5-2015

Computational methods to predict and enhance decision-making with biomedical data.

Behnaz Abdollahi
University of Louisville

COMPUTATIONAL METHODS TO PREDICT AND ENHANCE DECISION-MAKING WITH BIOMEDICAL DATA

By

Behnaz Abdollahi
B.S., University of Louisville, 2000
M.S., Sharif University of Technology, 2006

A Dissertation
Submitted to the Faculty of the
J.B. Speed School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering
University of Louisville
Louisville, Kentucky

May 2015

COMPUTATIONAL METHODS TO PREDICT AND ENHANCE DECISION-MAKING WITH BIOMEDICAL DATA

By

Behnaz Abdollahi
B.S., University of Louisville, 2000
M.S., Sharif University of Technology, 2006

A Dissertation Approved On

March 30, 2015

by the following Dissertation Committee:

Hermann Frieboes, Dissertation Director

Karla Conn Welch, Dissertation Co-Director

Ayman El-Baz, Dissertation Committee Member

Jacet M. Zurada, Dissertation Committee Member

Cindy Harnett, Dissertation Committee Member

DEDICATION

To all beloved people in my life.

ABSTRACT

COMPUTATIONAL METHODS TO PREDICT AND ENHANCE DECISION-MAKING WITH BIOMEDICAL DATA

Behnaz Abdollahi

March 30, 2015

The proposed research applies machine learning techniques to healthcare applications. The core ideas were using intelligent techniques to find automatic methods to analyze healthcare applications. Different classification and feature extraction techniques on various clinical datasets are applied. The datasets include: brain MR images, breathing curves from vessels around tumor cells during in time, breathing curves extracted from patients with successful or rejected lung transplants, and lung cancer patients diagnosed in US from in 2004-2009 extracted from SEER database. The novel idea on brain MR images segmentation is to develop a multi-scale technique to segment blood vessel tissues from similar tissues in the brain. By analyzing the vascularization of the cancer tissue during time and the behavior of vessels (arteries and veins provided in time), a new feature extraction technique developed and classification techniques was used to rank the vascularization of each tumor type. Lung transplantation is a critical surgery for which predicting the acceptance or rejection of the transplant would be very important. A review of classification techniques on the SEER database was developed to analyze the survival rates of lung cancer patients, and the best feature vector that can be used to predict the most similar patients are analyzed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The domain of the research is to design and implement machine leaning models and automatic evaluation of biomedical data. Automatic classification and decision making models are a big help in determining the best treatment and its effect on patients before making the final decision. In this dissertation several solutions are developed to health care applications using applied machine learning and data analytical techniques. The input data were raw and unstructured, so they needed at the first step to clean the data and convert them to structured data; the second step was to extract the features and select the best feature vector; and the third step was to choose the optimized classification technique. The nature of the data leads to use a unique procedure for each dataset. The given datasets include: database of lung cancer patients in US between 2004-2009 to analyze their survival rate, perfusion curves generated from different types of breast cancer that were growing in mice, with measurements of vessels around the tumor to define tumor ranking; perfusion curves of patients to predict rejected or accepted renal transplants; brain MRA images to detect and extract blood vessels. Machine learning techniques are divided into three groups: classification techniques (supervised), clustering techniques (unsupervised) and semi-supervised methods. This research is focused on unsupervised and supervised learning. The labelled data are needed in advance to apply the classification algorithms, so the mathematical models are trained based on labelled data. The labelled data of lung cancer patients was not given in advance, so the records were categorized and labelled based on their survival time, and a novel labelling procedures was developed. Unsupervised techniques do not need any labelled data in advance and they categorize similar

instances in the same cluster. In this section, the techniques that have been previously applied on similar datasets are reviewed. Previous works on vessel segmentation image analysis, integration of mathematical modelling of cancer growth and image analysis, perfusion curves analysis, and finally the machine learning techniques that were applied on the SEER database are reviewed. Medical image analysis is high demand research areas which can help clinicians find abnormalities in a more accurate way. The quality, slice thickness, resolution, and tissue location are different parameters that make it much harder finding a robust solution to segment tissues. The research started with focus on reconstruction of blood vessels. The parameters that make the extraction of blood vessels challenging are anatomical variability of the vasculature, location of the blood vessel, image contrast, resolution an also the imaging modality. Scale space smoothing is developed, which smooths images at different scales by employing diffusion equations. Peronal and Malik [134] proposed a new scale space edge detection method based on diffusion equations. Weickert added orientation to the diffusion filter to be able to enhance small vessels and coherence structure in images [183, 182]. Subsequent methods replaced the diffusion scalar by diffusion tensor employing the Hessian matrix configuration and analysing eigenvalues of the Hessian matrix. Different geometric interpretation extracted from Hessian matrix and its eigenvalues configurations [81, 104, 148, 46, 24]. For review of anisotropic diffusion, please refer to [164, 181, 39]. Krissian [86] and Manniesing [113] proposed anisotropic diffusion filters to segment vessels in 3D, based on a tensor structure filter. Fischl proposed a new method to indicate the best kernel function that matches the image [42]. For further review on vessel analysis read [97]. The proposed technique uses the scalar diffusion function, and is mainly based on the conventional Perona and Malik nonlinear diffusion filter and improves the efficiency of the algorithm using EM technique, it is more simple and it needs less computational calculation considering 3D vessel analysis[6]. One of the main issues with the proposed technique on vessel segmentation was lack of robust knowledge about the disease and also the tissue materials, so image analysis techniques cannot

