS/MS. A total of 132 proteome datasets were generated by 2-DE expression profiling analysis, and each with 230 consolidated protein expression intensities. Both the data-mining algorithms successfully distinguished the HCC phenotype from other non-malignant liver samples. The detection sensitivity and specificity of ANN were 96.97% and 87.88%, while those of CART were 81.82% and 78.79%, respectively. The three biological classifiers in the CART model were identified as cytochrome b5, heat shock 70 kDa protein 8 isoform 2, and cathepsin B. The 2-DE-based proteomic profiling approach combined with the ANN or CART algorithm yielded

satisfactory performance on identifying HCC and revealed potential candidate cancer biomarkers (Luk et al., 2007, pp. 68-73).

Cancer gene search with data-mining and genetic algorithms is searched which indicates that Cancer leads to approximately 25% of all mortalities, making it the second leading cause of death in the United States. Early and accurate detection of cancer is critical to the well being of patients. Analysis of gene expression data leads to cancer identification and classification, which will facilitate proper treatment selection and drug development. Gene expression data sets for ovarian, prostate, and lung cancer were analyzed in this research. An integrated gene-search algorithm for genetic expression data analysis was proposed. This integrated algorithm involves a genetic algorithm and correlation-based heuristics for data preprocessing (on partitioned data sets) and data mining (decision tree and support vector machines algorithms) for making predictions. Knowledge derived by the proposed algorithm has high classification accuracy with the ability to identify the most significant genes. Bagging and stacking algorithms were applied to further enhance the classification accuracy. The results were compared with that reported in the literature. Mapping of genotype information to the phenotype parameters will ultimately reduce the cost and complexity of cancer detection and classification (Shah and Andrew, 2007, pp. 251-261).

Tumor tissue identification based on gene expression data using DWT feature extraction and PNN classifier which is proposed the joint use of discrete wavelet transform (DWT)-based feature extraction and probabilistic neural network (PNN) classifier to classify tissues using gene expression data. In the feature extraction module, gene expression data are firstly transformed into time-scale domain by DWT and then the reconstructed signals by using wavelet transform are reduced to a lower dimensional feature space. In the module of tissue classification, the outputs of the extractor are fed into a PNN classifier, and the class labels are given finally. Some test and comparison experiments have been made to evaluate the performance of the proposed classification scheme, using the features extracted with as well as without wavelet transform processing procedure. Correct rates of 92% and 98.7% in tumor vs. normal classification have been obtained using the proposed scheme on two well-known data sets: a colon cancer data set and a human lung carcinomas data set (Sun, Dong and Xu, 2006, pp. 387-402).

Prediction of colon cancer using an evolutionary neural network insists that Colon cancer is second only to lung cancer as a cause of cancer-related mortality in Western countries. Colon cancer is a genetic disease, propagated by the acquisition of somatic alterations that influence gene expression. DNA microarray technology provides a format for the simultaneous

measurement of the expression level of thousands of genes in a single hybridization assay. The most exciting result of microarray technology has been the demonstration that patterns of gene expression can distinguish between tumors of different anatomical origin. Standard statistical methodologies in classification and prediction do not work well or even at all when N (a number of samples) <p (genes). Modification of conventional statistical methodologies or devise of new methodologies is needed for the analysis of colon cancer. Recently, designing artificial neural networks by evolutionary algorithms has emerged as a preferred alternative to the common practice of selecting the apparent best network. In this paper, it is proposed an evolutionary neural network that classifies gene expression profiles into normal or colon cancer cell. Experimental results on colon microarray data show that the proposed method is superior to other classifiers (Kim and Cho, 2004, pp. 361-379).

Clinical decision support systems for intensive care units using artificial neural networks subject which provides an overview of applications of artificial neural networks (ANNs) to various medical problems, with a particular focus on the intensive care unit environment (ICU). Several technical approaches were tested to see whether they improve the ANN performance in estimating medical outcomes and resource utilization in adult ICUs. These experiments include: 1- use of the weight-elimination cost function; 2- use of 'high' and 'low' nodes for input variables; 3- verifying the effect of the total number of input variables on the results; 4- testing the impact of the value of the constant predictor on the performance of the ANNs. The developments presented intend to help medical and nursing personnel to assess patient status, assist in making a diagnosis, and facilitate the selection of a course of therapy (Frize et al., 2001, pp. 217-225).

Artificial neural networks for early detection and diagnosis of cancer asks why use neural networks, the reasons commonly cited in the literature for using artificial neural networks for any problem are many and varied. They learn from experience. They work where other algorithms fail. They generalize from the training examples to perform well on independent test data. They reduce the number of false alarms without increasing significantly the number of false negatives. They are fast and are easier to use than conventional statistical techniques, especially when multiple prognostic factors are needed for a given problem. These factors have been overly promoted for the neural techniques. The common theme of this paper is that artificial neural networks have proven to be an interesting and useful alternate processing strategy. Artificial neural techniques, however, are not magical solutions with mystical abilities that work without good engineering. With good understanding of their capabilities and limitations they can be applied productively to problems in early detection and diagnosis

of cancer. The specific cancer applications which will be used to demonstrate current work in artificial neural networks for cancer detection and diagnosis are breast cancer, liver cancer and lung cancer (Rogers, Ruck and Kabrisky, 1994, pp. 79-83).

2. MATERIALS AND METHODS

2.1 COMPUTERIZED LUNG TOMOGRAPHY DATA

In this thesis, there is a set of 126 computerized lung tomographies with 128x128 pixels were taken. These images have been taken from Istanbul Cerrahpasa University Hospital, Yeditepe University Hospital, TDV 29 Mayis Hospital, Esenyurt Government Hospital and Sisli Etfal Education and Research Hospital, and which of them use GE Medical Systems (Centric DICOM Viewer) with technical properties 120 kV, 800 mA, 0.40s/HE+ 39.4 mm/rot Rot, 0.6 mm 0.984:1/0.6 SP, TiH 0.0, SIENET Sky (DICOM CD Viewer) with technical properties 120 kV, changing values with 226 to 255 mA, ST:8, CS:1.50 TI:500, MERGE MED (eFilm Lite) with technical properties 120 kV, 50 mA, ST:1.2 mm, and one more system with 2.5 mm ST value and changing SP value which changes one in an every 2.5 mm.

At all 85 patients have cancer mass and 41 patients have not cancer mass in their lungs. As shown in Figure 2.1, the segments of images exists the grayscale image originally and cancer masses show systematical and well defined surroundings on images in generally, when non cancer masses commonly filtered neighbor tissues showing non-systematically and angled edges.

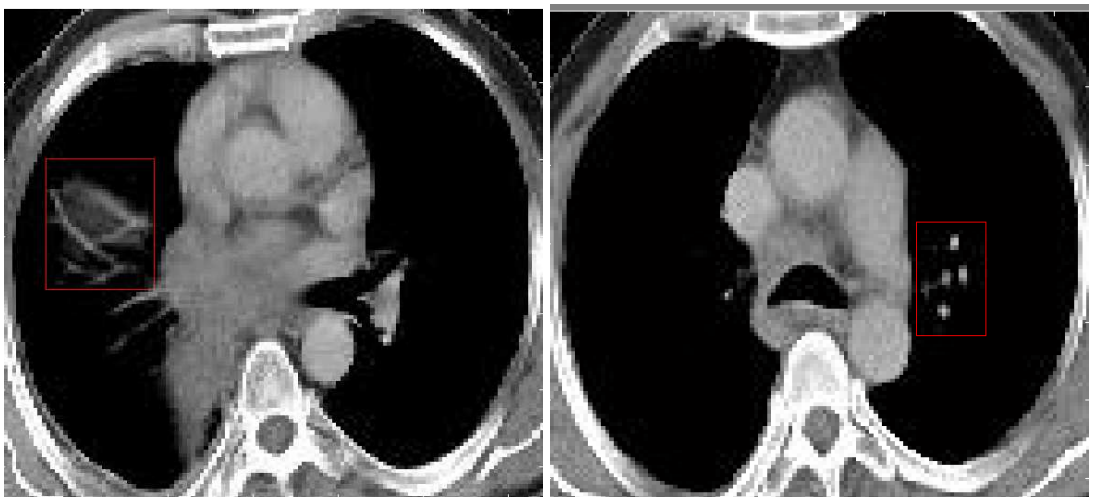

Figure 2.1: A sample Computerized Lung Tomography Image from the Data Set. a) Non Cancer Lung Mass b) Cancer Lung Mass

2.2 FEATURE EXTRACTION

Feature extraction is the specification of a feature matrix from a pattern. For pattern processing problems to be traceable are abridged representation of patterns, ideally including only main information.

