3-19-2012

Developing Predictive Models for Lung Tumor Analysis

Satrajit Basu
University of South Florida, satrajit@mail.usf.edu

Developing Predictive Models for Lung Tumor Analysis

by

Satrajit Basu

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Lawrence O. Hall, Ph.D.
Dmitry Goldgof, Ph.D.
Sudeep Sarkar, Ph.D.

Date of Approval:
March 19, 2012

Keywords: Radiomics, Classifiers, CT-scan, Image Features, Texture Features, Support Vector Machine, Decision Trees, Ensemble, Feature Selection, Parameter Tuning

DEDICATION

To my advisor, Dr. Lawrence Hall, for all of his guidance and support.

To my parents, Sujay and Anita Basu, who are my constant source of strength and inspiration.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank Dr. Lawrence Hall and Dr.Dmitry Goldgof for their invaluable guidance and support in my research. I would particularly like to thank them for their patience and belief in my work. I would like to thank Dr. Sudeep Sarkar for taking the time to be a part of my committee and providing valuable inputs regarding my thesis. I would like to thank Dr. Yuhua Gu and Dr. Virendra Kumar for their invaluable contributions towards my work. I would also like to thank Dr. Robert Gatenby and Dr. Robert Gillies from Moffitt Cancer Center for their support in this work. I would like to thank Dr. Nagarajan Ranganathan, Dr. Abraham Kandel, Dr. Rangachar Kasturi, Dr. Adriana Iamnitchi, Dr. Rafael Perez, Dr. Xiaoning Qian, Dr. Yicheng Tu, Theresa Collins, Yvette Blanchard and all other members of the CSE department for their support. I would like to thank John Korecki and everyone in my research group for their invaluable inputs. I would also like to thank my past and present colleagues in CSE and my friends, Saurabh, Diego, Caitrin, Ashish, Ingo, Mehrgan, Yue, Himanshu, Anand, Ravi, Soumyaroop and others.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A CT-scan of lungs has become ubiquitous as a thoracic diagnostic tool. Thus, using CT-scan images in developing predictive models for tumor types and survival time of patients afflicted with Non-Small Cell Lung Cancer *(NSCLC)* would provide a novel approach to non-invasive tumor analysis. It can provide an alternative to histopathological techniques such as needle biopsy. Two major tumor analysis problems were addressed in course of this study, tumor type classification and survival time prediction. CT-scan images of 109 patients with NSCLC were used in this study. The first involved classifying tumor types into two major classes of non-small cell lung tumors, Adenocarcinoma and Squamous-cell Carcinoma, each constituting 30% of all lung tumors. In a first of its kind investigation, a large group of 2D and 3D image features, which were hypothesized to be useful, are evaluated for effectiveness in classifying the tumors. Classifiers including decision trees and support vector machines *(SVM)* were used along with feature selection techniques (wrappers and relief-F) to build models for tumor classification. Results show that over the large feature space for both 2D and 3D features it is possible to predict tumor classes with over 63% accuracy, showing new features may be of help. The accuracy achieved using 2D and 3D features is similar, with 3D easier to use. The tumor classification study was then extended by introducing the Bronchioalveolar Carcinoma *(BAC)* tumor type. Following up on the hypothesis that Bronchioalveolar Carcinoma is substantially different from other NSCLC tumor types, a two-class problem was created, where an attempt was made to differentiate BAC from the other two tumor types. To make a three-class problem a two-class problem, misclassification amongst Adenocarcinoma and Squamous-cell Carcinoma were ignored. Using the same prediction models as the previous study and just 3D image features, tumor classes were predicted with around 77% accuracy. The final study involved predicting two

year survival time in patients suffering from NSCLC. Using a subset of the image features and a handful of clinical features, predictive models were developed to predict two year survival time in 95 NSCLC patients. A support vector machine classifier, naive Bayes classifier and decision tree classifier were used to develop the predictive models. Using the Area Under the Curve *(AUC)* as a performance metric, different models were developed and analyzed for their effectiveness in predicting survival time. A novel feature selection method to group features based on a correlation measure has been proposed in this work along with feature space reduction using principal component analysis. The parameters for the support vector machine were tuned using grid search. A model based on a combination of image and clinical features, achieved the best performance with an AUC of 0.69, using dimensionality reduction by means of principal component analysis along with grid search to tune the parameters of the SVM classifier. The study showed the effectiveness of a predominantly image feature space in predicting survival time. A comparison of the performance of the models from different classifiers also indicate SVMs consistently outperformed or matched the other two classifiers for this data.

CHAPTER 1

INTRODUCTION AND RELATED WORK

A Computed Tomography *(CT)* scan is an extensively used imaging technique, vital in the field of thoracic radiology [1]. Recent advances in both image acquisition and image analysis techniques allow semi-automated tumor segmentation and extraction of numerous features from images (e.g. texture). This falls under the broader category of techniques termed "Radiomics". Radiomics involve the high throughput extraction of quantitative imaging features from radiological images with the intent of creating mineable data. The data can then be used to build descriptive and predictive models relating image features to phenotypes or gene-protein signatures. The core hypothesis of radiomics is that these models, which can include biological or medical data, can provide valuable diagnostic, prognostic or predictive information.

The study being presented here is restricted to the analysis of Non-Small Cell Lung Cancer *(NSCLC)*. Of the possible analytical studies in Radiomics, this work concentrates on two aspects of predictive models, tumor type classification and survival time prediction. Identification of tumor type or class is essential in risk assessment and determining treatment options. Classifying tumor based on CT-scan images could provide an opportunity for faster diagnosis of tumor types without the need for invasive procedures. Predicting survival time of a patient is also essential in terms of determining aggressiveness of a tumor along with possible prognosis.

1.1 Related Work

In this section we will review some of the work that has already been done by using image features for tumor analysis. This will provide an idea of the scope of this work and

will also be helpful in the understanding of the different choices being made during the study.

First, some of the work done in the realm of lung cancer analysis using CT-features will be looked at. Then the focus will shift to the specific problem domains of tumor type prediction and survival time analysis.

1.1.1 Image Based Tumor Analysis

Ganeshan et al. [2] has shown that features extracted from CT images of lung tumors can be used to find a correlation with glucose metabolism and stage information. Extensive work has been done in the study of pulmonary nodules in the lung. The work by Samala et al. [3] looked at finding the optimum selection of image features to represent lung nodules. Those features were then implemented into a classification module of a computer-aided diagnosis system. Way et al. [4] wanted to distinguish benign nodules from malignant ones based solely on texture based image features. Lee et al. [5] also performed a detailed study on the usefulness of image features in the classification of pulmonary nodules based on CT-scan images. The work by Zhu et al. [6] shows the effectiveness of a support vector machine based classifier in classifying benign and malignant pulmonary nodules. Work has also been done by Al-Kadi et al. [7] in differentiating between aggressive and non-aggressive malignant lung tumors using texture analysis of Contrast Enhanced *(CE)* CT scan images. The use of fractal image features in tumor analysis can be found in the work of Kido [8]. The high level of information content within CT scans was highlighted by correlating imaging features with global gene expression in hepatocellular carcinoma [9]. Segal et al [9] showed that combinations of twenty-eight image features obtained from CT images of liver cancer could reconstruct 78% of the global gene expression profiles.

1.1.2 Survival Analysis

In terms of survival time prediction in cancer patients, Burke et al.[10] showed the effectiveness of using Area Under the ROC Curve *(AUC)* as a performance measure by means of 5 year survival prediction for breast and colorectal cancer patients.

Work on predicting survival time for patients suffering from NSCLC is being conducted by Hugo Aerts et al. at MAASTRO, Netherlands. The work titled, "Using Advanced Imaging Features for the Prediction of Survival in NSCLC" [11], dealt with building a predictive model built on proximal support vector machine [12]. Three sets of features, histogram, texture and shape features were used in developing the predictive model. These features were used alongside clinical features. The experiment performed was to predict two year survival in 412 patients with NSCLC, stage I-III, treated with either radiotherapy or chemo-radiotherapy. Each patient had an CT scan for treatment planning. All the scans have fixed resolution, with slice thickness of 3 mm and voxel size of $0.98x0.98$ mm.

On the given dataset, a total of 101 features, from the three classes discussed earlier, were extracted. The feature values formed the attributes for each patient instance. The working of the predictive model can be looked upon as a blackbox as represented in Figure 1.1. The experiment performed was a 90% training and 10% test data split, with random splits repeated over 1000 iterations. The combination of the clinical and image features resulted in an average AUC of 0.70 over 1000 iterations.

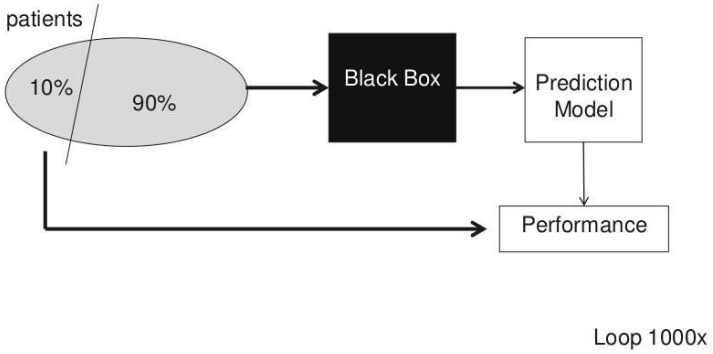


Figure 1.1: Analysis Setup: Prediction of Two Year Survival

1.2 Contributions

Review of existing literature has shown that studies have concentrated mostly on the analysis of CT-scan images to detect tumors and other anomalies of the lungs. However, minimal work has been done in attempting to classify tumor classes based on these images for which new ground is broken here. The common practice to determine the tumor class is to perform a histopathological analysis on tissue samples obtained by invasive techniques such as a needle biopsy. As time and cost are crucial factors when it comes to the treatment of a lung tumor, an automated image based classifier could act as a precursor to histopathological analysis, thus enabling the kick-starting of class specific treatment procedures. Based on CT-scan images of 74 patients with Non-Small Cell Lung Cancer (NSCLC) the first task was to develop an effective model to classify it into two subtypes, Adenocarcinoma and Squamous-cell Carcinoma [13]. These two tumor types constitute 30% of all lung tumor types [14]. This study made use of 4 different classifier algorithms along with relief-F and wrapper feature selection methods. Building upon a vast number of image features, this study helped to present a robust comparative analysis of the effectiveness of 2D and 3D image features as a basis for developing classifier models.

The study was then extended to include Bronchioalveolar Carcinoma [15], and using the same 3D feature set, the effectiveness as well as stability of the classifier models were evaluated. As part of this experiment and later, in a modified manner, for survival time prediction, a novel approach to feature selection based on feature correlation is presented. The study on survival time analysis was carried out using AUC as the performance metric. Though, working with a limited data-set, using a combination of feature space reduction using principal component analysis and parameter tuning of an SVM using grid search, a near comparable result was achieved to the work being conducted by Aerts et al. 1.1 on a much larger and uniform data-set.

1.3 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 describes the methodology followed, in processing the CT-scan images, including segmentation of the tumor object. In the same chapter, the features used in the course of this study has been detailed. In Chapter 3, we discuss the underlying algorithms upon which the various predictive models are developed. Classification algorithms and feature selection techniques are described here. In Chapter 4, the basic experimental setup followed by a detailed description of considerations, implementation, and results for each problem are presented. Finally, Chapter 5 contains the conclusions.

CHAPTER 2

IMAGE PREPROCESSING AND FEATURE EXTRACTION

In this chapter we discuss preparation of the image data, followed by a detailed discussion of the features that were extracted from the images. In Section 2.1 the process of image segmentation, tumor identification and isolation of the tumor object is presented. Then in Section 2.2 the vast range of image features which were analyzed for the study are presented. The four major image feature sets and their component features are each briefly described. The clinical features utilized in the survival time analysis are presented in Section 2.3. The general workflow of the process of developing and using predictive models is represented in Figure 2.1.

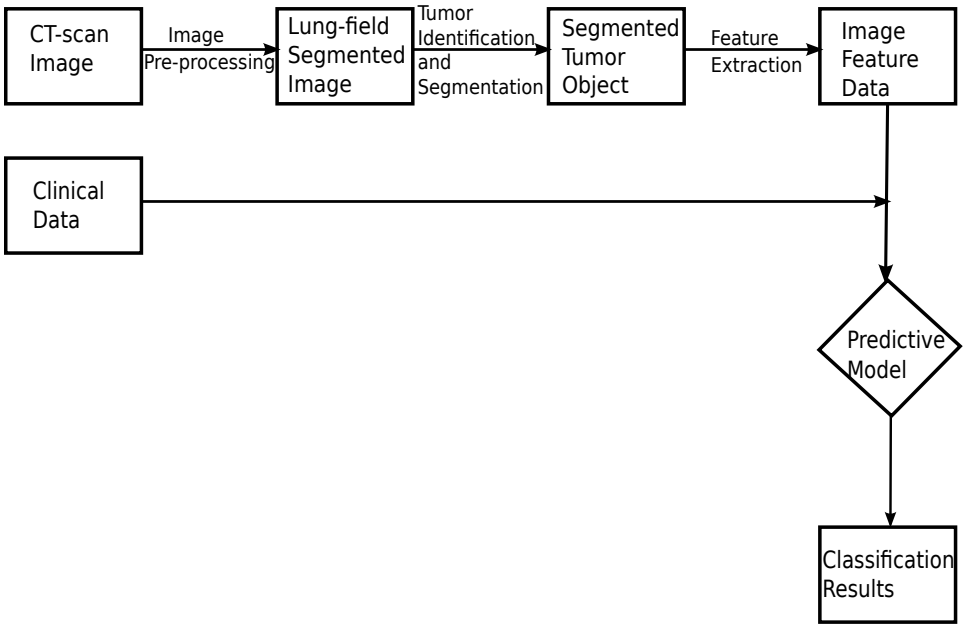
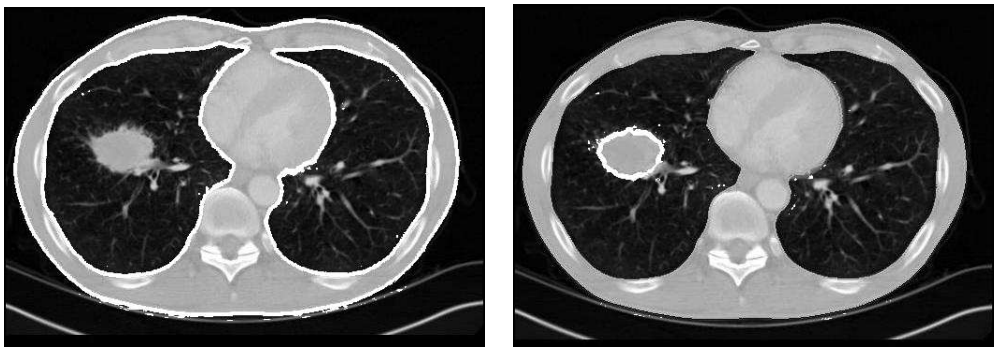


Figure 2.1: Schematic Representation of the Workflow Involved in Preparing Data for Predictive Models.