be trusted as an accurate and personalized package for each patients on its own. The developed technique needed to be able to capture the vessels around the tumor to be able to predict the tumor growth. A new approach to solve this limitation issue with medical image analysis would be a combination of mathematical cancer modelling and image analysis which can personalize and predict the cancer growth in patients. The mathematical cancer growing models use biological factors to calculate the velocity of the tumor growth or the diffusion of the cancer cells in particular tissues. Medical image analysis would be able to predict the tumor growth based on biological information taken from the mathematical modelling and also the location of the tumor given in the images. The new approach seems more practical, however it assumes biological parameters to be the same for all the patients. The next perfect solution would be using supervised learning and train the classifier with a huge number of patients so the biological parameters are trained based on realistic tumor growth and also a large number of images would be trained with different parameters. This solution is one the most practical and personalized solution to predict the growth of the disease in near future for each patient. The idea did not get the chance to develop because enough number of data could not get collected, so only a review was written of the techniques in a book chapter [5] was published. The main advantage of this combination is personalization of tumor evolution. Tumor growth factors are extracted for each patient based on the location of the tumor and its surrounded tissues and also the tumor growth rate might be different for each patient. Image series are one of the input information that is given for each individual patient. As discussed one important aspect of the tumor growth cancer is analyzing the blood vessels that are in the tissue. Perfusion imaging is typically used as a method for determining prognosis in the clinic [163]. Imaging of perfusion through a tissue has been used to measure vascular geometry and histological features of tumor angiogenesis, and also to estimate micro-vascular flow through capillaries and venules [100]. The provided perfusion curves were generated from intra-vital microscopy (IVM), which injects a fluorescent tracer to measure the blood volume and tissue permeability. The fluorescence intensity of

each vessel was measured over time to yield a heterogeneous set of arterial and venous perfusion curves on a tumor-by-tumor basis. Two features were considered: the time to arterial peak and the venous delay, which acted as inputs for a Fuzzy C-Mean(FCM) clustering. FCM technique was chosen because it uses a membership function for every instance and it calculates a number for each instance in each cluster. The data was classified into three defined groups (poorly vascularized, well vascularized, and in between vascularized), which were correlated to experimental nanoparticle accumulation measurements. This approach enables an automated ranking of tumor vascular perfusion in order to model the delivery of nano-therapeutics. Using an independent validation set, demonstration of technique shows that new samples can be mapped into the feature space to determine their perfusion ranking and hence estimate their nanoparticle retention. The feature using a mathematical technique was extracted, which is called gamma variate function; this finds the best fit for each curve and extract the mathematical features of the fitted curve as a feature vector. The same feature extraction algorithm was used on another similar database which is a dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) and is a non-invasive imaging technique that has been explored in perfusion-related concerns in many clinical applications, e.g., in evaluation of the kidney, brain and heart. The pioneering work of Larson, Tofts and Brix enabled the modeling of tracer kinetics using DCE-MRI [94, 93]. Advances in MRI technology in later years enabled models to estimate perfusion and capillary permeability more accurately. Recent studies have focused on revealing physiological characteristics of the tissues. The main idea is to extract a relation between the perfusion and vascular functionality of the tissue, which enables measuring blood volume and capillary permeability. Newer models have focused on extracting these critical tissue features [106]. Machine learning approach was applied on the given dataset and developed a prediction package. At short times (up to about two minutes) after administration at DCE-MRI, parameters can be derived that reflect the agent delivery to the tissue bed. A function-based model was used to analytically classify the perfusion of renal transplant patients in order to determine

their prognosis in terms of transplant acceptance or rejection. These curves quantify the average intensity of renal perfusion for up to four minutes. The algorithm first identifies a model function that can be consistently fitted to all of them, as discussed before the algorithm extracts the mathematical parameters of the gamma variate function. Then, the algorithm classifies the output of the model into two groups, namely, non-rejection (successful) or rejection (unsuccessful) transplants. After selecting the features of the function classes and training the data classifier, this classifier was used to classify new (unknown transplant outcome) curves.