The variation of patterns to features that are summarized explanation of patterns, which is desired, inherits only fundamental information, considers necessary to be traceable for pattern processing problems. The variation is materialized by Discrete Wavelet Transform for this thesis.

2.3 WAVELET FILTER

The Wavelet Transform which is currently has obtained seniority concerning in various applications similar compression, noising and denoising of data or images (Brislawn, 1995, pp. 1278-1283). The Wavelet Transform interests with the frequency and time components of the signal showing synchronously different from the Fourier Trasform which gives scientific information just about the frequency components of a signal, not determination of the time at these frequency's formation (Mathsoft Wavelet resources 2003).

Wavelet is being used to decompose an image into four subbands which is low-high (LH), low-low (LL), high-high (HH), and high-low (HL) components. Moreover, the LH subband is decomposed into another four subbands, and the Low-High-Low-High (LHLH) from this second decomposition subband is decomposed once more and continuous like this as shown in Figure 2.2.

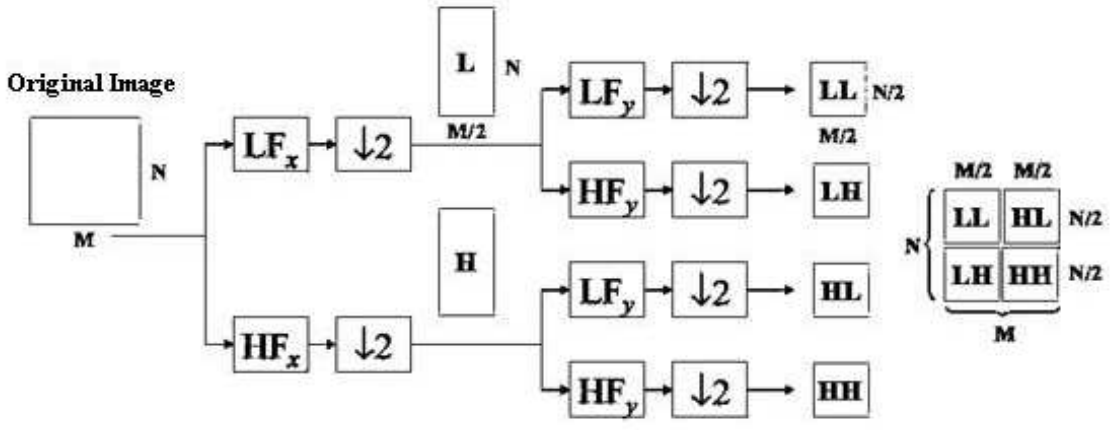

Figure 2.2: Image Decomposition with Wavelet Transform

There is a mandatory to disclosure lung cancer masses because of their natural properties. In this study it used two-level decomposition (Figure 2.3). There are various of wavelet transform types which differs from application to application. For continuous signals the continuous wavelet transform (CWT) can be used, which is the situation for both time and scale are continuous.

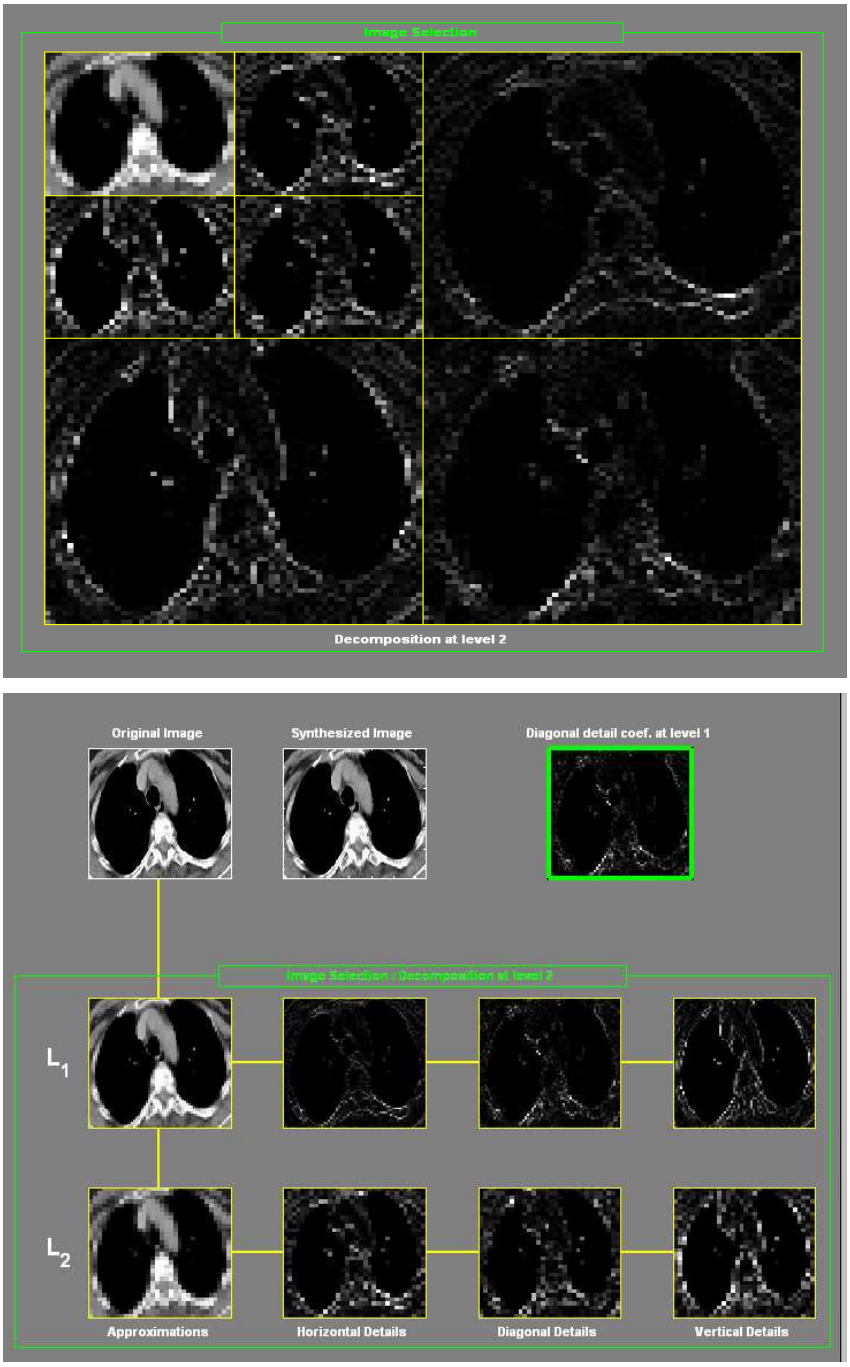

Figure 2.3: Two-Level DWT Decomposition a) LF_x & HF_x images b) Two-Level Decompositions