2.1 Image Preprocessing

The initial segmentation of the CT-scan images which segments out the lung region from the rest of the body was done using the built-in segmentation algorithm provided in the Lung Tumor Analysis (LuTA) software suite of Definiens [16]. On completion of the lung field segmentation, tumor identification was manually conducted by one of the radiologists at the H. Lee Moffitt Cancer Center or a person with expertise in identifying lung tumors. The tumor, upon identification, was segmented out using LuTA's built-in region growing algorithm. The initial seed point for the algorithm was provided by the expert. The algorithm finds the tumor boundary across the image sequences. This boundary contains the tumor objects in each slice of the CT-image sequence. Figures 2.2(a) and 2.2(b) show the lung with tumor and with the tumor boundary outlined after region growing, respectively.



(a) Segmentation of CT image to define the lung region

(b) Defining tumor boundary through region growing

Figure 2.2: Sample CT-Image Slice

2.2 Image Features

The image feature extraction algorithms were written in C++ using Visual Studio 2003 and the executables were embedded into the LuTA software. 102 2D features and 215 3D features were developed for the study. The image feature extraction was done only on the tumor objects. The major customized 2D/3D features types are the following,

- Geometric features
- Morphological features
- Texture features
- Intensity based features/ Histogram features

In this section, each feature subtype and the individual features that belong to it are discussed in brief. Some features have been developed for both 2D and 3D image spaces, which is indicated in the feature description.

2.2.1 Geometric Features

The first set of features to be looked at are geometric features. Geometric features, in both 2D and 3D provide vital structural information about the tumor object being analyzed. The geometric features evaluated were deemed useful in analyzing and quantifying biomedical images such as CT-scans [17].

- *Area (2D, 3D)*

  The number of pixels forming an image object rescaled by using unit information. In scenes that provide no unit information, the area of a single pixel is 1 unit. Consequently, the area of an image object is the number of pixels forming it. If the image data provides unit information, the area of an image object is the true area covered by one pixel times the number of pixels forming the image object. Area is measured as,

  $$A_v = \#P_v * u^2$$

  where, A_v is the area of image object v; P_v is the set of pixels of the image object v; $\#P_v$ is the total number of pixels contained in P_v; u is the pixel size in coordinate system units. u=1, when unit is pixel.

- *Border Length/Surface Area (2D, 3D)*

  The border length of an image object is defined as the sum of edges of the image object that are shared with other image objects or are situated on the edge of the entire scene. For a torus and other image objects with holes the border length sums the inner and outer border.

  For a 3D image object the corresponding feature is the surface area. It is measured as the sum of border lengths of all image object slices multiplied by the spatial distance between the slices. For torus and image objects with holes the surface area is the sum of the inner and outer surface areas, as in 2D. The expressions for border length in 2D and surface area in 3D are the following,

  – 2D

  $$b_v = b_o + b_i$$

  – 3D

  $$s_v = (\sum_{n=1}^{\#(slices)} b_v(Slice))u_{slices} + b_v(Z)$$

  where, b_v is the border length of image object v; s_v is the surfa area of the image object v; b_o is the length of outer border; b_i is the length of inner border; $b_v(Slice)$ is the border length of image object slice; $b_v(Z)$ is the border length of image object in Z-direction; u_{slices} is the spatial distance between slices in the coordinate system unit.

- *Length/Thickness (3D)*

  For this feature, the length-to-thickness ratio of the image object is measured.

- *Length/Width (2D, 3D)*

  The length-to-width ratio of an image object is measured in both 2D and 3D space. There are two methods to approximate the length/width ratio of an image object:

- The ratio length/width is identical to the ratio of the eigenvalues of the covariance matrix, with the larger eigenvalue being the numerator of the fraction:

$$\gamma_v^{EV} = \frac{\lambda_1(V)}{\lambda_2(V)}$$

- The ratio length/width can also be approximated using the bounding box:

$$\gamma_v^{BB} = \frac{K_v^{bb'}}{\#P_v}$$

where, λ_i represents the i^{th} eigenvalue; P_v is the set of pixels of the image object v; $\#P_v$ is the total number of voxels contained in P_v.

Both calculations are done and compared, with the smaller of the two results being returned as the feature value.

- *Length (2D, 3D)*

The length of an image object in 2D is calculated using the length-to-width ratio. The length of an image object is the largest of three eigenvalues of a rectangular 3D space that is defined by the same volume as the image object and the same proportions of eigenvalues as the image object. The length of an image object can be less than or equal to the largest dimension of the smallest rectangular 3D space enclosing the image object. Length is expressed as,

$$l_v = \sqrt{\#P_v \cdot \gamma_v}$$

where, P_v is the set of pixels of the image object v; $\#P_v$ is the total number of pixels contained in P_v; γ_v is the length/width ratio of an image object v.

- *Thickness (3D)*

The thickness of an image object is the smallest of the three eigenvalues of a rectangular 3D space that is defined by the same volume as the image object and the same proportions of eigenvalues as the image object.

The thickness of an image object can be either smaller than or equal to the smallest dimension of the smallest rectangular 3D space enclosing the image object.

- *Width (2D, 3D)*

  The width of an image object is the middle of the three eigenvalues of a rectangular 3D space that is defined by the same volume as the image object and the same proportions of eigenvalues as the image object. The width of an image object can be smaller than or equal to the middle dimension of the smallest rectangular 3D space enclosing the image object.

  The width of an image object is calculated using the length-to-width ratio.

$$w_v = \frac{\#P_v}{\gamma_v}$$

  where, P_v is the set of pixels of the image object v; $\#P_v$ is the total number of voxels contained in P_v; γ_v is the length/width ratio of an image object v.

- *Number of Pixels (2D, 3D)*

  In case of 2D, the number of pixels forming the tumor object for the slice under consideration is measured. In the 3D space, the number of pixels forming the entire tumor object for the volume is measured.

- *Volume (3D)*

  In 3D feature space, volume information provides substantial information regarding the tumor size. The number of voxels forming an image object rescaled by using unit information for the x and y coordinates and the distance information between slices. Volume is measured as,

$$V_v = \#P_v * u^2 * u_{slices}$$

  where, V_v is the volume of image object v; P_v is the set of pixels of the image object v; $\#P_v$ is the total number of voxels contained in P_v; u is the size of a slice pixel in the

coordinate system unit; u_{slices} is the spatial distance between slices in the coordinate system unit.

- *Asymmetry (2D, 3D)*

  The more elongated an image object, the more asymmetric it is. For an image object, an ellipse is approximated. In the case of 2D image feature, asymmetry can be expressed by the ratio of the lengths of the minor and the major axis of this ellipse. In the case of 3D features, asymmetry is calculated from the ratio between the smallest and the largest eigenvalues of the image object. The feature value increases with the asymmetry. The 2D and 3D measure for Asymmetry are,

  – 2D

  $$Asymmetry = \frac{\sqrt{(VarX+VarY)^2+(VarXY)^2-VarX \cdot VarY}}{VarX+VarY}$$

  – 3D

  $$Asymmetry = 1 - \frac{\sqrt{\lambda_{min}}}{\sqrt{\lambda_{max}}}$$

  where, λ represents the corresponding eigenvalue; $VarX$ and $VarY$ represent the variance of X and Y respectively.

- *Compactness (2D, 3D)*

  In the 2D domain, this feature is similar to the border index feature, but instead of the border, it is based on the area. However, the more compact an image object is, the smaller its border appears to be. The compactness of an image object is calculated by the product of the length and the width and divided by the number of its pixels.

  The measure for the compactness of a 3D image object is calculated by a scaled product of its three eigenvalues divided by the number of its pixel/voxel. A factor of 2 is included with each eigenvalue, since λ_i* eigenvectors represent otherwise half axes of an ellipsoid defined by its covariance matrix. The chosen approach thus provides an estimate of a cuboid occupied by the object. Compactness is measured as,

– 2D

$$Compactness = \frac{l_v * w_v}{\#P_v}$$

– 3D

$$Compactness = \frac{2\lambda_1 * 2\lambda_2 * 2\lambda_3}{V_v}$$

where, λ represents the corresponding eigenvalue; l_v is the length of the image object; w_v is the width of the image object; P_v is the set of pixels of the image object v; $\#P_v$ is the total number of voxels contained in P_v.

- *Density (2D, 3D)*

  A density feature describes the spatial distribution of the pixels of an image object. In 2D, the ideal compact shape on a pixel raster is the square. The more an image object is shaped like a square, the higher its density, while a filament like structure is indicative of lower density. The density is calculated by the number of pixels forming the image object divided by its approximated radius based on the covariance matrix.

  In 3D, the ideal compact shape on a pixel raster is the cube. As in the case of 2D, the more cuboid the shape, the higher the density and the more the object is shaped like a filament, the lower is its density. It is calculated by the edge of the volume fitted cube divided by the fitted sphere radius.

  – 2D

$$Density = \frac{\sqrt{\#P_v}}{1+\sqrt{VarX+VarY}}$$

  – 3D

$$Density = \frac{\sqrt[3]{V_v}}{\sqrt{Var(X)+Var(Y)+Var(Z)}}$$

where, P_v is the set of pixels of the image object v; $\sqrt{\#P_v}$ is the diameter of a square object with $\#P_v$ pixels; $\sqrt{VarX + VarY}$ is the diameter of the ellipse; V_v is the volume of the image object v; $\sqrt[3]{V_v}$ is the edge of the volume fitted cube; $VarX$ and $VarY$ represent the variance of X and Y respectively; $\sqrt{Var(X) + Var(Y) + Var(Z)}$ is the radius of the fitted sphere.

- *Shape Index (2D, 3D)*

  A shape index feature provides perceptual representation to the coverage of the shape of an image object. The smoother the border of an image object, the lower is its shape index. In 2D space it is calculated from the border length feature of the image object divided by four times the square root of its area. In the 3D space it is measured by dividing the border length feature by the volume of the image object.

  – 2D

  $$Shape\ Index = \frac{b_v}{4\sqrt{\#P_v}}$$

  – 3D

  $$Shape\ Index = \frac{b_v}{V_v}$$

  where, b_v is the border length of the image object v; $4\sqrt{\#P_v}$ is the border of the square with area $\#P_v$; V_v is the volume of the image object v.

- *Border Index (2D, 3D)*

  The border index feature is similar to shape index feature, but it uses a rectangular approximation instead of a square. The smallest rectangle enclosing the image object is created. The border index is then calculated as the ratio of the Border length feature of the image object to the border length of this smallest enclosing rectangle.The more rough or jagged an image object is, the higher its border index.

  $$Border\ Index = \frac{b_v}{2(l_v + w_v)}$$

where, b_v is the border length of the image object v; l_v is the length of the image object; w_v is the width of the image object.

- *Elliptic Fit (2D, 3D)*

  Elliptic fit describes how well an image object fits into, an ellipse/ellipsoid (2D/3D), of similar size and proportions. While 0 indicates no fit, 1 indicates a complete fitting image object. The calculation is based on an ellipse/ellipsoid (2D/3D) with the same area/volume (2D/3D) as the considered image object. The proportions of the ellipses/ellipsoids (2D/3D) are equal to the proportions of the length-to-width/length-to-width-to-thickness (2D/3D) of the image object. The area/volume (2D/3D)of the image object outside the ellipse/ellipsoid (2D/3D) is compared with the area/volume (2D/3D) inside the ellipse/ellipsoid (2D/3D) that is not filled out with the image object.

  – 2D

  $$\phi = 2 \cdot \frac{\#\{(X,Y) \in P_v : \epsilon_v(X,Y) \leq 1\}}{\#P_v} - 1$$

  – 3D

  $$\phi = 2 \cdot \frac{\#\{(X,Y,Z) \in P_v : \epsilon_v(X,Y,Z) \leq 1\}}{\#P_v} - 1$$

  where, ϕ is the elliptic fit; $\epsilon_v(X,Y)$ is the elliptic distance at pixel (X,Y); $\epsilon_v(X,Y,Z)$ is the elliptic distance at pixel (X,Y,Z); P_v is the set of pixels of the image object v; $\#P_v$ is the total number of pixels contained in P_v

- *Main Direction (2D, 3D)*

  Main direction is defined as the direction of the eigenvector belonging to the larger of the two eigenvalues derived from the covariance matrix of the spatial distribution of the image object.

  The main direction feature of a three-dimensional image object is computed as follows:

- For each image object slice (a 2D pieces of the image object in a slice) the centers of gravity are calculated.
- The coordinates of all centers of gravity are used to calculate a line of best fit, according to the Weighted Least Square method.
- The angle α between the resulting line of best fit and the z-axis is returned as feature value.

$$Main\ Direction = \frac{180^\circ}{\pi} \tan^{-1}(VarX, \lambda_1 - VarY) + 90^\circ$$

where, VarX and VarY are the variance of X and Y respectively; λ_1 is the eigenvalue.