The final research was to apply machine learning on the SEER database. The original database has more than 100 features with most of them having overlap information, so those features was excluded and only considered those that have the lowest overlap information. The features were measured directly from the patients and showed the physical basic information of each patient. Different classification techniques was applied and results are compared. The main goal was to categorize the patients based on their survival rate, so analyzed the relation between the survival time and the patients in the same class were analyzed. Clinicians typically use a combination of features which include the tumor size, the distribution of tumor in other organs, and being a primary or non-primary lesion to predict the stage and finally the survival range of the patients; however the statistical analysis distinguishes the best feature vectors that can predict the survival rate of the patients in a more accurate way.

CHAPTER II

OVERVIEW OF MEDICAL IMAGE SEGMENTATION

This chapter reviews most of the important papers in vessel segmentation techniques and it will focus on novel idea on vessel segmentation using multi enhancement technique and probabilistic information.

The shape of vessels is unique, therefore requiring application of special image analysis techniques to extract its structure accurately. In diagnostic imaging, magnetic resonance angiography (MRA) and X-ray Computed Tomography Angiography (CTA) are two main 3D modalities that are primarily used to image network of vasculatures. In both types of acquisition, typically the vessels are brighter than other organs. However, the vessels are embedded in organs, making their automated extraction a difficult task.

Numerous approaches have been proposed for vessel extraction in the past. Previously published vessel segmentation methods are categorized into 5 classes: 1) Multiscale, 2) Ridge-based, 3) Skeleton-based, 4) Region-Growing, and 5) Active Contours (including both parametric active contours and level set technique.) In each group, only a few of some of the previously published papers was discussed, representing principal features of each group.

A Multi-Scale Approaches

Multiscale framework by itself is not utilized to segment the vessels. The method can handle different vessel width, so this critical feature making it as a good alternative to be combined with other approaches.

An image results from a physical measurement. Imaging devices fix the scale of the outside world that is observed [193]; the captured images are only available in

one scale. Scale space theory provides an appropriate framework for analyzing images at different resolutions. Analyzing images in different scales is an appropriate technique for bringing into focus only objects that have a specific size, thereby increasing the accuracy of computations. Indeed, multi-scale analysis is biologically inspired and is the way that human eye visualizes the outside world.

If vessels are considered in only one slice of a CTA or MRA dataset, it will be found that a continuum of vessel sizes is present. Therefore segmentation needs to be implemented in different scales. Large vessels need to be segmented in large scales (low resolution) and small vessels should be analyzed in small scales (high resolution). In the linear scale space approach [193] a Gaussian kernel at a range of scales is convolved with the image.

The symmetrical two dimensional Gaussian kernel is:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{(x^2+y^2)}{2\sigma^2}} \tag{1}$$

where x and y are the distance from the origin in the horizontal and vertical axis and is the standard deviation of the Gaussian kernel. If the standard deviation in both x and y axis are the same, the equation of the Gaussian function is symmetry, its equation is the same as the above one. Figure 1 shows a two dimensional symmetrical Gaussian function.

σ is the important parameter of the Gaussian function, if is chosen high then the image will become blur, and if it is chosen as a very low value the image will have higher resolution and less blurring is occurred. Selecting an appropriate scale is a challenging problem. Here some techniques are considered which are as a representative for multiscale vessel segmentation.

If larger scale are chosen (σ of the Gaussian) the blurrier (at a low resolution) the image would be, however, the smaller scale would results in higher resolution and less blurring.

Applying the multi scale approach might not be a good solution for segmentation by itself; so the method is combined with other approaches. Frangi [46] is a good representative for utilizing multiscale technique as a feature to