The discrete wavelet transform (DWT) can be used for discrete signals. In this study, discrete wavelet transform (DWT) was used because of DWT s discrimination. The DWT deals with a discrete set of the wavelet measurements and managing by some definite directions which is a validation of the wavelet transform generally.
CWT of a signal, function f is:

$$f(a,b) = \int f(x)\, \psi_{a,b}(x)\, dx \quad \ldots\ldots (2.1)$$

$$\psi_{a,b}(x) = 1/\sqrt{a}\ \psi(x-b / a) \quad \ldots\ldots (2.2)$$

The fundamental wavelet transform is $\psi(x)$. In wavelet transform, the basis functions are derived from translation and scaling of a unique function, called mother wavelet. The basis of wavelet function is presented by scaling and shifting mother wavelet function. w(x) signal is discremination into a family of composition wavelets as given in Formula 2.3.

$$w(x) = \sum_a \sum_b < w(x), \psi_{a,b}(x) > \psi_{a,b}(x) \quad \ldots\ldots (2.3)$$

$$w[b] = \sum_{i=1 \text{ to } I} \sum_{k\epsilon Z} c_{i,k}\ g[b-2^i k] + \sum_{k\epsilon Z} d_{I,k}\ h_I[b-2^I k] \quad \ldots\ldots (2.4)$$

w[b] is a discrete time signal, $c_{i,k}$ where i=1,…,I are wavelet coefficients and $d_{i,k}$ where i=1,…,I are scaling coefficients.

$$c_{i,k} = \sum_b w[b]\ g_i * [b-2^i k]\ ,\ d_{i,k} = \sum_b w[b]\ h_I * [b-2^I k] \quad \ldots\ldots (2.5)$$

2.4 TWO DIMENSION OF DISCRETE WAVELET TRANSFORM

In this study, a feature matrix is derived from computerized lung tomographies supported by multi-level wavelet decomposition. These matrices are used to train a SVM for classification of computerized lung tomographies. The DWT is applied to every dimension one by one (Chaplot and Patnaik, 2006, pp. 86-92). This returns a multi resolution decomposition of the signal into four subbands called the approximation which is low frequency component and details which represents high frequency component. The approach α demonstrates a low

resolution of the original image. The detail coefficients are *h* for horizontal, *v* for vertical and *d* for diagonal. An image Y is being decomposed into a first level approach component Y_{α}^{1} and detailed components Y_h^1, Y_v^1, and Y_d^1 (Gonzales and Woods, 2002) is shown in Figure 2.4. The approach component Y_α includes low frequency components of the image when the detailed components Y_h, Y_v, Y_d comprise high frequency components.

So,

$$Y = Y_\alpha + \{Y_h^1 + Y_v^1 + Y_d^1\} \quad \text{.....} \quad (2.6)$$

When it has applied DWT to Y_α^1, the second level approach and detailed components are taken. When the manipulation goes on after and after up to N levels, the image Y can be mentioned by the Nth approach component Y_α^N and every detailed components as shown like Formula 2.7.

$$Y = Y_\alpha^N + \sum_{i=1 \text{ to } N} \{Y_h^1 + Y_v^1 + Y_d^1\} \quad \text{.....} \quad (2.7)$$

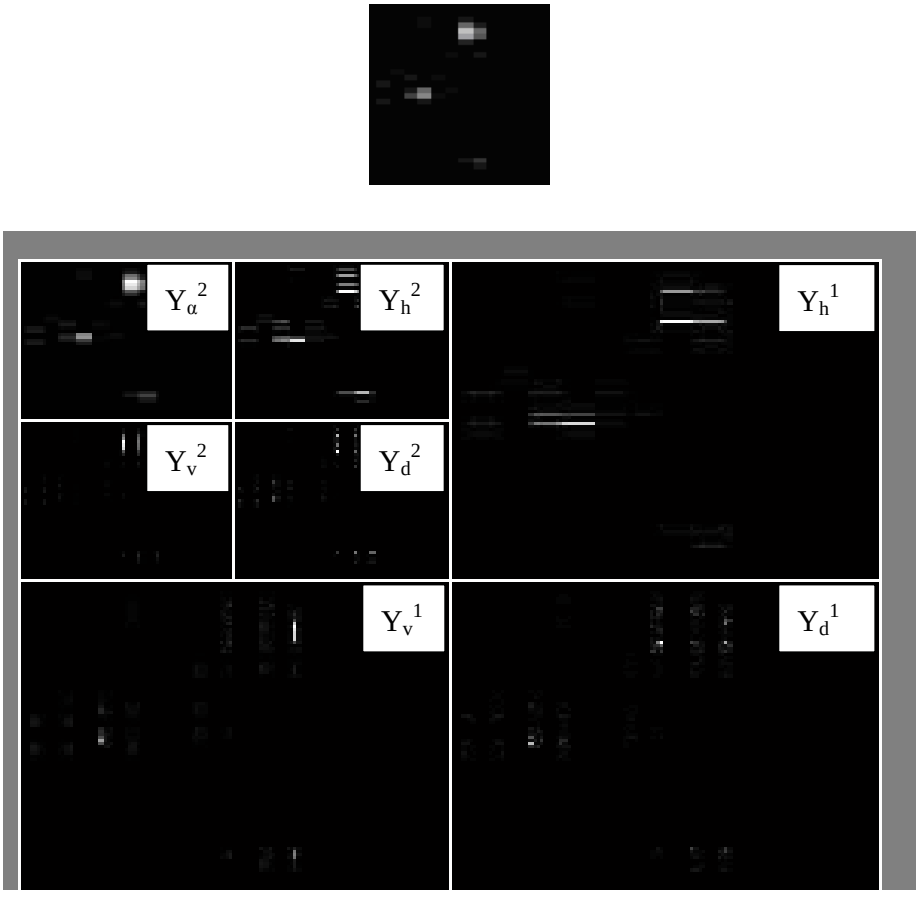

Figure 2.4: Two-Level DWT Lung Tomographic Image a) Region of Interest (ROI) Image b) DWT application to ROI Image

In every decomposition level, the equal share band filters produce signals which are spanning just equal share band of the frequency band. This makes the frequency resolution two when the undetermination in frequency becomes ½ less. The decomposed signals' lengthiness reduces ½ of the lengthiness of the signal in the stage before. So, the first level decomposition of an NxN image is N/2xN/2 when the second level decomposition is N/4xN/4 and goes on like this. When there exists the augmentation of the level of decomposition, frequently but it has taken rough approach of the image. It is attained that wavelets procure a basic rank order fundamental for commentary of the image definition (Cybern, 1984, pp. 363-373).

2.5 SUPPORT VECTOR MACHINES

Nowadays, Support Vector Machine (SVM) which was matured by (Vapnik, 1999, pp. 988-999) has been used in various problems inherited pattern recognition, bioinformatics and text classification (Haykin, 1999, pp. 329-339). The classification system using in medical diagnosis is augmenting day by day and acquires it's notorious by the means of assorted fetching preculiarities, and experimental performance expectation.

2.5.1 SVM CLASSIFIER

SVM is a learning tool derived from the last statistical learning algorithms. On the exaggeration power of learning functioned machines SVM gives certain advantage for limiting. The algorithm of SVM establishes by surveying the input space interior distinct extreme surfaces in the input space a high dimensional peculiarity space internal certain non linear surveying preferred a precedence (internal) or producing in this peculiarities space the Maximal Margin Hyperplane (Bazzani and Bevilacqua, 2000).

As shown in Figure 2.5, in every edge of the hyperplane it severances the data. The hyperplane enlarges the distance between the two parallel hyperplanes which is the distinction of hyperplane. A hypothesis is constructed from the majorities of the margin or distance between these parallel hyperplanes will be more genius generalization error of the classifier.