- *Radius of Largest Enclosed Ellipse (2D, 3D)*

  This feature describes how much the shape of an image object is similar to an ellipse/ellipsoid (2D/3D). The calculation is based on an ellipse/ellipsoid (2D/3D) with the same area/volume (2D/3D) as the object and based on the covariance matrix. This ellipse/ellipsoid (2D/3D) is scaled down until it is totally enclosed by the image object. The ratio of the axis length of the largest enclosed ellipse/ellipsoid (2D/3D) to the axis length of the original ellipse/ellipsoid (2D/3D) is returned as the feature value. Calculations in 2D and 3D are the following,

  - 2D

$$\epsilon_v(X_0, Y_0) = \min \epsilon_v(X, Y), (X, Y) \notin P_v$$

  - 3D

$$\epsilon_v(X_0, Y_0, Z_0) = \min \epsilon_v(X, Y, Z), (X, Y, Z) \notin P_v$$

  where, ϵ_v is the elliptic distance; P_v is the set of pixels of the image object v.

- *Radius of Smallest Enclosing Ellipse (2D, 3D)*

  The calculation for this feature is based on an ellipse/ellipsoid (2D/3D) with the same area/volume (2D/3D) as the image object and based on the covariance matrix. This ellipse/ellipsoid (2D/3D) is enlarged until it entirely encloses the image object. The ratio of the axis length of this smallest enclosing ellipse/ellipsoid (2D/3D) to the axis length of the original ellipse/ellipsoid (2D/3D) is returned as feature value.

  – 2D

  $$\epsilon_v(X_0, Y_0) = \max \epsilon_v(X, Y), (X, Y) \in \sigma P_v$$

  – 3D

  $$\epsilon_v(X_0, Y_0, Z_0) = \max \epsilon_v(X, Y, Z), (X, Y, Z) \in \sigma P_v$$

  where, ϵ_v is the elliptic distance; P_v is the set of pixels of the image object v.

- *Rectangular Fit (2D, 3D)*

  This feature describes how well an image object fits into a rectangle/cuboid (2D/3D) of similar size and proportions. While 0 indicates no fit, 1 indicates a complete fitting image object. The calculation is based on a rectangle/cuboid (2D/3D) with the same area/volume (2D/3D) as the considered image object. The proportions of the rectangle/cuboid (2D/3D) are equal to the proportions of the length-to-width/length-to-width-to-thickness (2D/3D) of the image object. The area/volume (2D/3D) of the image object outside the rectangle/cuboid (2D/3D) is compared with the area/volume (2D/3D) inside the rectangle/cuboid (2D/3D) that is not filled out with the image object.

  – 2D

  $$Rectangular\ Fit = \frac{\#\{(X,Y) \in P_v : \rho_v(X,Y) \leq 1\}}{\#P_v}$$

– 3D

$$Rectangular\ Fit = \frac{\#\{(X,Y,Z) \in P_v : \rho_v(X,Y,Z) \leq 1\}}{\#P_v}$$

where, ρ_v is the elliptic distance; P_v is the set of pixels of the image object v.

- *Roundness (2D, 3D)*

  Roundness quantifies how much the shape of an image object is similar to an ellipse/ellipsoid (2D/3D). The more the shape of an image object is similar to an ellipse/ellipsoid (2D/3D), the lower its roundness. It is calculated by the difference of the enclosing ellipse/ellipsoid (2D/3D) and the enclosed ellipse/ellipsoid (2D/3D). The axis length of the largest enclosed ellipse/ellipsoid (2D/3D) is subtracted from the axis length of the smallest enclosing ellipse/ellipsoid (2D/3D).

$$Roundness = \epsilon_v^{max} - \epsilon_v^{min}$$

  where, ϵ_v^{max} is the axis length of the smallest enclosing ellipse/ellipsoid (2D/3D); ϵ_v^{min} is the axis length of the largest enclosed ellipse/ellipsoid (2D/3D).

- *Sphericity (3D)*

  The sphericity feature is used to quantify how spherical a tumor object is. This feature is useful in describing overall tumor geometry. The expression for sphericity is given as,

$$Sphericity = \frac{\sqrt[3]{\pi(6V_v^2)}}{A_v}$$

  where, V_v is the volume; A_v is the total surface area of the image object.

- *Number of Macrospiculations (3D)*

  This feature provides the number of countable spiculations of the tumor [18].

- *Distance of Center of Gravity to Border of Tumor (3D)*

  This feature set provides a measure for the distance from the center of gravity to the border of the tumor. It is reported in terms of Average, Standard Deviation, Minimum and Maximum.

- *Attachment of Tumor to other Anatomical Structures (3D)*

  This feature set provides information regarding the attachment of the tumor to other anatomical structures. It is reported in terms of Relative Border to Lung; Relative Border to Pleural Wall; Ratio of Free to Attached Surface Areas.

- *Fractional Anisotropy (3D)*

  This feature provides the measure for the Fractional Anisotropy of the long and the short axes of the tumor object.

$$Fractional\ Anisotropy = \sqrt{\frac{(l_v - w_v)^2 + (w_v - t_v)^2 + (t_v - l_v)^2}{l_v^2 + w_v^2 + t_v^2}} * \sqrt{\frac{1}{2}}$$

where, l_v, w_v, t_v are respectively, the length, width and thickness of the tumor object.

2.2.2 Morphological Features

The morphological features are all 2D features. This set of features help in identifying vital image characteristics by accounting for the form and structure of the image object.

- *Margin Gradient(2D)*

  Margin gradient measures the attenuation value from the centroid of the image object to the background. It fits data to extract the gradient at the edge of the tumor (in HU/mm). Then, every 1 degree from spokes emanating from the centroid is measured and the slope is obtained for each degree. From these data, dimensionality is further reduced to the mean and standard deviation.

- *Fractal Dimension(2D)*

  Fractal dimension is a useful measure of the morphology of complex patterns that are seen in nature [19]. Fractal geometry is a way to quantify natural objects, with a complex irregular structure, by regular Euclidean geometrical methods. The study of fractals has been extended to biological structures in the past, such as in the study of human retinal vessels [20] and cell structure [21]. The Fractal dimension was applied to quantitatively characterize the complexity of the 2D boundary of the tumor.

  Fractal dimension (FD) was measured as follows [8]. First, box counting method was employed to the image objects. For a given length d, the number of boxes in the grid required to cover the image object boundary was measured as $N(d)$. For fractal objects, $N(d)$ is proportional to d^{-FD} as,

  $$N(d) = \mu d^{-FD}$$

  where μ is a constant. Fractal dimension (FD) is then measured as the slope of the regression line generated by plotting $ln(d)$ and $lnN(d)$, using the least-squares method. The expression for fractal dimension is given as,

  $$FD = \frac{ln(\mu) - lnN(d)}{ln(d)}$$

- *Fourier Descriptor(2D)*

  A Fourier descriptor (FD) is another widely used shape descriptor [22]. The FD is obtained by applying the Fourier transform on a shape contour, where each contour pixel is represented by a complex number. On applying the Fourier transform, the tumor region is broken up into four equally spaced annular regions in the Fourier domain. For each region, the second moment or energy is measured. Five Fourier descriptors were developed.

  $$FD_{Energy} = \sum_{i,j} I^2(i,j)$$

  where, i,j are within a specific annular region.

2.2.3 Texture Features

The next set of image features used were texture features. Texture features are available in both 2D and 3D domain. Texture features provide essential information about the internal structure of an image object and have been previously shown to be useful in the analysis of CT-scan images of Non-small Cell Lung Cancer as well [2].

- *Co-occurrence Matrices (2D, 3D)*

  The co-occurrence matrix [23] is a matrix that contains the frequency of one gray level intensity appearing in a specified spatial linear relationship with another gray level intensity within a certain range. Computation of features requires first constructing the co-occurrence matrix, then different measurements [24] can be calculated based on the matrix. The measurements include: contrast, energy, homogeneity, entropy, mean and maximum probability.

  - Contrast

    $$CM_{Contrast} = \sum_{i,j} |i-j|^2 * p(i,j)$$

  - Energy

    $$CM_{Energy} = \sum_{i,j} p(i,j) * p(i,j)$$

  - Homogeneity

    $$CM_{Homogeneity} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|}$$

  - Entropy

    $$CM_{Entropy} = -\sum_{i,j} p(i,j) * \log(p(i,j))$$

  - Sum Mean

    $$CM_{Mean} = \sum_{i,j} (i+j) * p(i,j)$$

– Maximum Probability

$$CM_{MaxProb} = \max(p(i,j))$$

where $p(i,j)$ is the element of the co-occurrence matrix and i,j are the gray level intensities.

- *Run-length Analysis (2D, 3D)*

  The run-length texture features [25] are created from runs of similar gray values in an image. Runs may be labeled according to their length, gray value, and direction (either horizontal or vertical). Long runs of the same gray value correspond to coarser textures, whereas shorter runs correspond to finer textures. In this study, texture information was quantified through the computation of 11 features [26] derived from the run-length distribution matrix. The measurement of Run-length analysis is conducted as follows,

  $p(i,j)$ is the element of run-length matrix, let M be the number of gray levels, N be the maximum run length, n_r is the total number of runs, n_p is the number of pixels in the image. Three new matrices are first defined,

$$p_p(i,j) = p(i,j) * j$$
$$p_g(i) = \sum_{j=1}^{N} p(i,j)$$
$$p_r(j) = \sum_{i=1}^{M} p(i,j).$$

  The generated features are,

  – Short Run Emphasis (SRE)

$$SRE = \frac{1}{n_r} \sum_{j=1}^{N} \frac{p_r(j)}{j^2}$$

  – Long Run Emphasis (LRE)

$$LRE = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j) * j^2$$

– Gray-Level Non-uniformity (GLN)

$$GLN = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i)^2$$

– Run Length Non-uniformity (RLN)

$$RLN = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j)^2$$

– Run Percentage (RP)

$$RP = \frac{n_r}{n_p}$$

– Low Gray-Level Run Emphasis (LGRE)

$$LGRE = \frac{1}{n_r} \sum_{i=1}^{M} \frac{p_g(i)}{i^2}$$

– High Gray-Level Run Emphasis (HGRE)

$$HGRE = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i) * i^2$$

– Short Run Low Gray-Level Emphasis (SRLGE)

$$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 * j^2}$$

– Short Run High Gray-Level Emphasis (SRHGE)

$$SRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) * i^2}{j^2}$$

– Long Run Low Gray-Level Emphasis (LRLGE)

$$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) * j^2}{i^2}$$

– Long Run High Gray-Level Emphasis (LRHGE)

$$LGHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) * j^2 * i^2$$

Volume features are often calculated as a series of 2D images and 2D texture features are usually computed for pixels in slices. However this kind of processing will result in losing information across slices. As in [27], co-occurrence matrices and run-length analysis features can be obtained in 3D, the features are calculated in 13 different directions, with each direction processing is done by plane instead of slice. Hence, information between slices is not ignored.

- *Laws Features*

  The Laws features [28] were constructed from a set of five one-dimensional (1D) filters, each designed to detect a different type of structure in the image. These 1D filters are defined as E5 (edges), S5 (spots), R5 (ripples), W5 (waves), and L5 (low pass, or average gray value). By using these 1D convolution filters, 2D filters are generated by convolving pairs of these filters, such as L5L5, E5L5, S5L5, W5L5, R5L5, etc. Thus a total of 25 different 2D filters can be generated. 3D Laws filters were constructed similarly by convolving 3 types of 1D filter, such as L5L5L5, L5L5E5, L5L5S5, L5L5R5, L5L5W5, etc. The total number of 3D filters is 125. After the convolution with their respective filters, the energy [29] of the texture feature was computed by the following equation:

  – 2D

  $$Laws^{2D}_{Energy} = \frac{1}{R} \sum_{i=N+1}^{I-N} \sum_{j=N+1}^{J-N} h^2(i,j)$$

  – 3D

  $$Laws^{3D}_{Energy} = \frac{1}{R} \sum_{i=N+1}^{I-N} \sum_{j=N+1}^{J-N} \sum_{k=N+1}^{K-N} h^2(i,j,k)$$

  where R is a normalizing factor, I, J and K are image dimensions in the 3D space and h(i,j,k) is derived from the convolution filters and original image.

- *Wavelet Decomposition*

  A wavelet transform decomposes an image into several components iteratively [30] based on the frequency, content and orientation. The discrete wavelet transform can iteratively decompose an image into four components. Each iteration splits the image both horizontally and vertically into low-frequency (low pass) and high-frequency (high pass) components. Thus, four components are generated: a high-pass/high-pass component consisting of mostly diagonal structure, a high-pass/low-pass component consisting mostly of vertical structures, a low-pass/high-pass component consisting mostly of horizontal structure, and a low-pass/low-pass component that represents a blurred version of the original image. Subsequent iterations then repeat the decomposition on the low-pass/low-pass component from the previous iteration. These subsequent iterations highlight broader diagonal, vertical, and horizontal textures. And for each component, we calculated the energy feature. For each iteration, the wavelet decomposition of a 2D image can be achieved by applying the 1D wavelet decomposition along the rows and columns of the image separately, while for 3D, 1D wavelet transform is applied along all the three directions (x,y,z).

  – 2D

$$WD^{2D}_{Energy} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} I^2(i,j)$$

$$WD^{2D}_{Entropy} = \frac{-1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{I^2(i,j)}{norm^2} \log(\frac{I^2(i,j)}{norm^2})$$

  – 3D

$$WD^{3D}_{Energy} = \frac{1}{M \times N \times L} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{L} I^2(i,j,k)$$

$$WD^{3D}_{Entropy} = \frac{-1}{M \times N \times L} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{L} \frac{I^2(i,j,k)}{norm^2} \log(\frac{I^2(i,j,k)}{norm^2})$$

  where, I(j,k) and I(j,k,l) are the sub-block elements for 2D and 3D respectively. The dimensions of each sub-block are represented by M, N and L as per 2D/3D. The

number of features is dependent on the number of decomposition levels selected. A single level generated 16 features while, 2 levels yielded 30 features.

2.2.4 Intensity Based Features/Histogram Features

The histogram features take into account the intensities of the pixels forming the image. The intensity histogram h(a), of an image object, represents the number of pixels for brightness level "a" plotted against their brightness level. The probability distribution of the brightness P(a) is also calculated. Six histogram features were developed. The expression for each are briefly shown here.

- Mean

$$HF_{Mean} = \sum_{i=1}^{range} i * P(i)$$

- Standard Deviation

$$HF_{SD} = \sqrt{\sum_{i=1}^{range} (i - HF_{mean})^2 * P(i)}$$

- Skewness

$$HF_{Skew} = \frac{\sum_{i=1}^{range} (i-HF_{mean})^3 * P(i)}{(\sum_{i=1}^{range} (i-HF_{mean})^2 * P(i))^{1.5}}$$

- Kurtosis

$$HF_{Kurt} = \frac{\sum_{i=1}^{range} (i-HF_{mean})^4 * P(i)}{(\sum_{i=1}^{range} (i-HF_{mean})^2 * P(i))^2}$$

- Energy

$$HF_{Energy} = \sum_{i=1}^{range} P(i) * P(i)$$

- Entropy

$$HF_{Entropy} = -\sum_{i=1}^{range} P(i) * \log(P(i))$$

where, range indicates range of intensity (normalized), $P(i) = \frac{h(i)}{\sum h(i)}$, h(i) is the frequency of intensity i.

2.3 Clinical Features

In addition to image features, two clinical features were used,

- Gender
- Tumor Location

The clinical features used, were obtained from the histopathological report for the patients.

- *Gender*

  The gender information was simply represented by feature values 1 for female and -1 for male.

- *Tumor Location*

  Determination of tumor location was done based on the clinical report rather than on observation. The location information in the clinical report was represented using the $3rd$ edition of the International Classification of Diseases for Oncology (ICD-O-3) codes. The location of the tumor was split into the following categories:

  - Lung Upper Lobe (C341)
  - Lung Middle Lobe (C342)
  - Lung Lower Lobe (C343)
  - Lung Overlapping Lesion (C348)
  - Lung NOS (C349).

  NOS indicates not otherwise stated. Since the SVM classifier considers the distance measure for each feature, it was not possible to represent location by a single attribute with multiple values. Hence 5 attributes were created for representing the location information as represented below,

$$
\begin{array}{c}
 \\
\textit{upper lobe} \\
\textit{middle lobe} \\
\textit{lower lobe} \\
\textit{overlap} \\
\textit{NOS}
\end{array}
\begin{array}{c}
\begin{array}{ccccc} C341 & C342 & C343 & C348 & C349 \end{array} \\
\left[\begin{array}{ccccc}
1 & -1 & -1 & -1 & -1 \\
-1 & 1 & -1 & -1 & -1 \\
-1 & -1 & 1 & -1 & -1 \\
-1 & -1 & -1 & 1 & -1 \\
-1 & -1 & -1 & -1 & 1
\end{array}\right]
\end{array}
$$

This chapter thus presented the vast array of image features that were successfully extracted from the CT images of the lung. The clinical features provided a brief glimpse of the possibilities of expanding the feature space using histopathological reports. The scope and use of these features in tumor analysis are described in Chapter 4.

CHAPTER 3

PREDICTION MODELS

In this chapter a general description of the classification algorithms and the feature selection techniques used for developing predictive models is presented. The classifier models used in the study are presented in Section 3.1. Section 3.2 describes the feature selection techniques. Tuning of the parameters of SVM is described in Section 3.3.

3.1 Classifier Models

Classifier models evaluated here are decision trees [31], random forests [32], nearest neighbor [33], support vector machines (SVM) [34] and naive Bayes [35].

3.1.1 Decision Tree

The decision tree classifier [31] model consists of a structure that is either a leaf, indicating a class or a decision node, that specifies some test to be carried out on a single attribute value, with one branch and subtree for each possible outcome of the test. The performance measure used in these tests is gain-ratio. In a decision tree model a case is classified by starting at the root of the tree and then traversing through it until a leaf is encountered. At each of the non-leaf decision nodes, the outcome of the test at the node is determined and it in turn becomes the root of the subtree that corresponds to this outcome. This continues until a leaf node is reached and the class of the instance is predicted to be the class label associated with the leaf. The decision tree used in the survival analysis study, described in Section 4.5, was the C4.5 library, with release 8 patches, developed by J.R. Quinlan [36]. The decision tree used for the tumor type classification experiments described in Section

4.2 was J48, a Java implementation of C4.5. The parameter, confidence factor, was set to 0.25.

3.1.2 Random Forests

Random forests [32] consist of an ensemble of decision trees. In this method, the training data is bagged a specified number of times and then for each training set a decision tree is built. At every decision tree node, a random set of attributes are chosen and the best among them are used as a test. In this work, the forest contained 200 decision trees and randomly chose $log_2(n) + 1$ features from n total features at each node. The class predicted comes from a vote of the trees. The implementation of random forests used here is part of the Weka-3.6 data mining tool [37].

3.1.3 Nearest Neighbor

The nearest neighbor algorithm [33] used was the IB-k Weka implementation. It is a modified version of K nearest neighbors. The nearest neighbor search can be done employing brute force linear search or by using other data structures such as KD-trees. In this work the simple linear search algorithm was used. The distance metric was the Euclidean distance. Since the scale of the attributes determines the distance measure, attributes with larger ranges would dominate. Hence, the weka implementation of the algorithm performs normalization on the attributes before measuring the distance. The number of nearest neighbors chosen for this particular set of experiments was 5.

3.1.4 Support Vector Machine

Support vector machines are based on statistical learning theory [38] and have been shown to obtain high accuracy on a diverse range of application domains [39]. The idea behind SVMs is to non-linearly map the input data to a higher dimensional feature space and construct a hyper plane so as to maximize the margin between classes. In this feature space a linear decision surface is constructed. The hyper plane construction can be reduced

to a quadratic optimization problem which is determined by subsets of training patterns that lie on the margin, termed support vectors [40]. Special properties of the decision surface ensure high generalization ability of the learning machine. The hyper plane in the input space is in the form of a decision surface, the shape of which is determined by the chosen kernel. Figure 3.1 illustrates the choice of support vectors and the generation of the hyper plane to separate class boundaries in a support vector machines. Different kernels can be chosen for SVMs, such as, Linear Kernels, Radial Basis Function Kernel and Sigmoid Kernel. The *Radial Basis Function* (RBF) Kernel was used here. The expression for the RBF kernel is given by $exp(-\gamma|u-v|^2)$. All data was scaled to be in the range $[-1, 1]$. Support vector machines have previously been used efficiently on CT scan image data of the lungs [41] in a Computer-Assisted Detection (CAD) system for automated pulmonary nodules detection in thoracic CT-scan images. For the support vector machine, libSVM [42] was used.

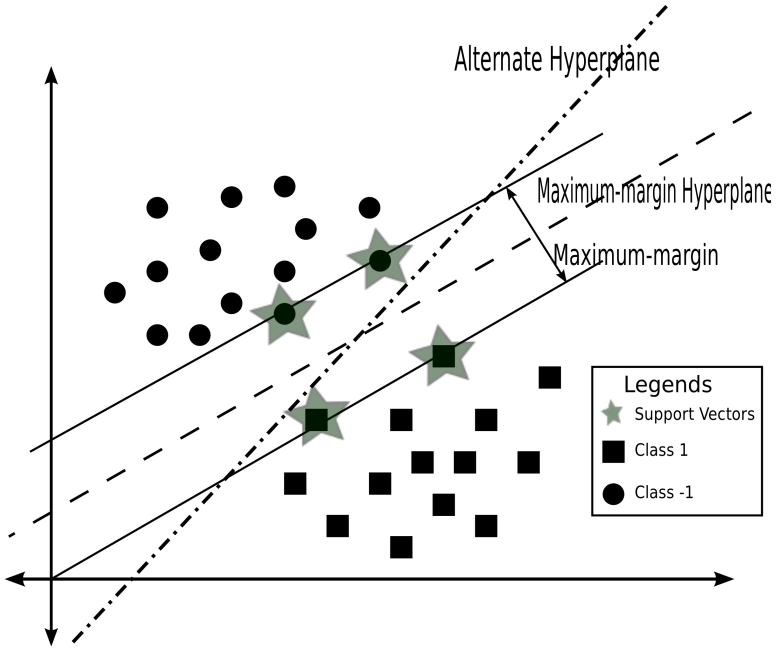


Figure 3.1: An Example of the Use of a Support Vector Machine for a Separable Problem in a 2D Space. The Support Vectors Define the Margin of Largest Separation Between the Two classes. (Based on figure by Vapnik and Cortes [40] pp. 275.)

3.1.5 Naive Bayes

The naive Bayes [35] classifier is designed to be used when features are independent of one another within each class. However, it has been shown in practice that it works well even when the independence assumption is not valid. The naive Bayes classifier estimates the parameters of a probability distribution, assuming features are conditionally independent given the class using the training samples. It then computes the posterior probability of a sample belonging to each class and classifies the test sample according the largest posterior probability.

The class-conditional independence assumption greatly simplifies the training step since it allows for an individual estimate of the one-dimensional class-conditional density for each feature. Even though, the class-conditional independence between features does not hold for most data sets, this optimistic assumption works well in practice. The assumption of class independence allows the naive Bayes classifier to better estimate the parameters required for accurate classification while using less training data. The implementation for naive Bayes used for this work is from the Matlab Statistical Toolbox [43].

3.2 Feature Selection Methods

Given the large parameter space for both 2D and 3D features and limited number of examples, it was necessary to perform feature selection. relief-F and wrapper methods were used for tumor type classification, while feature correlation based feature selection and principal component analysis were used in survival analysis study.

3.2.1 Relief-F

Relief-F [44], which stands for Recursive Elimination of Features, chooses instances at random and changes the weights of the feature relevance based on the nearest neighbor. The ranker search algorithm, used along with relief-F, assigns a rank to each individual feature. The number of features to be chosen for evaluation can be done by means of a parameter. For the tumor type classification study, 50 features were chosen at first and

then the feature space was further reduced to 25 features. The choice of the feature space was based on multiple trials to effect substantial changes in classification results.

3.2.2 Wrappers

The second feature selection technique employed was wrapper feature selection [45]. This involves evaluating attribute subsets with an underlying classifier model. In this case, the underlying model was the same as the classifier being evaluated. That is, if the classifier in use was a support vector machine with an RBF kernel, the underlying classifier for wrappers was also a support vector machine with an RBF kernel. Feature selection using wrappers was done by using best-first forward selection search. Wrapper feature selection was not used for random forests classifiers. The reason being, both approaches involve selecting of subset of features for evaluation.

3.2.3 Feature Selection Based on Correlation

In his study of Test-Retest analysis on an independent, publicly available Reference Image Database to Evaluate Response [1](RIDER) data set, Kumar et al. [46] evaluated the reproducibility of the 3D image features we had developed. For imaging biomarkers to be effective in prognostication, prediction or therapy response studies, standardization and optimization of the feature space is necessary. One of the key steps in the qualification process of potential biomarkers is to assess the intra-patient (test-retest) reproducibility and biological ranges of these features. The importance of highly reproducible features is that they are potentially the most informative. The more reproducible a feature, the higher its ability to identify subtle changes with time, pathophysiology and response to therapy. The RIDER data set consisted of 32 patients. The baseline scan was followed up, within 15 minutes, by a follow-up scan, acquired using the identical CT scanner and imaging protocol.

The reproducibility measure used was the concordance correlation coefficient *(CCC)* and dynamic range *(DR)*. The concordance correlation coefficient *(CCC)* was measured for each

[1]RIDER. Available from: https://wiki.nci.nih.gov/display/CIP/RIDER

feature by comparing the feature values extracted for the two sets of scans. Dynamic range *(DR)* was determined by comparison of the reproducibility to the entire biological range available for a feature. The greater the dynamic range, the more useful is a feature. The study determined that the reproducibility criteria of CCC > 0.85 and DR > 100 produced useful features and that 33 of the 215 3D image features met the aforementioned thresholds. Only this reduced feature set of 33 stable image features was used in the survival analysis study described in Section 4.5.

One of the results of the Test-Retest study was the measure of correlation amongst features. For the subset of 33 stable image features the meeting the correlation criteria, Pearson's correlation between the features of the subset was measured to generate correlation matrices. In order to further reduce the number of features, it can be stated intuitively that choosing a single representative feature from a group of highly correlated features should reduce the redundancy. The goal is to identify a set of uncorrelated features that can be used to analyze lung tumors. This is achieved through the setting of a threshold for Pearson's correlation measure. Features which correlate together (have Pearson correlation values greater than a threshold) form a group. For this work, the threshold was set to 0.8. The threshold was chosen so that the groups formed were neither too large nor too small. Each sub-group formed consisted of features that had a higher correlation measure with others in the group than the threshold amongst each other. The grouping was done in such a way that if a feature appeared in two groups, it was removed from the group with a smaller number of features.

The features with very high correlation could be represented by a single feature, significantly reducing the number of features needed. To perform feature selection based on correlation measure, the features under consideration were ranked based on the relief-F algorithm [44] making use of the training data. The ranker algorithm assigns a rank to each individual feature. Now, based on the groups generated, the highest ranked feature from each group is chosen for analysis on test data. The workflow can be represented using the schematic shown in Figure 3.2.

3.2.4 Principal Component Analysis

Principal component analysis *(PCA)* is a widely used multivariate statistical technique. PCA is useful in analyzing data tables representing observations represented by dependent and generally inter-correlated variables. The goal of PCA is to extract the important information for the data table and express it as a set of new orthogonal variables which are termed *principal components* obtained as linear combinations of the original variables. The first principal component essentially has the largest possible variance. The second component is restricted to be orthogonal to the first principal component and have the largest possible variance of the remaining set. Other principal components are computed similarly. The values for the new variables for the observation are obtained through projections of the observations onto the principal components.

3.3 Parameter Tuning

The performance of classifiers is often governed by the choice of parameters. In this study, parameter tuning has been restricted to the SVM classifier, since the parameters used for the other classifier models were often standard with not much room for variation. The parameter tuning method used in this study is grid search.