If the data is trained (x_i, y_i), $i = 1,\ldots,\ell$ are distinction of $w \cdot x + b = 0$ hyperplane, it occurs when $y_i(w \cdot x_i + b) \geq 1$, where $y_i = \pm 1$ are the categories. The margin is $2 / \|w\|$, thus the hyperplane, with maximal margin data severance is:

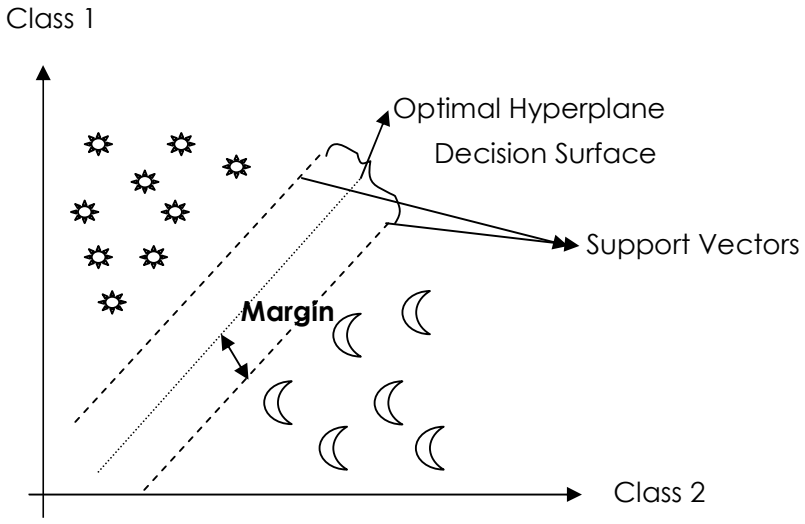

Figure 2.5: A hypothetical classification involving two feature variables ☾s and ☼ s (Bishop, 1995).

Moons denote patterns from Class 1 and Sun symbols denote patterns from Class 2. The decision boundary (shown by the line) is able to provide good separation of the two classes, although there are still a few patterns which would be incorrectly classified by this boundary.

The establishment of two parallel hyperplanes;

{

- Minimize $\|w\|^2 / 2$
- With $y_i(w \cdot x_i + b) \geq 1$

} ……..……..... (2.8)

Constraints are expanded to $y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$ ……. (2.9)
In the sequence of tolerance misclassification errors. (Formula 2.1) transforms to Figure 2.6.

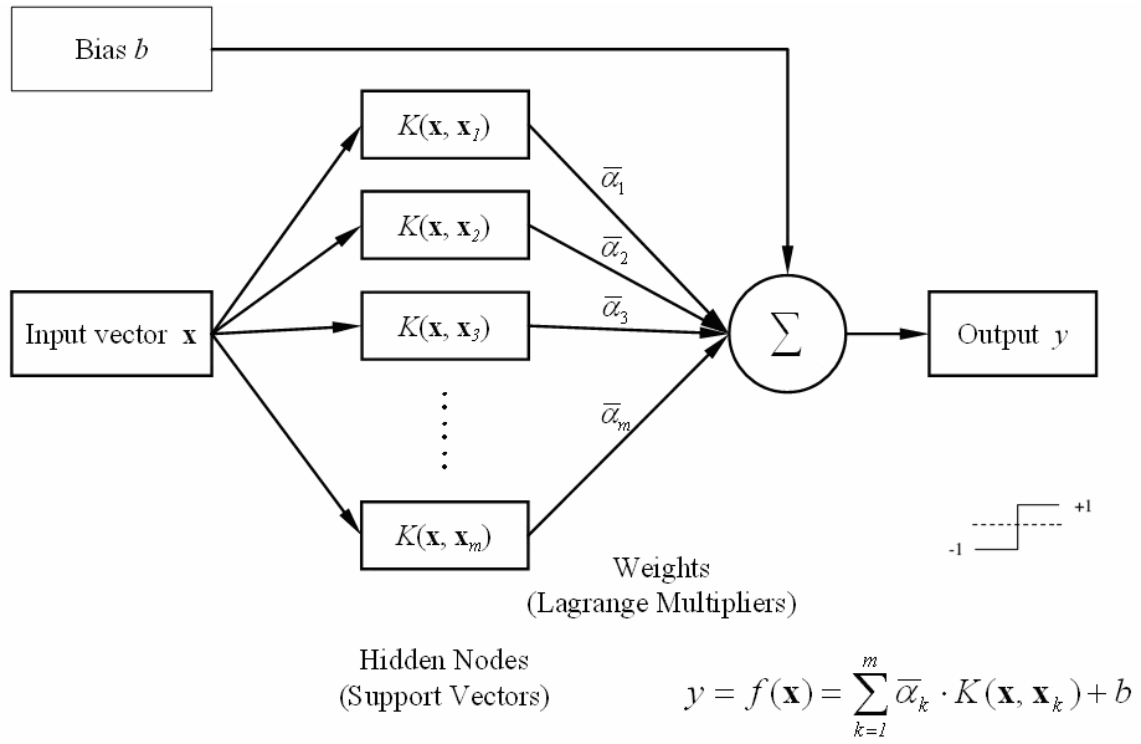

Figure 2.6: Architecture of the Support Vector Machine (Ks indicate Hidden Layer and m is the number of support vectors) (National Cheng Kung University Taiwan Researches).

2.5.2 DESIGN OF SVM CLASSIFIER FOR CLASSIFICATION

2.5.2.1 SVM Kernel Functions

SVM is not capable of accomplish the classification duties in the non linear condition. To conquer this boundary on SVM, kernel approximations are proposed. The kernel function in SVM is the main function of completely surveying the input matrix into a high dimensional peculiarity space. Main types for kernel function are: Gaussian Radial Basis Function (RBF), Sigmoidal, Polynomial, Inverse Multi quadratic and so on. In this study, The Gaussian RBF and polynomial kernels are used.

Polynomial Kernel Function:

$K(x_i - x_j) = (x_i^T x_j + 1)^p$ …….. (2.10)

where p>0 is a constraint.

Gaussian RBF Kernel Function:

$K(x, x_i) = \exp[\gamma|x - x_i|]$ ……..... (2.11)

This two kernel functions are sufficient for the cases, and they have the most majority in use in between SVM functions in SVM.

2.5.2.2 Training Examples and SVM Model Selection

Binary Support Vector Machine has separation capability which can separate positive samples from negative samples in training. Lagrange multipliers α_i parameters of every binary support vector machine find out by reducing the cost function in Formula 2.12 below:

$P(w) = ½ \|w\|^2$ …….. (2.12)

Managed by:

$y_i \cdot f(x_i) \geq +1 - \varepsilon_i\varepsilon_i \geq 0, i = 1,2,\ldots,k$ …….. (2.13)

The cost function L_D is convex and quadratic in terms of the unknown parameters α_i. The aim is to increase the classification margin that managing by constraints. So, in this matter it can be analyzed with dual formula explained as

$L_D = \sum\alpha_i - ½ \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j$ ……..…… (2.14)

Managing by

$0 \leq \alpha \leq C$ and $\sum_i \alpha_i y_i = 0$ ……. (2.15)

When appreciative C, systematic parameter which checks the tolerance for the errors of classification in training step is load, and because of this, the errors will have to be a pay fine. The training vector x_i which corresponds α_i is non zero is called Support Vector.
In this study, 66% percentage split is used which means 66% of the instances would be used for training and the other part of the percentage (34%) would be used for testing. So, 43

instances are used for training in this data set. Percentage Split is used to select model with varied moderation, every one is for conjecture of generalization error.

43 instances are used for training and 83 instances are used for testing in this data set. In this method, it is important that what sort the data obtains partitioned. Each data point obtains to be in a test set and in a training set several times.

As model definition percentage split method is considered in this study. In training and testing RBF kernel and Polynomial (Linear) were used. For the best error level will be reached, kernel parameters were selected like C=100 and γ=0.1.

3. FINDINGS AND TEST RESULTS

3.1 EXPERIMENTAL RESULTS

The inputs of the wavelet coefficients of computerized lung tomographies have been selected to use in MATLAB tool to put on images to computerized data. After then wavelet filter has applied to this data set in WEKA's (Frank and Witten, 2005) (The histogram values of one of the image from the data set are shown in Figure 3.1) SVM SMO (Sequential Minimal Optimization) with RBF Kernel and Polynomial Kernel for training and the classification. The level 2 HAAR wavelet approximation coefficients are taken in consideration (Figure 3.2) and they have been the input of the SVM classifier. In this study, it is used a SVM classifier to find out if it is cancer or not cancer mass of computerized lung tomographic images. For classification purpose, it is used and tested some techniques like shown in Figure 3.3, Figure 3.4 and Figure 3.5.

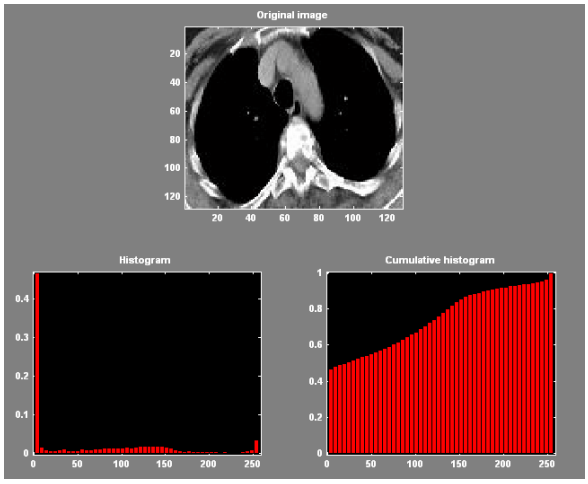

Figure 3.1: The histogram values of one of the images from the data set with Two-Level & HAAR

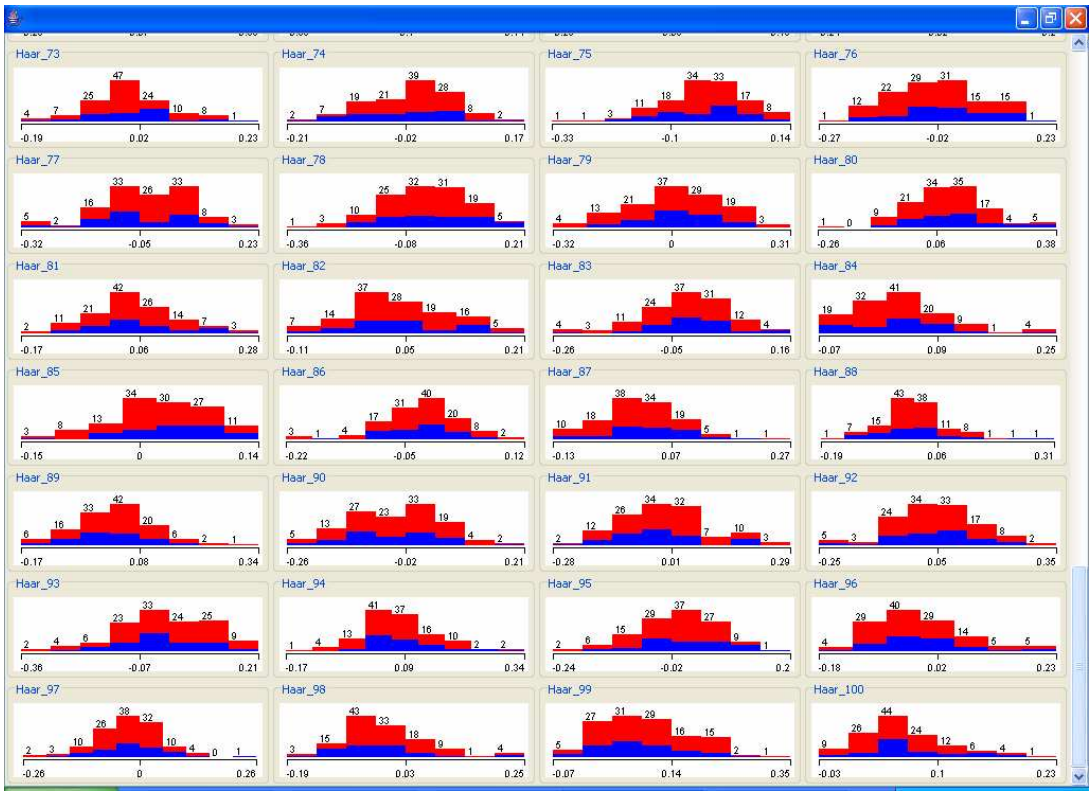

Figure 3.2: HAAR wavelet approximation coefficients

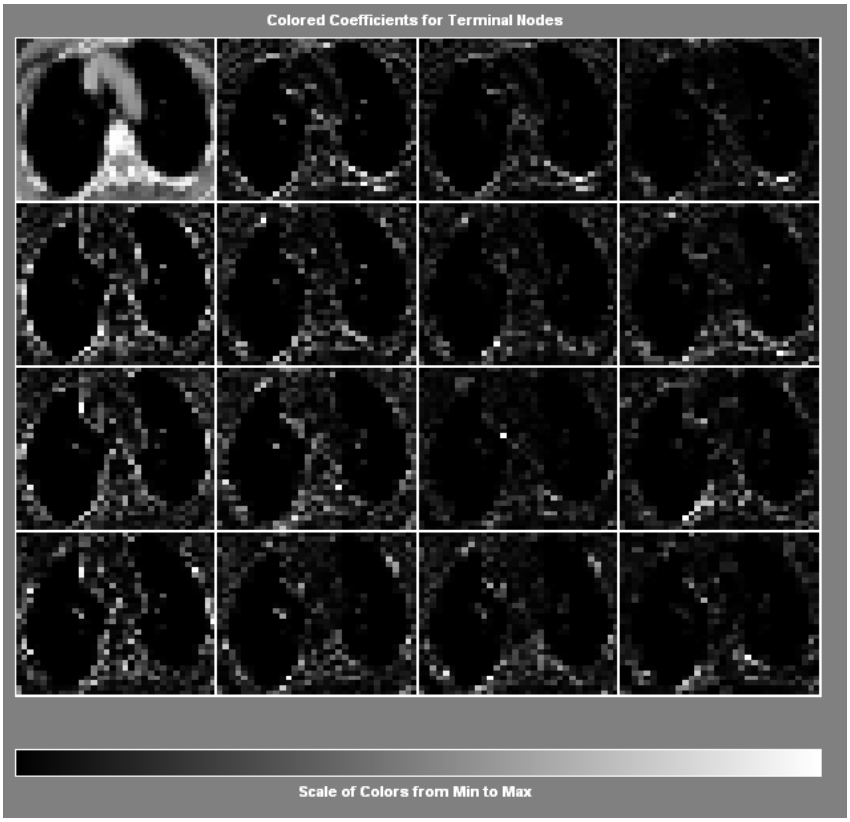

Figure 3.3: Colored Coefficients for Terminal Nodes

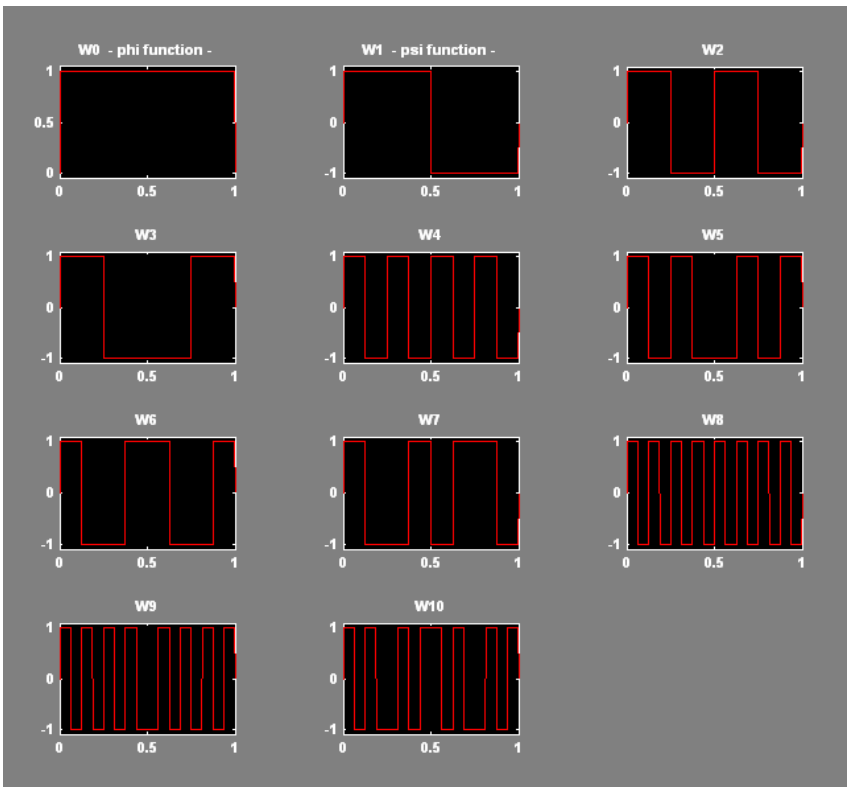

Figure 3.4: Signals of HAAR Wavelet Coefficients for Terminal Nodes

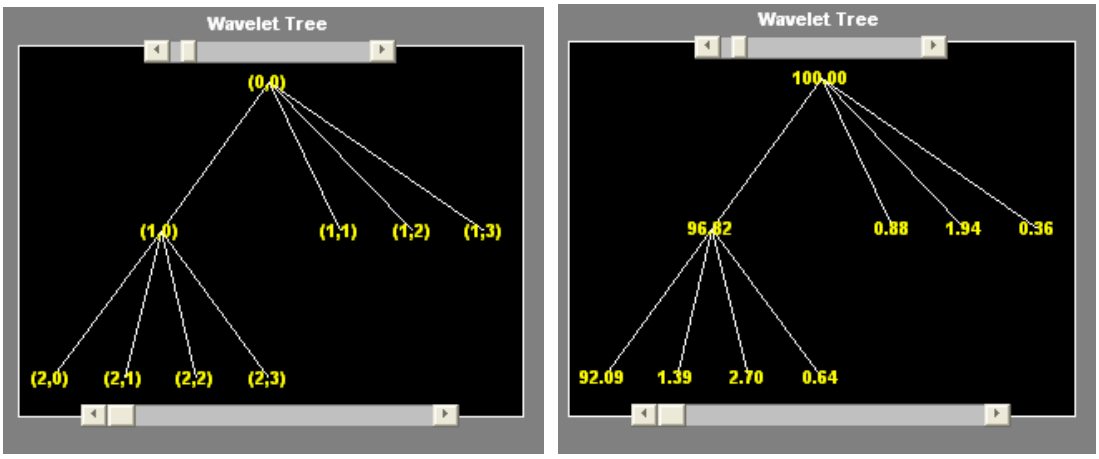

Figure 3.5: Wavelet Tree a) For Depth Position b) For Enegy

Fusion application of Original Image is shown in Figure 3.6.

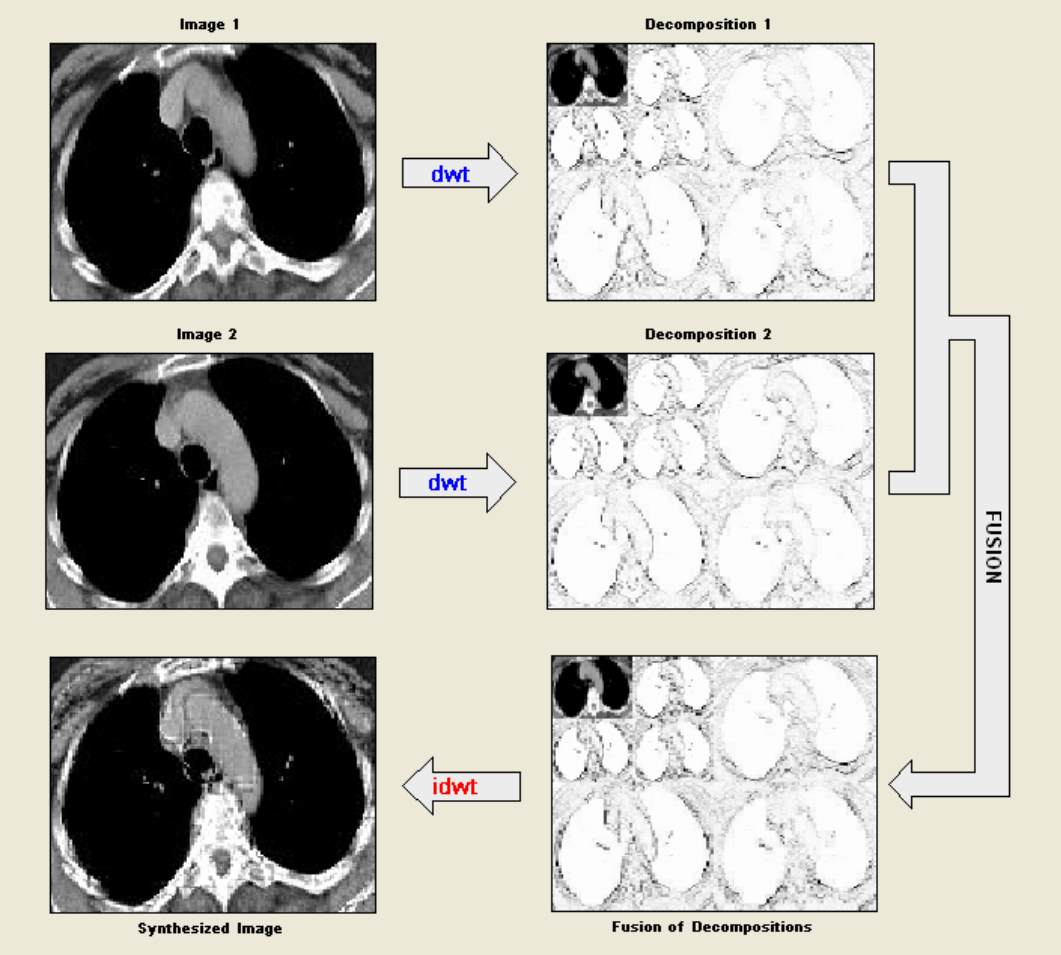

Figure 3.6: Image Fusion for HAAR Level 2.

3.2 CLASSIFICATION AND TEST ACTIVITIES

The usage of SVM inherits training and testing ranks. In this study, for training and testing with specific kernel function, percentage split approval method is decided. For training and testing the polynomial (linear) and RBF kernels were used. The conclusion of classification of the SVM SMO (Sequential Minimal Optimization) classifier was shown with a confusion matrix. The confusion matrices are shown in Table 3.1 of the conclusions of classification of test values for the 43 computerized lung tomographic mass images. It is shown the classification ratio of RBF kernel is bigger than the polynomial (linear) kernel. As it is shown in confusion matrix, if it is taken only training values which is 43 test images from 126 images, when it is deal with 43 instances, 3 of 6 non cancer mass were classified wrong which supposed like cancer mass by the net like cancer mass, in spite of this situation, 7 of 37 cancer mass were classified wrong which supposed like non cancer mass by the net.

Table 3.1: Classification results for the computerized lung tomographic mass images.

Kernel Types of SVM Classifier	Desired Result	Output Result	
		Cancer Mass	Non Cancer Mass
RBF Kernel	Cancer Mass	30	7
	Non Cancer Mass	3	3
Poly(Linear) Kernel	Cancer Mass	24	5
	Non Cancer Mass	9	5

SVM SMO (Sequential Minimal Optimization) classifiers' test performance can be supposed to the computation of sensitivity, specificity and total classification accuracy. The explanation of features listed as follows:

Sensitivity: number of correct classified non cancer mass / total number of non cancer masses

Specificity: number of correct classified cancer mass / total number of cancer masses

Total classification accuracy: number of classified mass / total number of masses

The specificity, sensitivity and total classification accuracy reached from the using of the SVM classifier for classification of computerized lung tomographic mass images are shown in Table 3.2.