3.3.1 Grid Search

For tuning the parameters of the SVM classifiers, grid search was used. The parameters tuned were the *cost* (c) and *gamma* (γ) parameters of the radial basis function (RBF) kernel. The cost parameter was evaluated in the range from -5 to 15, while γ was evaluated over the range of 10^{-15} to 10^{3}. The optimization of the parameters was conducted by running a 5 fold cross validation over each training fold and optimized for the highest AUC. The AUC reported is the average of the highest AUC for each fold. Thus cost and γ parameters are optimized over each individual fold and no specific optimal parameter for the entire model can be reported.

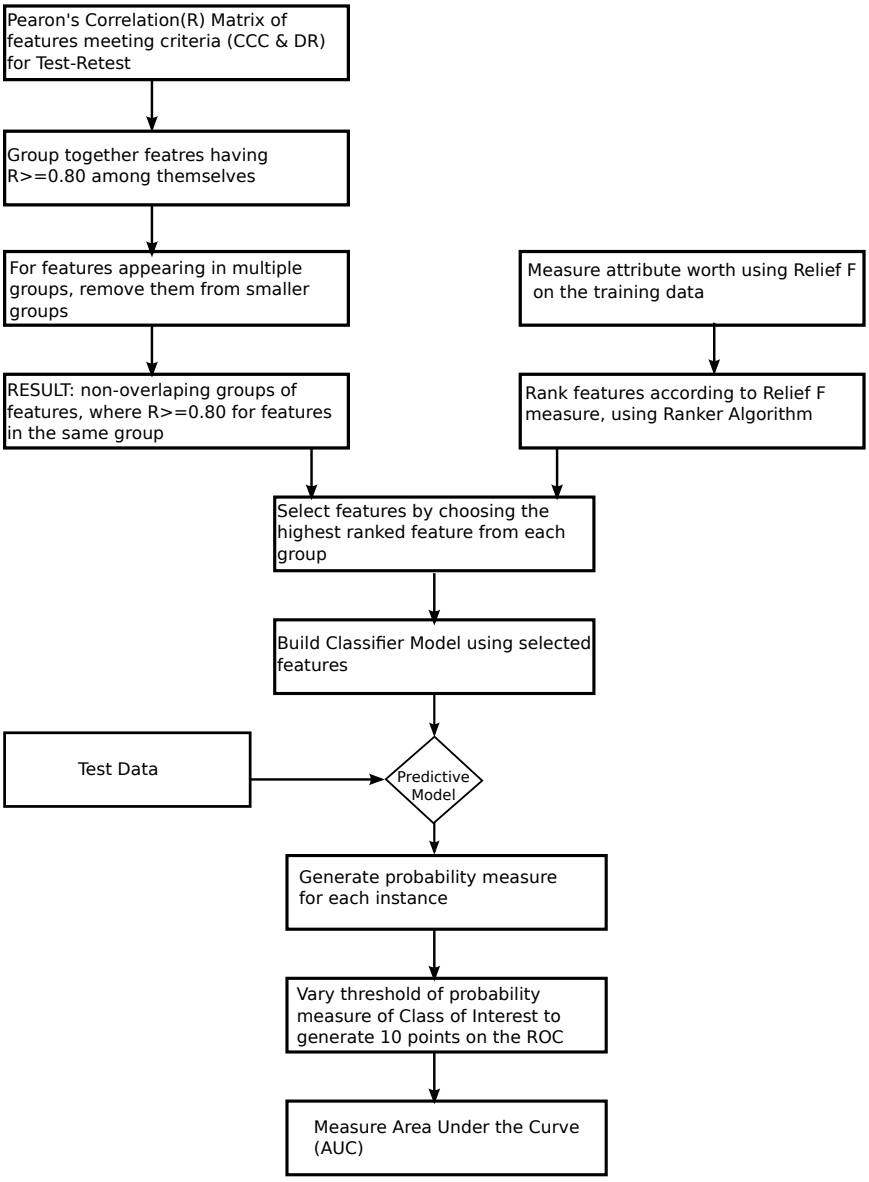


Figure 3.2: Schematic Representation of Feature Selection Using Feature Correlation

CHAPTER 4

EXPERIMENTS AND RESULTS

In this chapter, the implementation and results for the experiments conducted are presented. First, the data sets used for the experiments are discussed. The class distribution and basic filtering of the data set for different experiments are presented in 4.1. Section 4.2 presents a basic overview of the method used to generate predictive models for tumor type classification. Results and analysis of the two-class classification problem are presented in Section 4.3 followed by the results of introducing a third tumor class and analyzing it as a 2 class problem in Section 4.4. Section 4.5 covers the results of 2 year survival time prediction.

4.1 Data Set

The data used here are CT-scan images from the H. Lee Moffitt Cancer Center and Research Institute, Tampa. The images are in the DICOM (Digital Imaging and Communications in Medicine) format. The slice thickness of the acquired CT-images ranged from 3mm to 6mm. However, the pixel spacing amongst the scanned images varied greatly, as illustrated by Figure 4.1.

CT-scans of 109 patients were used for this study. The distribution of the tumor type is given as,

- Adenocarcinoma = 38
- Squamous-cell Carcinoma = 36
- Bronchioalveolar Carcinoma = 35

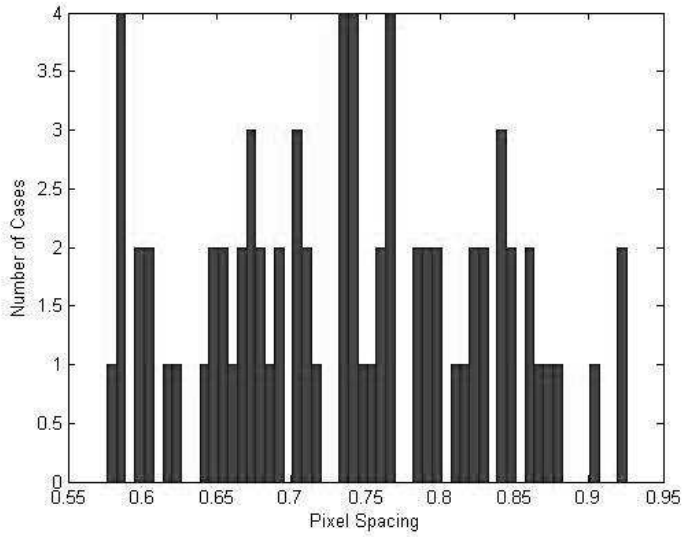


Figure 4.1: Representation of the Variability in Pixel-Spacing over 109 CT-Scan Images

In each of the cases only one tumor was present in the lung. All patient identification information had been removed.

For the first set of experiments concerning classification of tumor types into Adenocarcinoma and Squamous-cell Carcinoma, as per the class distribution of available data, CT-scan images of 74 patients were used for analysis. CT-scans of all 109 patients were utilized when the analysis was classifying Bronchioalveolar Carcinoma from the other tumor types.

For survival time analysis, the data-set consisted of patients across three tumor tumor types, Adenocarcinoma, Bronchioalveolar Carcinoma, Squamous-cell Carcinoma. However for two year survival analysis, patients alive but with survival time less than 24 months had to be removed. That reduced the size of the dataset to 95. The class distribution of survival time in months is as follows,

- patients with survival time ≥ 24 [Class 1] = 63
- patients with survival time < 24 [Class -1] = 32

4.2 Tumor Type Classification Outline

The basic schematic of tumor classification is that, once the tumor objects have been identified and segmented out in the CT-scan images, rulesets are run to extract the required feature values from the tumor objects. In the case of the 2-class problem, the features extracted were both 2D and 3D features. While for the remaining two experiments, only the 3D features were used. These feature values formed the mineable data upon which the classifier models were built. For the tumor classification problem, the classifiers used were, J-48, random forests, IB5 and SVM. These classifiers were first run without any feature selection being applied. Then these classifiers were run along with relief-F feature selection algorithm which reduced the feature space to 50 and 25 features. The final feature selection method used was wrapper. These feature selection algorithms were used in combination with classifier algorithms to generate predictive models. The results and further discussion about the two classification problems are given in the next section.

4.3 Tumor Classification for Adenocarcinoma and Squamous-cell Carcinoma

In this section, details of the experiment concerning classification of Adenocarcinoma and Squamous-cell Carcinoma using 2D and 3D image features is presented.

4.3.1 Experimental Outline

The feature values acquired through feature extraction formed the base set from which the tumors were classified. For the 2D slices, a certain amount of filtering was done before the data was evaluated. For our study as stated earlier, only a single tumor volume was identified for each patient. Thus a segmented tumor volume was represented by 2D tumor objects over the sequence of slices containing the tumor. Observing the segmentation of the tumor objects we found multiple very small tumor objects, in terms of pixels. These mainly consisted of either tumor fragments or the objects identified in slices marking the beginning of a sequence containing the tumor. Many feature extraction algorithms fail for such tumor

objects. Hence, during feature extraction a threshold value of 30 pixels was set for tumor objects. This threshold was identified based on observation. From the 74 volumes a total of 710 (Adenocarcinoma: 347; Squamous-cell Carcinoma: 363) tumor objects from 672 (Adenocarcinoma: 324; Squamous-cell Carcinoma: 348) slices were identified and features were extracted from them. Still, there were a few tumor objects for which one or many of the feature extraction algorithms failed. It was observed that the size in terms of area (measured as number of pixels in the object) for which the feature extraction process failed, varied from one volume to another. Hence a definitive area threshold eludes us at the moment.

However, it was observed that leaving out those objects for which feature extraction was not possible did not result in a choice of random slices within a tumor, but instead resulted in, a contiguous sequence of slices containing tumor objects in each of the volumes. Thus successful feature extraction was done on the dataset as follows: for 74 volumes, 675 tumor objects (Adenocarcinoma: 323; Squamous-cell Carcinoma: 352) present in 592 (Adenocarcinoma: 260; Squamous-cell Carcinoma: 332) slices. More filtering of the data was done to select only one tumor object for each slice of the CT-image sequence. This filtering was done by choosing the largest tumor object for each slice in terms of area in pixels. By keeping only the largest tumor object, when there are multiple objects in a single slice, the number of tumor objects under consideration was reduced to 592 (Adenocarcinoma: 260; Squamous-cell Carcinoma: 332) for 74 volumes. Given the small number of volumes, the evaluation of the classifier was done by performing a *'leave-one-volume-out'* experiment. We also performed a 10-fold cross validation experiment for two reasons. It can give a reasonable estimate of accuracy on unseen data when there is enough data and it avoids potential pitfalls of leave one out.

There is a complication for 2D features. Each instance in a volume represents one tumor object which will be the largest tumor object in a slice (rarely there will be fragments). The prediction for the entire volume is more important than that for each individual tumor object. The leave-one-volume-out experiment involves training the classifier model on ex-

amples from all but one volume and testing on the remaining volume. Since each instance in the test set represents a tumor object for the given volume, class prediction for the entire volume is done by means of voting, where class prediction for the majority of objects is taken to be the tumor class. In case of ties, tie breaking was explored using two methods. First by choosing the class of the tumor object on the $(\frac{n}{2}+1)^{th}$ slice in the sequence, where n is the number of slices having a tumor object (i.e. the middle slice). A second method of tie-breaking involved choosing the class of the largest tumor object for the given volume. This mode of tie-breaking was employed since it is reasonable to believe that feature extraction from the largest tumor object for a given volume would contain the maximum possible information. Accuracy by volume for both tie-breaking techniques will be shown. For each of 74 folds, the training set consisted of 73 tumor volumes and the trained model was evaluated on the one remaining volume. For 10-fold cross validation on 2D features, using the same distribution of volumes in each fold as for 3D features, individual volumes in a single fold were tested against data from the remaining 9 folds. The average performance for each fold is reported.

4.3.2 Feature Merit using F-test

It is of interest to determine whether there is statistically significant improvement if 3D features are selected over 2D features. This is done by means of performing an F-test on results obtained from a 5x2 fold cross validation. The effectiveness of combining 5x2 fold cross validation with the F-test has been used to compare supervised learning algorithms [47]. The goal here is to, using the same classifier models, compare classifier accuracy using 3D features with that of 2D features.

For comparison, 5x2 fold cross validation was first run for 3D features. Then making use of the volumes in each fold, 2D feature evaluation was done on the same 10 folds. For this evaluation only a single tie-breaking method, choosing the class of the largest tumor object, was employed for 2D features.

4.3.3 Results

A leave-one-volume-out *(LOVO)* experiment for 2D and 3D features using classifiers with the previously described settings was done with the results shown in Table 4.1 for 2D features and Table 4.2 for 3D features. The results of the 10-fold cross validation for 2D and 3D features are shown in Tables 4.3 and 4.4 respectively.

Table 4.1: Performance of Classifier Models on 2D Features Performing Leave-One-Volume-Out (The highest accuracy for each classifier is in bold)

Classifier	Feature Selection	Accuracy (slice)	Accuracy ($Volume^1$)	Accuracy ($Volume^2$)
J48	None	49.32	45.94	47.30
	Relief-F (50 features)	47.63	47.30	48.65
	Relief-F (25 features)	47.47	43.24	45.95
	Wrapper	56.92	56.76	**56.76**
Random Forests	None	49.32	**52.70**	50.00
	Relief-F (50 features)	49.48	51.35	48.64
	Relief-F (25 features)	47.30	44.59	47.29
IB5	None	49.59	47.30	48.65
	Relief-F (50 features)	50.82	51.35	55.41
	Relief-F (25 features)	56.09	**58.11**	56.76
	Wrapper	53.27	56.76	56.76
SVM	None	53.48	48.64	51.35
	Relief-F (50 features)	52.68	52.70	54.05
	Relief-F (25 features)	58.47	54.05	52.70
	Wrapper	53.29	59.46	***60.81***

In Table 4.1, the average accuracy over slices in the 74 volumes is shown in the first column. The next two columns show the percentage of volumes that were correctly identified. This is the result of voting among the tumor objects that constitute each tumor volume. The fourth column shows accuracy when the tie-breaking for voting is done by choosing the class of the $(\frac{n}{2}+1)^{th}$ slice in the sequence, where n is the number of slices having a tumor object. The fifth column represents accuracy when the method of tie-breaking was done choosing the class of the largest tumor object for the given volume. In the case of 2D features, the number objects of the two classes of tumor were, Adenocarcinoma: 260

[1]Breaking Tie by choosing the class of the middle slice
[2]Breaking Tie by choosing the class of the largest tumor object

and Squamous-cell Carcinoma: 332. However, the majority class distribution in terms of volume was 51.35%, with Adenocarcinoma being the majority class.