Table 3.2: Values of statistical parameters of SVM classifiers.

Kernel Types of SVM Classifier	Statistical parameters (%)		
	Sensitivity	Specificity	Total Classification Accuracy
RBF Kernel	50	81.1	76.74
Poly(Linear) Kernel	35.71	82.76	67.44

The examination of Table 3.2, the best results have achieved in sensitivity, specificity and total classification accuracy when it used RBF Kernel as 76.74 % and Polynomial Kernel as 67.44 % in decreasing order. It can be decided like SVM with RBF kernel could be one of the

hopeful procedures in the classification of computerized lung tomographic mass as (From the values of Table 3.3 and Table 3.4) with the values; Kappa statistic 0.243, Mean absolute error 0.2326, Root mean squared error 0.4822, Relative absolute error 53.5939 %, Root relative squared error 108.053 % with 33 correctly classified instances. In other side, Polynomial Kernel (PolyKernel) error values are Kappa statistic 0.1995, Mean absolute error 0.3256, Root mean squared error 0.5706, Relative absolute error 75.0315 %, Root relative squared error 127.85 % with 29 correctly classified instances.

Table 3.3: Detailed accuracy by class of SVM with RBF Kernel.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.3	0.091	0.5	0.3	0.375	0.605	no
	0.909	0.7	0.811	0.909	0.857	0.605	yes
Weighted Average	0.767	0.558	0.739	0.767	0.745	0.605	

Table 3.4: Detailed accuracy by class of SVM with Poly Kernel.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.5	0.273	0.357	0.5	0.417	0.614	no
	0.727	0.5	0.828	0.727	0.774	0.614	yes
Weighted Average	0.674	0.447	0.718	0.674	0.691	0.614	

In this thesis also the other classifiers have been tested as an experiment, and the results of the other classifiers are demonstrated in Table 3.5.

Table 3.5: The other classifiers.

Classifier	Desired Result	Output Result	
		Cancer Mass	Non Cancer Mass
Logistic	Cancer Mass	23	4
	Non Cancer Mass	10	6
Multilayer Perceptron	Cancer Mass	27	5
	Non Cancer Mass	6	5
Simple Logistic	Cancer Mass	22	7
	Non Cancer Mass	11	3
Voted Perceptron	Cancer Mass	31	9
	Non Cancer Mass	2	1

The other classifiers' sensitivities, specificities and accuracies are shown in Table 3.6. If it is checked by the accuracies (as Table 3.6, Table 3.7, Table 3.8 and Table 3.9) the SVM classifier with RBF Kernel is victorious with the other classifiers that it is examined as it is used 66% Percentage Split as training.

Table 3.6: Values of statistical parameters of different classifiers from SVM classifier.

Classifier	Statistical parameters (%)		
	Sensitivity	Specificity	Total Classification Accuracy
Logistic	37.50	85.19	67.44
Multilayer Perceptron	45.46	84.38	74.42
Simple Logistic	21.43	75.86	58.14
Voted Perceptron	33.33	77.50	74.42

Detailed accuracy by Logistic classifier is shown in Table 3.7. And the errors in Logistic Classifier, which is 67.44 % successful that have revealed, are Kappa statistic 0.2456, Mean absolute error 0.3459, Root mean squared error 0.5634, Relative absolute error 79.7253 %, and Root relative squared error 126.2262 % with 29 correctly classified instances.

Table 3.7: Detailed accuracy by Logistic classifier.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.6	0.303	0.375	0.6	0.462	0.639	no
	0.697	0.4	0.852	0.697	0.767	0.639	yes
Weighted Average	0.674	0.377	0.741	0.674	0.696	0.639	

Table 3.8 shows detailed accuracy by Multilayer Perceptron classifier. Errors of the Classifier which is reached 74.42 % success and which has algorithm is based on nodes related to classes, Kappa statistic 0.3075, Mean absolute error 0.3134, Root mean squared error 0.4781, Relative absolute error 72.214 %, Root relative squared error 107.1232 % with 32 correctly classified instances.

Table 3.8: Detailed accuracy by Multilayer Perceptron classifier.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.5	0.182	0.455	0.5	0.476	0.658	no
	0.818	0.5	0.844	0.818	0.831	0.658	yes
Weighted Average	0.744	0.426	0.753	0.744	0.748	0.658	

Simple Logistic Classifier which has 58.14 % success rate and 25 instances correctly founded, detailed accuracy is shown in Table 3.9 and the errors are Kappa statistic -0.0293, Mean absolute error 0.4041, Root mean squared error 0.5647, Relative absolute error 93.1207 %, Root relative squared error 126.5269 %.

Table 3.9: Detailed accuracy by Simple Logistic classifier.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.3	0.333	0.214	0.3	0.25	0.485	no
	0.667	0.7	0.759	0.667	0.71	0.485	yes
Weighted Average	0.581	0.615	0.632	0.581	0.603	0.485	

At last, in Voted Perceptron classifier which has 74.42 % success rate and 32 correctly classified instances with using perceptrons, detailed accuracy by classifier is shown in Table 3.10 and the with the classifier's error values are Kappa statistic 0.0521, Mean absolute error 0.2582, Root mean squared error 0.5017, Relative absolute error 59.5076 %, Root relative squared error 112.4158 %.

Table 3.10: Detailed accuracy by Voted Perceptron classifier.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.1	0.061	0.333	0.1	0.154	0.474	no
	0.939	0.9	0.775	0.939	0.849	0.474	yes
Weighted Average	0.744	0.705	0.672	0.744	0.688	0.474	

Root Mean Squared Error (RMSE), accuracy and correctly classified instances are very important variations. RMSE (Root Mean Squared Error) values of the classifiers vary between 0.4781 and 0.5706, where accuracy is between 58.14 % and 76.74 % in classifiers. As it is shown in Table 3.11, Although RMSE values of SVM with RBF Kernel and Multilayer Perceptron classifiers have both bigger values than the other classifiers, and nearly

same values with each other, the number of correctly classified instances of SVM with RBF Kernel classifier is bigger than the Multilayer Perceptron classifier.

Table 3.11: Root mean squared error, accuracy and correctly classified instances comparison.

Classifier	Root Mean Squared Error	Accuracy	Correctly Classified Instances
SVM with RBF Kernel	0.4822	76.74	33
SVM with PolyKernel	0.5706	67.44	29
Logistic	0.5634	67.44	29
Multilayer Perceptron	0.4781	74.42	32
Simple Logistic	0.5647	58.14	25
Voted Perceptron	0.5017	74.42	32

Also, RMSE values could be a good indicator of reliability of classifiers. The smaller Root mean squared errors of the classifiers are SVM with Kernel's and Multilayer Perceptron's errors and SVM with Kernel's and Multilayer Perceptron's sensitivities are bigger than the other classifiers (as shown in Table 3.12). In conclusion, SVM with Kernel's both Root mean squared error and sensitivity values have the best condition, and that makes SVM with Kernel classifier the best from inside of the other classifiers.

Table 3.12: Root mean squared error values with statistical parameters.

Classifier	Statistical parameters (%)			Root Mean Squared Error
	Sensitivity	Specificity	Accuracy	
SVM with RBF Kernel	50	81.1	76.74	0.4822
SVM with PolyKernel	35.71	82.76	67.44	0.5706
Logistic	37.50	85.19	67.44	0.5634
Multilayer Perceptron	45.46	84.38	74.42	0.4781
Simple Logistic	21.43	75.86	58.14	0.5647
Voted Perceptron	33.33	77.50	74.42	0.5017

4. DISCUSSION AND CONCLUSION

This thesis has been based on the classification of the lung masses that is cancer or not. The investigation proved that 76.74 % total classification accuracy can be reached by using SVM and RBF kernel function trained with the wavelet approximation coefficients of decomposed signals. Discrete Wavelet Transformation method has been used for the extraction of features from computerized lung tomographic images.