Table 4.2: Performance of Classifier Models for 3D features performing Leave-One-Volume-Out (The highest accuracy for each classifier is in bold)

Classifier	Feature Selection	Accuracy (L-O-O)
J48	None	***67.57***
	Relief-F (50 features)	60.81
	Relief-F (25 features)	48.65
	Wrapper (forward selection)	45.95
Random Forests	None	59.46
	Relief-F (50 features)	54.05
	Relief-F (25 features)	**60.81**
IB5	None	51.35
	Relief-F (50 features)	51.35
	Relief-F (25 features)	51.35
	Wrapper (forward selection)	**51.35**
SVM	None	52.70
	Relief-F (50 features)	56.76
	Relief-F (25 features)	56.76
	Wrapper (forward selection)	**66.22**

4.3.3.1 Leave One Out

In Table 4.1, when the classifiers were run without any feature selection employed, SVM had relatively higher accuracy than the rest of the models. We see that the most accurate classifier was SVM after wrappers feature selection at 60.81% accuracy. Feature selection usually improved accuracy for all but random forests. There is little difference in results from the two tie-breaking approaches.

In Table 4.2, results when performing leave one volume out with 3D features are shown. Without any feature selection being employed, J48 achieves an accuracy of 67.57%. With an accuracy of 66.22%, SVM has the next highest accuracy using wrapper feature selection. Feature selection improved the accuracy of SVM, but resulted in performance degradation for J48. Random forests, achieved its highest accuracy of 60.81% with relief-F feature selection using a subset of 25 features. IB5 achieves the same accuracy of 51.35% in each

case, even though the misclassification of the minority class was observed to improve with feature selection.

Table 4.3: Performance of Classifier Models on 2D Features Performing 10-Fold Cross Validation (The highest accuracy for each classifier is in bold)

Classifier	Feature Selection	Accuracy ($Volume^1$)	Accuracy ($Volume^2$)
J48	None	**52.70**	48.65
	Relief-F (50 features)	45.95	47.30
	Relief-F (25 features)	47.30	48.65
	Wrapper	47.30	49.32
Random Forests	None	49.32	48.65
	Relief-F (50 features)	60.81	***62.16***
	Relief-F (25 features)	55.41	56.76
IB5	None	54.05	56.76
	Relief-F (50 features)	49.32	52.70
	Relief-F (25 features)	56.76	56.76
	Wrapper	**59.46**	58.11
SVM	None	**56.76**	**56.76**
	Relief-F (50 features)	47.30	49.32
	Relief-F (25 features)	45.95	48.64
	Wrapper	49.32	48.64

4.3.3.2 10-Fold Cross Validation

The 10-fold cross validation result is shown for 2D features in Table 4.3 and for 3D features in Table 4.4. In Table 4.3, without feature selection being applied, SVM had the best result with an accuracy of 56.76%. Also, apart from IB5 all other classifier models had fairly low accuracy on reduction of the feature pool from 50 to 25 with relief-F feature selection. The highest accuracy of 62.16% was achieved with random forests with 50 features selected using relief-F feature selection.

In Table 4.4 the 3D 10-fold CV results are reasonably consistent with the evaluation for Leave One Out, with J-48 having the highest accuracy of 67.57%. Running classifiers without feature selection seems to yield better accuracy in case of J48 and IB5, while feature selection improves performance for random forest and SVM.

[1]Breaking Tie by choosing the class of the middle slice
[2]Breaking Tie by choosing the class of the largest tumor object

Using the relief-F feature selection method to reduce the feature set to 50 features reduced accuracy for J48 (around 6% less accuracy) but improved performance on SVM (around 7% greater accuracy). It had no effect on the performance of random forests, while for IB5 accuracy was vastly reduced (over 12% less accuracy) to 50%. When the feature set was further reduced to 25 features using relief-F some increase in accuracy was obtained for random forest, while all other classifiers had lower accuracy. Wrapper forward selection did not have much of an effect in improving accuracy of the classifiers, except for SVM, which achieved accuracy of 62.16% using this feature selection method.

Table 4.5 shows percentage accuracy from performing a 5x2 fold cross validation on classifiers using 2D features, with the difference in accuracy of 3D features given within parentheses. The results indicate that using 2D features provides higher average accuracy over 3D features except for four instances, that have been highlighted in the table. Table 4.6 shows the result of performing an F-test on results of the 5x2 cross validation on each classifier model, using 3D and 2D features respectively. An F-measure over 4.7351 indicates statistical significance over 95%. This level was not reached. In the cases where the 2D feature based classifiers had a higher F-value like J48 and wrappers, the overall accuracy is less than achieved with J48 and 3D features (albeit with no feature selection).

4.4 Tumor Classification involving Bronchioalveolar Carcinoma

The results from the experiments in Section 4.3.3 indicated that using 3D image features yielded performance on par if not better than 2D features all the while being much easier to use. Next, to extend the work, a tumor type different to Adenocarcinoma and Squamous-cell Carcinoma was introduced. Bronchioalveolar Carcinoma, though a sub class of Adenocarcinoma, differs significantly in terms of structure and malignancy compared to other Adenocarcinoma subtypes [48]. Hence, it can be considered a separate class for classification.

Table 4.4: Performance of Classifier Models for 3D Features Performing 10-Fold Cross Validation. (The highest accuracy for each classifier is in bold)

Classifier	Feature Selection	Accuracy (10-fold CV)
J48	None	***67.57***
	Relief-F (50 features)	60.81
	Relief-F (25 features)	52.70
	Wrapper (forward selection)	55.41
Random Forests	None	55.41
	Relief-F (50 features)	55.41
	Relief-F (25 features)	**58.11**
IB5	None	**62.16**
	Relief-F (50 features)	50.00
	Relief-F (25 features)	54.05
	Wrapper (forward selection)	45.96
SVM	None	55.41
	Relief-F (50 features)	60.81
	Relief-F (25 features)	58.11
	Wrapper (forward selection)	**62.16**

4.4.1 Evaluating 3-class Problem as 2-class Problem

In the previous section, Section 4.3.3, experiments on Adenocarcinoma and Squamous-cell Carcinoma yielded classifiers that could be improved for greater accuracy. It has been hypothesized that Bronchioalveolar Carcinoma (BAC) is substantially different from other classes of Non-small Cell Lung Cancer (NSCLC) classes. Hence a 2 class problem was created, where an attempt was made to tell BAC (35 volumes) apart from the other two classes (74 volumes). To treat this as a two-class problem, misclassification amongst Adenocarcinoma and Squamous-cell Carcinoma was ignored.

4.4.2 Experimental Outline

The experiment conducted was similar to that done with 2 tumor types. Since BAC is the tumor type of interest in this study and a significantly smaller class, apart from general accuracy, recall on BAC is also considered. Recall can be expressed as follows,

$$Recall = \frac{true\ positive}{true\ positive + true\ negative}$$

Table 4.5: Performance of Classifier Models Performing 5x2 Fold Cross Validation for 2D Features. Difference for 3D features given within parentheses. (The case where 3D features perform better has been highlighted)

Classifier	Feature Selection	Accuracy (10-fold CV)
J48	None	**53.11 (-6.35)**
	Relief-F (50 features)	**51.75 (-0.14)**
	Relief-F (25 features)	**52.57 (-0.13)**
	Wrapper (forward selection)	53.92 (+3.16)
Random Forests	None	52.84 (+3.92)
	Relief-F (50 features)	52.30 (+1.76)
	Relief-F (25 features)	52.57 (+0.95)
IB5	None	52.16 (+3.51)
	Relief-F (50 features)	51.75 (+4.45)
	Relief-F (25 features)	50.00 (+1.90)
	Wrapper (forward selection)	53.65 (+5.26)
SVM	None	**53.11 (-0.94)**
	Relief-F (50 features)	52.30 (+2.50)
	Relief-F (25 features)	49.73 (+0.81)
	Wrapper (forward selection)	50.00 (+1.61)

Table 4.6: F-test on 5x2 Cross Validation Results Between 2D Features and 3D Features.

Classifier	None	Relief-F (50 features)	Relief-F (25 features)	Wrapper
J-48	3.89	2.43	3.79	3.18
Random Forests	3.25	2.43	2.08	X
IB5	2.43	2.20	1.73	2.63
SVM	1.53	3.26	3.21	3.44

4.4.3 Results

The results of performing classification to identify BAC from the other two tumor classes is provided in this section. First we analyze the performance on classifiers in a 10-fold cross validation. Then in the next subsection, the effectiveness of introducing a correlation based feature selection procedure is evaluated.

4.4.3.1 10-Fold Cross Validation

Table 4.7, lists the results for the classifier models using the image features described earlier.

Table 4.7: Performance of Classifier Models Performing 10-Fold Cross Validation for the BAC Study (The highest accuracy and the highest recall % on Bronchioalveolar carcinoma for each classifier is in bold)

Classifier	Feature Selection	Accuracy (10-fold CV)	Recall On BAC
J48	None	69.37	54.00
	Relief-F (50 features)	65.57	54.00
	Relief-F (25 features)	69.37	51.00
	Wrapper	**70.27**	**63.00**
Random Forests	None	70.27	63.00
	Relief-F (50 features)	**74.77**	**71.00**
	Relief-F (25 features)	72.97	63.00
IB5	None	65.77	23.00
	Relief-F (50 features)	65.77	29.00
	Relief-F (25 features)	69.34	40.00
	Wrapper	**74.77**	**51.00**
SVM	None	71.17	66.00
	Relief-F (50 features)	***76.58***	66.00
	Relief-F (25 features)	71.17	63.00
	Wrapper	71.17	***74.00***

J-48 with wrapper feature selection yields the best result for decision tree. Random forests performs better for some feature selection criteria. IB5 produces relatively high accuracy, but very low recall on BAC. SVM without feature selection has highest accuracy, though when used along with wrapper has the highest recall measure for BAC.

4.4.3.2 Feature Selection using Concordance Correlation Coefficient

Given the large feature space it is important to find a suitable feature selection method. The relief-F and wrapper feature selection techniques have been shown to allow for some improvement but their performance has not been consistent. Feature correlations, specifically, the concordance correlation coefficient (CCC) measure can be used. CCC is useful in finding the reproducibility amongst features. For feature selection, the CCC value for each pair of 3D features was measured. The hypothesis is that high CCC values indicate features that are highly correlated and hence one feature from each such group can be selected. The CCC Heat Map of 3D features on 109 volumes in Fig 4.2 helps visualize the correlation

amongst features. Features are grouped into those having CCC ≥ 0.85 amongst each other. There is no overlap of features in a group. Feature selection is done by randomly selecting one feature from each of the groups. Thus selecting one random feature from each group gives 98 image features. The classifier performance based on those features is shown in Table 4.8. Random forests without feature selection achieves highest accuracy along with the joint highest recall on BAC. IB5 similar to previous results struggles to accurately classify BAC.

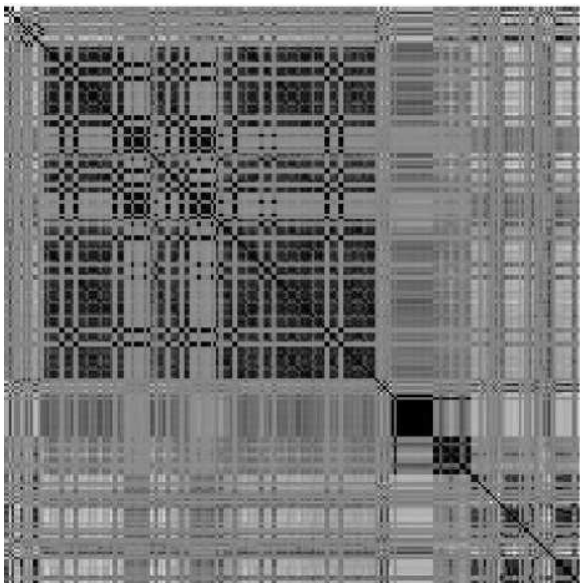


Figure 4.2: Gray-Level Heatmap Representing Concordance Correlation Amongst 3D Image Features Obtained Over 109 Volumes

4.5 Survival Time Prediction

In this section, the study involving 2 year survival time prediction of 95 patients is presented. The experimental outline details 10-fold cross validation and 90-10 split experiment setups. The significance test used for the study is described next, followed by the results.

Table 4.8: Performance of Classifier Models, Using 98 Features, Performing 10-Fold Cross Validation for the BAC Study (The highest accuracy and the Highest Recall % on Bronchioalveolar Carcinoma for each classifier is in bold)

Classifier	Feature Selection	Accuracy (10-fold CV)	Recall On BAC
J48	None	**73.87**	60.00
	Relief-F (50 features)	72.07	57.00
	Relief-F (25 features)	71.17	57.00
	Wrapper	70.27	**63.00**
Random Forests	None	***77.48***	***74.00***
	Relief-F (50 features)	72.07	71.00
	Relief-F (25 features)	65.77	51.00
IB5	None	**76.58**	51.00
	Relief-F (50 features)	74.77	54.00
	Relief-F (25 features)	71.17	40.00
	Wrapper	74.77	51.00
SVM	None	71.17	63.00
	Relief-F (50 features)	**75.68**	66.00
	Relief-F (25 features)	70.27	57.00
	Wrapper	71.17	***74.00***

4.5.1 Experimental Outline

Our experiments are designed to evaluate the effectiveness of image features and clinical features in predicting two year survival. As stated earlier, the effectiveness is measured in terms of the AUC. The experiments were conducted using two approaches. First, a 10-fold cross validation *(CV)* was conducted on the data using the predictive model. The next set of experiments involved a random 90%-10% split of training to test *(90-10 split)* data repeated over 100 iterations. The intuition behind conducting 100 iterations of the 90-10 split was to test the stability of the predictive models on data.

The predictive models were built based on support vector machine, naive Bayes and decision tree classifiers. The evaluation was first done exclusively using image features, followed by introduction of clinical features.

For the predictive model using image features, a novel feature selection approach was evaluated which involved the use of an analysis of feature correlation measured during the Test-Retest analysis, as introduced in Section 3.2.3. Feature selection using principal

component analysis *(PCA)* was applied to the stable features for both exclusively image features and a combination of image and clinical features. Parameter tuning for SVMs was conducted using grid search where the performance was optimized over the cost and γ parameters. The experiments conducted are described in detail in this section.

As discussed earlier in Section 3.2.3, the Test-Retest study on the RIDER dataset, helped reduce the feature space. Using a stricter reproducibility criteria of Concordance Correlation Coefficient (CCC) > 0.85; Dynamic Range (DR) > 100, only 33 features were available for analysis. The available features were then grouped according to their correlation measure. A visual representation of the feature correlation is shown in Figure 4.3 by means of a heatmap. The lighter shade represents high correlation. For the 33 image feature, the grouping was as follows,

- 8 features which were uncorrelated to the other features
- 4 features forming two separate highly correlated pairs.
- 4 features forming a single highly correlated group.
- 8 features forming a single highly correlated group.
- 9 features forming a single highly correlated group.

The 33 features and the groups to which they belong are shown in Table 4.9. Thus 13 groups of features were identified which were highly correlated within their individual group but had low correlation among elements of different groups.

To analyze whether the choice of the highest ranked feature from each highly correlated group was an effective technique, an experiment was also conducted by choosing a random feature from each of the 13 resultant groups. Since the feature being selected was repeated for each fold, the random selection of the features from each of the group was done 10 times to get an average for each fold.

PCA, as described in Section 3.2.4, was conducted on the entire data for the two feature sets, image features and the combination of image and clinical features. The number of

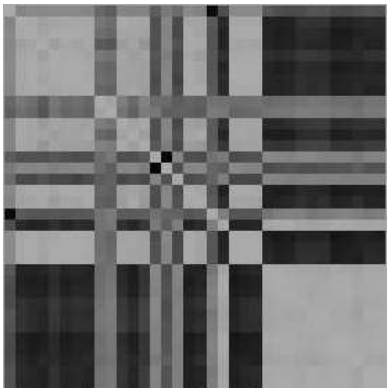


Figure 4.3: Gray-Level Heatmap Representing Pearson's Correlation Amongst Image Features Obtained Through Test-Retest Analysis on RIDER Data Set

principal components to be chosen is determined by the following observations. Figure 4.4 represents the eigenvalue curve for the dataset involving two feature sets. In each case, a drop in eigenvalues around the 10 principal components mark and the eigenvalues of the principal components beyond 15 are almost negligible. Also, Figure 4.5(a) and Figure 4.5(b) represent the variance covered by the principal Components in either case. Based on the two observations, 10 and 15 principal Components were chosen to be used in developing predictive models.

4.5.2 Results

In this section, the results from the experiments conducted are explained in detail. The results are presented in two sections as the experiments were conducted. First presented is the discussion of the results of the 10-fold cross validation followed by result from the 90-10 split.

4.5.2.1 10-Fold Cross Validation

The results from conducting a 10-fold cross validation using the predictive models developed are shown in Table 4.10. The first set of results represent the performance of models built using just image features. The results indicate that a predictive model built on image features using a support vector machine yields an average AUC of 0.67 thus out-

Table 4.9: Features Meeting Reproducibility Criteria: CCC > 0.85 and DR > 100 (Grouped based on Pearson's Correlation > 0.80)

Number of Features in a Group	Number of Groups	Features
1	8	Mean [HU]
		Density
		Rectangular Fit
		Relative Border To Lung
		Relative Border To Pleural Wall
		Ratio Free To Attached
		Av Volume Air Spaces [mm^3]
		Laws Feature S5W5L5
2	2	Volume [cm^3], Area [mm^2]
		Laws Feature W5E5S5, Laws Feature W5S5S5
3	0	. . .
4	1	Avg. GLN, Avg. RLN, Avg. RP, Avg. Coocurrence-contrast
. . .	. . .	. . .
8	1	Laws Feature E5E5E5, Laws Feature E5S5E5, Laws Feature E5W5E5, Laws Feature L5W5S5, Laws Feature S5E5W5, Laws Feature S5R5W5, Laws Feature S5S5E5, Laws Feature S5W5E5
9	1	Avg. Dist. COG To Border [mm], Length (Pxl), Max Dist. COG To Border [mm], Thickness (Pxl), Volume (Pxl),Width (Pxl), Compactness, Longest Diameter [mm], Short Axis [mm]

performing both naive Bayes and decision tree which generated average AUC of 0.61 and 0.63 respectively. The feature selection technique proposed based on correlation does not show improvement over the use of all the features in building the model. However, the method of choosing the highest ranked feature from a group of highly correlated features gives slightly improved performance over selecting random features from each group. This is true for models developed using all the three classifiers. Implementing principal component analysis for reduction of the feature space results in a slight improvement with AUC being 0.68. But, increasing the number of principal components from 10 to 15 does not further

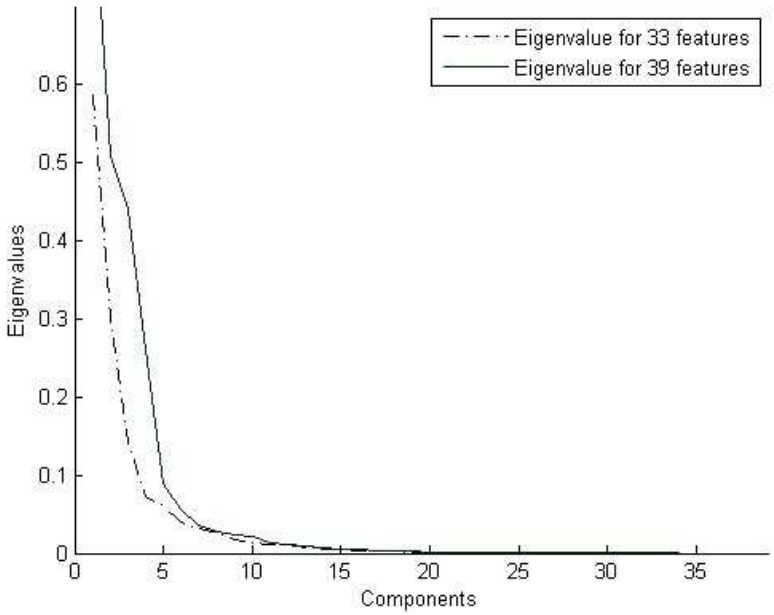


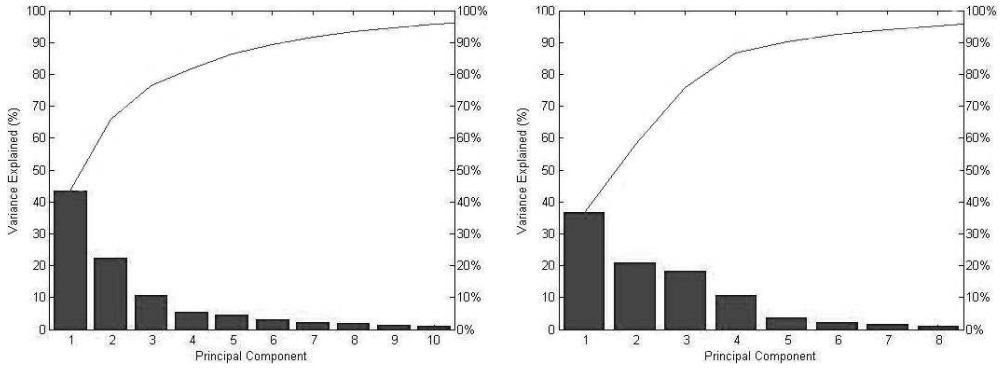
Figure 4.4: Eigvenvalue Plot for PCA

improve the performance of SVM and decision trees, but shows significant improvement in the case of naive Bayes. SVM with an AUC of 0.68 is the highest.

The introduction of clinical features to the feature space has little effect on the average AUC of the predictive models. Using all the features, the AUC remains the same at 0.67 for SVM but improves slightly for naive Bayes and decision tree. The results show similar improvement as shown for only image features after performing principal component analysis. The highest average AUC from a 10-fold CV remains 0.68 for a SVM.

4.5.2.2 90-10 Split

Table 4.11 represents the results from conducting 100 iterations of a random 90%-10% split on the data and averaging the results. As discussed in the case of 10-fold CV, the first set of experiments involved only image features, followed by the combination of image and clinical features. The average AUC achieved using all 33 image features was 0.66 for SVM. Decision tree comes in next with an AUC of 0.63 followed by naive Bayes at 0.62. Unlike in the case of 10-fold CV, feature selection through correlation using feature ranking did not result in a lower AUC for SVM, however, when random features were selected, the



(a) Image Features

(b) Image and Clinical Features

Figure 4.5: Variance Covered by Principal Components

AUC went down to 0.65. Naive Bayes and decision tree showed a noticeable decrease in the AUC when subject to feature selection using correlation. PCA as in the previous case of 10-fold CV resulted in improved performance for SVM and naive Bayes, but not for decision tree. However with an AUC of 0.68 with no variation in regards to the number of principal components, SVM again was the better performing classifier.

For the 90%-10% split, the AUC for using all 39 features was 0.67 for SVM, 0.62 for naive Bayes and 0.64 for decision tree. Unlike previous instances, a SVM model built on 10 principal components failed to show any improvement over using the entire feature space. However, there was comparatively improved performance on introduction of 15 principal components with average AUC of 0.69 in the case of SVM. This was the highest measured AUC for this study. Decision tree shows a slight drop in performance on using 10 PCs but then improves to 0.64 on selecting 15 PCs.

4.5.2.3 Significance Test

The resultant average AUC for 100 random splits for the different classifier models, as shown in Table 4.11 do not vary a large amount. Thus to test whether the difference in performance of each classifier model is statistically significant, Wilcoxon's Signed Rank Test [49] was performed. The signed rank test is a paired, two-sided test based on the

observations for two samples. In this case two classifier models form the sample pair, while the AUC in each of fold of the 100 random 90-10 split form the observations. The signed rank test starts by measuring the difference between the observations for a pair of samples. The differences were ranked without regard to the sign of the difference. All zero differences were ignored. The original sign of the difference was then affixed to the rank numbers. All pairs with equal absolute differences (ties) were assigned the same rank. This was done by ranking them with the mean of the rank numbers that would have been assigned if they would have been different. The sum of all the positive ranks ($W+$), the sum of all the negative ranks ($W-$) and the total number of pairs (N) were determined. The level of significance was calculated by dividing the number of all distributions of signs over the ranks. These values were tabulated and the level of significance can be looked up. For large values of N, (N $\geq$ 15), significance Z is considered to have an approximately standard normal distribution with W, the larger of $W+$ and $W-$. It is calculated as,

$$Z = \frac{(W-0.5-N*(N+1)/4)}{\sqrt{(N*(N+1)*(2*N+1)/24}}$$

In this case a test hypothesis was also defined, that the difference between the matched samples in the paired samples comes from a distribution, symmetric about its median. For each pair of classifier models compared, a test is conducted to test whether the null hypothesis ("median is zero") can be rejected at the 5% level.

The test for statistical significance was carried out between the model with the best average AUC ($M1$), and the two models which were tied for the next highest average AUC ($M2$ and $M3$). The first model, M1, being SVM built on image and clinical features, using 15 principal components. The other two, M2 and M3, being SVM built on exclusively image features, using 15 and 10 principal components respectively. The results are presented in Table 4.12. In the table, the test for rejection of null hypothesis is represented by 0 and 1. In case of successful rejection of the null hypothesis at the 5% level, H = 1, while a failure at rejecting null hypothesis is represented by H = 0.

The results indicate, that the null hypothesis is rejected when comparing 100 random runs of the model with the highest average AUC with the models with the next highest

average AUC. Thus, even though the difference in the average AUC between the models is not very high, there is significant statistical difference between them. Models, $M2$ and $M3$, when compared fail to reject the null hypothesis. Observing the signed ranks of the three comparisons also shed light on the fact $M1$ outperforms $M2$ and $M3$ at different folds of the 100 runs, generating higher rankings.