Wavelets have qualification to decompose various frequencies and to protect in hand signal properties in various resolutions.

The SVM classifier performed a superior performance when classification like mapping of the features to a higher dimensional space. Two types of kernel function is applied in SVM method, the best classification accuracy results are taken when RBF kernel is used and the worst classification accuracy results are taken when linear kernel is used between the two classifiers. 66 % of 126 computerized lung tomographies which means 43 instances are used for training and it is reached 33 correctly instances with using SVM with RBF Kernel. RMSE (Root mean squared error) of SVM with RBF Kernel was the smallest value than the other classifiers that have examined and sensitivity parameter value of SVM with RBF Kernel is the biggest value of all classifiers' sensitivity values. As a result, SVM with RBF Kernel gives the best results when classification of the computerized lung tomographic masses.

REFERENCES

Books

Bishop, C, 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press. ISBN: 0-19-853864-2.

Bracewell, R.N., 1999. *The Fourier Transform and its applications*. New York: McGraw-Hill.

Frank, E., and Witten, I.H., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition. San Francisco, CA: Morgan Kaufmann. ISBN: 0-12-088407-0.

Gonzales, R.C., and Woods, R.E., 2002. *Digital image processing*. New Jersey: Prentice Hall.

Haykin, S., 1999. *Neural Networks.* New Jersey: Prentice Hall. ISBN: 0-13-908385-5.

Vapnik, V., 1998. *Statistical learning theory*. New York: Wiley.

Periodical Publications

Adhami, R.R., and Bruce L. M., 1999. Classifying mammographic mass shapes using the wavelet transform modulus-maxima method. *IEEE T Med Imaging 18*, pp. 1170-1177.

Akay, M.F., 2009. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, Volume 36, Issue 2, Part 2, March 2009, pp. 6357-6361.

Ateşçi, Y.Z., Çınar, M., Engin, E. Z., and Engin, M., 2009. Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Systems with Applications*, Volume 36, Issue 3, Part 2, April 2009, pp. 6357-6361.

Bowman, R.V., Clarke, B.E., Duhig, E.E., Fry, M.J., McCowan, I.A., Moore, D.C., and Nguyen, A.N., 2007. Collection of Cancer Stage Data by Classifying Free-text Medical Reports. *Journal of the American Medical Informatics Association*, Volume 14, Issue 6, November-December 2007, pp. 736-745.

Brislawn, C.M., 1995. Fingerprints go digital. *Notices Amer Math Soc*., pp. 1278-1283.

Chaplot, S., and Patnaik, L.M., 2006. Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. *Biomedical Signal Processing and Control*, pp. 86-92.

Chen, L., Day, P.J., Fan, S.T., Ho, D.W., Lam, B.Y., Lee, N.P.Y., Leng, X., Luk, J.M., Peng, J., and Sham, P.C., 2007. Artificial neural networks and decision tree model analysis of liver cancer proteomes. *Biochemical and Biophysical Research Communications*, Volume 361, Issue 1, 14 September 2007, pp. 68-73.

Cho, S.B., and Kim, K.J., 2008. Evolutionary ensemble of diverse artificial neural networks using speciation. *Neurocomputing*, Volume 71, Issues 7-9, March 2008, pp. 1604-1618

Cho, S.B., and Kim, K.J., 2004. Prediction of colon cancer using an evolutionary neural network. *Neurocomputing*, Volume 61, October 2004, pp. 361-379.

Dong, X., Sun, G., and Xu, G., 2006. Tumor tissue identification based on gene expression data using DWT feature extraction and PNN classifier. *Neurocomputing*, Volume 69, Issues 4-6, January 2006, pp. 387-402.

Dy, S.M., Phillips-Wren, G., and Sharkey, P., 2008. Mining lung cancer patient data to assess healthcare resource utilization. *Expert Systems with Applications*, Volume 35, Issue 4, November 2008, pp. 1611-1619.

Ennett, C.M., Frize, M., Stevenson, M., and Trigg, H.C.E., 2001. Clinical decision support systems for intensive care units: using artificial neural networks. *Medical Engineering & Physics*, Volume 23, Issue 3, April 2001, pp. 217-225.

Gammerman, A., Hammer, B., Kostrzewa, M., Schleif, F.M., and Villmann, T., 2008. Cancer informatics by prototype networks in mass spectrometry. *Artificial Intelligence in Medicine*, 7 September 2008, pp. 15-75.

Kabrisky, M., Rogers, S.K., and Ruck, D.W., 1994. Artificial neural networks for early detection and diagnosis of cancer. *Cancer Letters*, Volume 77, Issues 2-3, 15 March 1994, pp. 79-83.

Koenderink, J., 1984. The structure of images. *Biol Cybern*, pp. 363-370.

Kusiak, A., and Shah, S., 2007. Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine*, Volume 37, Issue 2, February 2007, pp. 251-261.

Mallat, S.G., 1980. A theory of multiresolution signal decomposition: the wavelet representation. *IEEE T. Pattern Anal 11*, pp. 674–693.

Vapnik, V.N., 1999. An overview of statistical learning theory. *IEEE T. Neural Network*, pp. 988-999.

Other Publications

Bazzani, A., and Bevilacqua, A.D., 2000. Automatic detection of clustered microcalcifications in digital mammograms using an SVM classifier. *ESANN'2000 proceedings, European Symposium on Artificial Neural Networks*. Bruges, Belgium.

Cancer Care Ontario, (1964-2002). Percentage of lung cancer associated with smoking: Probably more than 90% of lung cancers in men and at least 70% in women are directly attributable to cigarette smoking. *In Proceedings of Cancer Incidence and Mortality in Ontario Conference*. Cancer Care Ontario, Ontario.

Cancer Index. 2009. *Cancer informations and statistics of the world*. Available from: http://www.cancerindex.com [cited 12 May 2009].

Mathsoft Wavelet resources. 1997. *A great collection of theory and application oriented articles*. Available from: http://mw.mthsof.codwavelets.html [cited 10 November 2008].

National Cheng Kung University Taiwan Researches. 2008. *Researches and articles*. Available from: http://research.ncku.edu.tw/re/articles/e/20081205/images/0810200222104gLmpu.jpg [cited 15 May 2009].

Özdemir, N., (2009). Erken evre küçük hücreli dışı akciğer kanseri tedavisini ne kadar biliyoruz?. *In Proceedings of TAKD (Türk Akciğer Kanseri Derneği) Symposium*. Ankara: TAKD.

Turkish Republic Ministry of Health, Annual Report, 2009, http://www.saglik.gov.tr [cited 15 May 2009].

CURRICULUM VITAE

Name Surname: Başak Sarıkaya

Address: Fazıl Hüsnü Dağlarca Street Filiz Apartment No: 4/7 Caferağa Quarter / Kadıköy

Birthplace and Birthday: Ankara, 11.09.1983

Foreign Language: English

Primary School: Arı College, 1994

Middle School: Arı College, 1998

High School: Arı College Science High School, 2001

Undergraduate: Çankaya University Computer Engineering, 2005

Graduate: Bahçeşehir University Computer Engineering, 2009

Institute Name: Institute of Sciences

Program Name: Computer Engineering Graduate Program

Work Experience: T.C. Ziraat Bankası, Project Management Unit, 2008-2009

T.C. Ziraat Bankası, Investment Applications/Stocks and Shares, Software Specialist, 2007-2008

Link Bilgisayar Sistemleri Yazilimi ve Donanimi Sanayi ve Ticaret Anonim Sirketi, Software Developer Specialist, 2006

Modular Hightech, Database and Web Design Assistance Specialist,2004

Iller Bankasi Genel Mudurlugu Bilgi Islem Daire Baskanligi Hardware Assistant, 2003

Çankaya University, Student Assistant, 2003-2005