Table 4.10: AUC for 10-Fold Cross Validation

Feature Set	Classifier	Feature Selection (Feature Space)	Parameter Tuning	Average AUC
Image	SVM	None (33 features)	Grid Search	0.67
		Feature Correlation Random (13 features)		0.65
		Feature Correlation Ranked (13 features)		0.66
		PCA (10 PCs)		0.68
		PCA (15 PCs)		*0.68*
	Naive Bayes	None (33 features)	None	0.61
		Feature Correlation Random (13 features)		0.58
		Feature Correlation Ranked (13 features)		0.60
		PCA (10 PCs)		0.63
		PCA (15 PCs)		*0.66*
	Decision Tree	None (33 features)	None	0.63
		Feature Correlation Random (13 features)		0.60
		Feature Correlation Ranked (13 features)		0.61
		PCA (10 PCs)		0.63
		PCA (15 PCs)		0.63
Image + Clinical	SVM	None (39 features)	Grid Search	0.67
		PCA (10 PCs)		0.68
		PCA (15 PCs)		*0.68*
	Naive Bayes	None (39 features)	None	0.63
		PCA (10 PCs)		0.66
		PCA (15 PCs)		0.66
	Decision Tree	None (39 features)	None	0.63
		PCA (10 PCs)		0.64
		PCA (15 PCs)		0.62

Table 4.11: Average AUC over 100 Iterations of Random 90-10 Splits

Feature Set	Classifier	Feature Selection (Feature Space)	Parameter Tuning	Average AUC
Image	SVM	None (33 features)	Grid Search	0.66
		Feature Correlation Random (13 features)		0.65
		Feature Correlation Ranked (13 features)		0.66
		PCA (10 PCs)		0.68
		PCA (15 PCs)		0.68
	Naive Bayes	None (33 features)	None	0.62
		Feature Correlation Random (13 features)		0.61
		Feature Correlation Ranked (13 features)		0.60
		PCA (10 PCs)		0.65
		PCA (15 PCs)		0.66
	Decision Tree	None (33 features)	None	0.63
		Feature Correlation Random (13 features)		0.58
		Feature Correlation Ranked (13 features)		0.60
		PCA (10 PCs)		0.62
		PCA (15 PCs)		0.62
Image + Clinical	SVM	None (39 features)	Grid Search	0.67
		PCA (10 PCs)		0.67
		PCA (15 PCs)		***0.69***
	Naive Bayes	None (39 features)	None	0.62
		PCA (10 PCs)		0.64
		PCA (15 PCs)		0.66
	Decision Tree	None (39 features)	None	0.64
		PCA (10 PCs)		0.63
		PCA (15 PCs)		0.64

Table 4.12: Wilcoxon's Signed Rank Test on Top 3 Classifier Models

Classifier 1	Classifier 1	W+	W-	W	N	Z	$p \leq$	H
M1	M2	1906.50	-109.50	1797	63	6.15	7.89e-10	1
M1	M3	2535	-93	2442	72	6.85	7.48e-12	1
M2	M3	283	-663.50	-380	63	2.288	0.02213	0

CHAPTER 5

SUMMARY AND DISCUSSION

In this chapter a summary of the observations made and the conclusions drawn from the experiments performed in course of the study is presented. A short discussion regarding future work is provided at the end.

5.1 Result Summary

To begin summarizing the study, first we discuss the experiments conducted to classify lung tumors into Adenocarcinoma and Squamous-cell Carcinoma. The feature selection approaches applied here were often ineffective for 3D features. It is likely that not all features are needed and some selection seems likely to improve the process. There was some improvement with 2D features. For support vector machines, finding the right parameter combination is essential, but our one attempt with 2D features showed no performance difference. Even though the number of 2D and 3D features is large, they are by no means exhaustive. While there are other features that might be of use, we believe this work generally covers the types of features needed to discriminate among lung tumor classes. Clinical features like tumor location within the lung and others can be taken into account along with image features to help uniquely identify tumors.

In the first set of experiments we tried two different simple approaches to feature selection, and both have advantages and disadvantages. Relief-F ignores feature dependencies, but runs very fast. Relief-F was relatively ineffective in choosing a good set of features for this data. Wrappers model feature dependencies, however they run the risk of over fitting. The wrapper forward selection approach was better in selecting features, but the stopping

criterion stops it too early, we believe. We can use backward best first search, but search time then becomes a very significant cost.

Tumors are certainly 3D, but feature selection could potentially improve performance in 2D. On the other hand, there are no ties to be broken with 3D features unlike the case where different slices of a tumor may be classified into separate classes. The most accurate classifiers often used 3D features. Feature selection improved accuracy more for 2D classifiers than 3D, however, it usually did not improve accuracy above that obtained by using all 3D features. There is no clear advantage in accuracy between 2D and 3D features, but 3D simplify classifier construction. Since a vast range of image features have been investigated here, it may be hypothesized that the two classes of tumors chosen for classification may require some specialized image features not developed here.

The attempt at classifying Bronchioalveolar Carcinoma against other NSCLC tumor types was somewhat successful but the significance of that result can be doubted. The attempt at using concordance correlation coefficient as a feature selection method, resulted in somewhat improved performance by the classifier models.

Introducing the reproducibility criteria for features over an independent data set like RIDER does help reduce the feature set and promises some stability in the features that makes the cut. The results indicate that the different models evaluated, though not varying by a large degree, are statistically significantly different from each other. In terms of performance, the support vector machine has consistently outperformed both decision tree and naive Bayes classifiers. It can also be inferred that applying principal component analysis on the feature space does improve the performance of the predictive model. The conditions of reproducibility of features seem to have restricted the performance of the models derived from them to a local maxima that would require other pertinent features to increase the AUC. The clinical features considered are just two, gender and tumor location. The effectiveness of adding other clinical feature is something to be evaluated. But unavailability of complete clinical data for patients for whom images are acquired is often a roadblock in this form of analysis. And finally the variability of the imaging parameters as documented

previously is a major hurdle when developing predictive models on predominantly image features. We note that the AUC here is approximately the same as in another study that had a homogeneous set of images in terms of slice thickness and field of view.

5.2 Future Work

Over the course of this work, one thing that has consistently been a sticking point is the lack of uniformity in the data. Given the fact that there was not much previous work to compare or build upon, the presence of a standardized data set as bench-mark would be of immense help. Apart from that, when working with a substantially large feature space, the amount of patient data should also be increased. Since a large image feature space and wide range of feature selection tools have been examined it may be that the the focus should go towards identifying features beyond the image domain.

REFERENCES

[1] I. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken. Computer analysis of computed tomography scans of the lung: a survey. *Medical Imaging, IEEE Transactions on*, 25(4):385–405, 2006. ID: 1.

[2] B. Ganeshan, S. Abaleke, Rupert C. D. Young, C. R. Chatwin, and Kenneth A. Miles. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging.*, 10(1):137–147, 2010.

[3] Ravi Samala, Wilfrido Moreno, Yuncheng You, and Wei Qian. A novel approach to nodule feature optimization on thin section thoracic ct. *Academic Radiology*, 16(4):418–427, 4 2009.

[4] T. Way, L. Hadjiiski, B. Sahiner, H. Chan, P. Cascade, E. Kazerooni, N. Bogot, and C. Zhou. Computer-aided diagnosis of pulmonary nodules on ct scans: Segmentation and classification using 3d active contours, 2006-06-19T12:51:52 2006.

[5] Michael C. Lee, Lilla Boroczky, Kivilcim Sungur-Stasik, Aaron D. Cann, Alain C. Borczuk, Steven M. Kawut, and Charles A. Powell. Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. *Artificial Intelligence in Medicine*, 50(1):43–53, 9 2010.

[6] Yanjie Zhu, Yongqiang Tan, Yanqing Hua, Mingpeng Wang, Guozhen Zhang, and Jianguo Zhang. Feature selection and performance evaluation of support vector machine (svm)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. *Journal of Digital Imaging*, 23(1):51–65, 2010.

[7] O. S. Al-Kadi and D. Watson. Texture analysis of aggressive and nonaggressive lung tumor ce ct images. *Biomedical Engineering, IEEE Transactions on*, 55(7):1822–1830, 2008. ID: 1.

[8] S. Kido, K. Kuriyama, M. Higashiyama, T. Kasugai, and C. Kuroda. Fractal Analysis of Internal and Peripheral Textures of Small Peripheral Bronchogenic Carcinomas in Thin-section Computed Tomography: Comparison of Bronchioloalveolar Cell Carcinomas with Nonbronchioloalveolar Cell Carcinomas. *Journal of Computer Assisted Tomography*, 27(1):56–61, 2003.

[9] E. Segal, C.B. Sirlin, C. Ooi, A.S. Adler, J. Gollub, X. Chen, B.K. Chan, G.R. Matcuk, C.T. Barry, H.Y. Chang, and M.D. Kuo. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nature Biotechnology*, 25:675–680, 2007.

[10] Harry B. Burke, Philip H. Goodman, David B. Rosen, Donald E. Henson, John N. Weinstein, Frank E. Harrell, Jeffrey R. Marks, David P. Winchester, and David G. Bostwick. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857–862, 1997.

[11] Hugo Aerts. Using Advanced Imaging Features for the Prediction of Survival in NSCLC. personal communication, 2011.

[12] Olvi L. Mangasarian and Edward W. Wild. Proximal support vector machine classifiers. In *Proceedings KDD-2001: Knowledge Discovery and Data Mining*, pages 77–86, 2001.

[13] S. Basu, L. O. Hall, D. Goldgof, Y. Gu, V. Kumar, J. Choi, R. J. Gillies, and R. A. Gatenby. Developing a Classifier Model for Lung Tumors in CT-scan Images. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 1306 – 1312, 9-12 Oct. 2011.

[14] Mitchell E. Garber, Olga G. Troyanskaya, Karsten Schluens, Simone Petersen, Zsuzsanna Thaesler, Manuela Pacyna-Gengelbach, Matt van de Rijn, Glenn D. Rosen, Charles M. Perou, Richard I. Whyte, Russ B. Altman, Patrick O. Brown, David Botstein, and Iver Petersen. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences*, 98(24):13784–13789, November 20 2001.

[15] S. Basu, L. O. Hall, D. Goldgof, Y. Gu, V. Kumar, R. J. Gillies, and R. A. Gatenby. Classifying Lung Tumors from CT-scan Images Based on 3D Image Features. Poster at 2011 World Molecular Imaging Congress, San Diego, 2011.

[16] Rene Korn, Arno Schaepe, Guenter Schmidt, Gerd Binnig, and Claus Bendtsen. Lung Tumor Analysis (LuTA) with Definiens Cognition Network Technology, 2010.

[17] M. Baatz, J. Zimmermann, and C.G. Blackmore. Automated Analysis and Detailed Quantification of Biomedical Images Using Definiens Cognition Network Technology. *Combinatorial Chemistry & High Throughput Screening*, 12(9):908–916, 2009.

[18] K. Kuriyama, R. Tateishi, O. Doi, K. Kodama, M. Tatsuta, M. Matsuda, T. Mitani, Y. Narumi, and M. Fujita. CT-pathologic correlation in Small Peripheral Lung Cancers. *American Journal of Roentgenology*, 149(6):1139–1143, December 01 1987.

[19] Benoit B. Mandelbrot. *The fractal geometry of nature*. Freeman, San Francisco., 1983.

[20] Fereydoon Family, Barry R. Masters, and Daniel E. Platt. Fractal pattern formation in human retinal vessels. *Physica D: Nonlinear Phenomena*, 38(1-3):98, 1989.

[21] T. G. Smith Jr., W. B. Marks, G. D. Lange, W. H. Sheriff Jr., and E. A. Neale. A Fractal Analysis of Cell Images. *Journal of neuroscience methods*, 27(2):173, 1989.

[22] R. Chellappa and R. Bagdazian. Fourier coding of image boundaries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(1):102–105, 1984. ID: 1.

[23] M. M. Mokji and S. A. R. Abu Bakar. Gray level co-occurrence matrix computation based on haar wavelet. *Computer Graphics, Imaging and Visualization, International Conference on*, 0:273–279, 2007.

[24] V. A. Kovalev, F. Kruggel, H. J Gertz, and D. Y. von Cramon. Three-dimensional texture analysis of mri brain datasets. *Medical Imaging, IEEE Transactions on*, 20(5):424, may 2001.

[25] D. H. Xu, A. S. Kurani, J. D. Furst, and D. S. Raicu. Run-length encoding for volumetric texture. In *Proceedings of The 4th IASTED International Conference on Visualization, Imaging, and Image Processing - VIIP 2004, Marbella, Spain.*, September 6-8 2004.

[26] Xiaoou Tang. Texture information in run-length matrices. *Image Processing, IEEE Transactions on*, 7(11):1602–1609, 1998. ID: 1.

[27] A. S. Kurani, D. H Xu, J. D. Furst, and D. S. Raicu. Co-occurrence matrices for volumetric data. In *Proceedings of The 7th IASTED International Conference on Computer Graphics and Imaging - CGIM*, 2004.

[28] K. Laws. Textured image segmentation. Technical Report USCIPI-940 Image Process. Inst. Univ. of Southem California, 1980.

[29] K. K. Benke, D. Cox, and D. R. Skinner. A study of the effect of image quality on texture energy measures. *Measurement Science and Technology*, 5(4):400, 1994.

[30] K. Jafari-Khouzani. Comparison of 2d and 3d wavelet features for tle lateralization, 2004-05-10T09:29:31 2004.

[31] J. R. Quinlan. Decision trees and decision-making. *Systems, Man and Cybernetics, IEEE Transactions on*, 20(2):339–346, 1990. ID: 1.

[32] L. Breiman, J. H. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.

[33] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[34] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998. ID: 1.

[35] David Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In Claire Ndellec and Cline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 4–15. Springer Berlin / Heidelberg, 1998. 10.1007/BFb0026666.

[36] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1993.

[37] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor.Newsl.*, 11(1):10–18, November 2009.

[38] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. 10.1023/A:1009715923555.

[39] K. A. Kramer, L. O. Hall, D. B. Goldgof, A. Remsen, and Tong Luo. Fast support vector machines for continuous data. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(4):989–1001, 2009. ID: 1.

[40] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[41] J. Dehmeshki, J. Chen, M. V. Casique, and M. Karakoy. Classification of lung data by sampling and support vector machine. In *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE*, page 3194, 1-5 Sept. 2004.

[42] Chih chung Chang and Chih jen Lin. Libsvm: a library for support vector machines, 2001.

[43] MATLAB. *version 7.10.0 (R2010a)*. The Mathworks Inc., Natick, Massachusetts, 2010.

[44] I. Kononenko and L. De Raedt. Estimating attributes: Analysis and extension of relief. In F. Bergadano, editor, *Proceedings European Conference on Machine Learning*, 1994.

[45] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273, 1997. Relevance.

[46] V. Kumar, Y. Gu, K. Jongphil, R. A. Gatenby, and R. J. Gillies. Test retest reproducibility of image features extracted from ct images of lung tumors. 2012.

[47] Ethem Alpaydin. Combined 5x2 cv f test for comparing supervised classification learning algorithms. *Neural Computation*, pages 1885–1892, 1999.

[48] David H.Garfield and Jacques L. Cadranel. The bronchioloalveolar carcinoma and peripheral adenocarcinoma spectrum of diseases. *Thoracic Oncology*, 1(4):1556–0864, 2006.

[49] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):pp. 80–83, Dec. 1945.