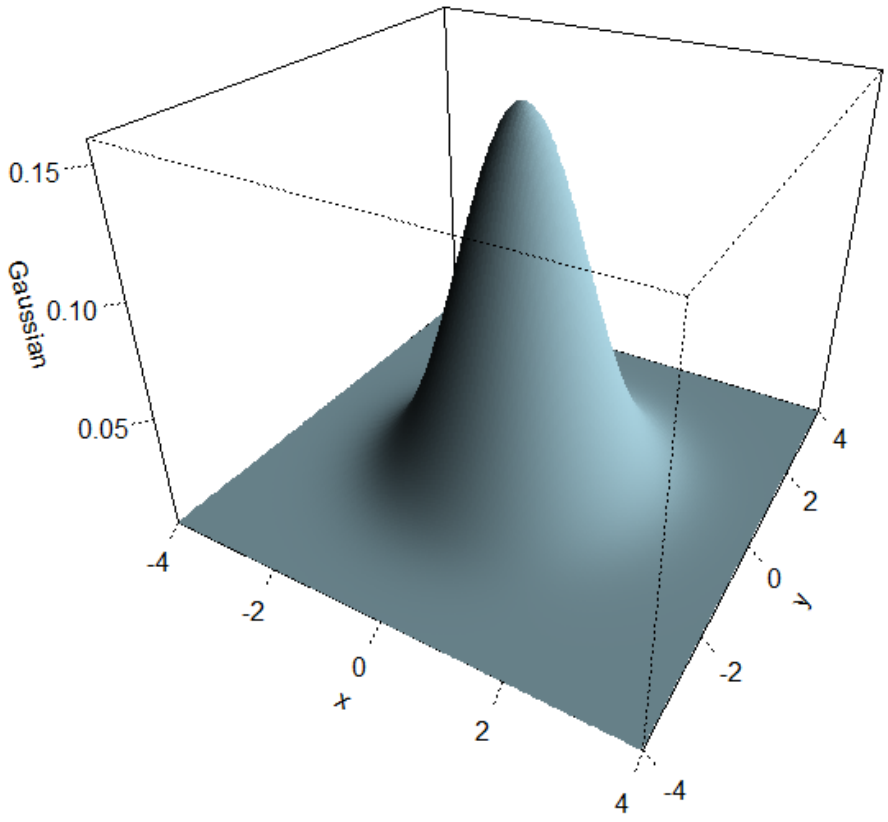

Figure 1. The symmetrical two dimension Gaussian function with mean(0,0) and standard deviation 1.

extract vessels.

Multiscale Hessian-based filters are commonly used for vessel enhancement and segmentation. Hessian-based filters are designed by calculating the eigenvectors and eigenvalues of the Hessian matrix in 2D or 3D. The filter discriminates between the plane and tubular structure and in so doing quantify a measure of vesselness. The filters are applied in different scales; so vessels with different sizes are segmented accurately. The best scale selection for each vessel size is defined by computing the filter response in different scales and choosing the highest filter response for each pixel. Frangi [46], Li [101], Shikata [149],[24] used the Hessian filter to enhance and segment the vessel structures.

Frangi [46] reconstructs the vessel network using a multiscale Hessian based filter. The response of the Hessian filter is calculated in different scales and the maximum response is chosen as the best scale for extracting the vesselness.

Most papers apply their segmentation method in different scales and the final result is a combination of the results. Alyward and Bullitt [15, 14] propose a ridge based approach, taking advantage of multiscale. ter Haar Romany [164] in a tutorial describes multi-scale methods for computer vision. He describes the difference between the human visual system and digital image capture, and describes the mathematical background of the multiscale methods. In [193], Sporring discusses the mathematical background of Gaussian scale space theory and its usage in medical image analysis.

Most of the scale-space segmentation methods apply linear scale space techniques though nonlinear scale space has also been utilized for vessel enhancement and segmentation. Nonlinear scale space filters are mostly employed as a preprocessing step in order to reduce the noise and homogenize the vessel regions. The idea being that preprocessing filters make the segmentation more robust to noise.

The principal difference between the linear scale space (also known as linear diffusion filter) and nonlinear scale space (also known as nonlinear diffusion filter) is the diffusion kernel function. Linear Gaussian scale space can refine the image to

different scales but the same kernel function is applied all over the image independent of the local image structure of the image. Nonlinear diffusion filter which was first proposed by Peron and Malik [134] blur the image taking into account local image structure in that important edge content is preserved. The diffusion function is based on the gradient of the image and the anisotropic diffusion function is defined as:

$$I_t = div(c(x, y, t)\nabla I) = c(x, y, t) \cdot \Delta I + \nabla_c \cdot \Delta I \tag{2}$$

where c is the diffusion coefficient, is gradient operator of the vector space, is the laplacian operator, div is the divergence. Divergence of a continuously differentiate vector field (F) is defined as $divF = \nabla F = \frac{\delta U}{\delta x} + \frac{\delta V}{\delta y} + \frac{\delta W}{\delta z}$ and ∇ is the symbol of divergence in mathematics and x, y and z are the Cartesian coordinates of a 3D Euclidean space. F is the vector field where: $F = Ui + Vj + Wk$ where i, j and k are the unit vectors in 3D Euclidean space. If c is a constant then the equation reduces to $I_t = c(x, y, t) \cdot \Delta I$ which it is an isotropic function. The non-linear diffusion would be achieved if for instance c is set to 1 in the interior of regions and 0 on the boundaries then the blurring would be applied only in the regions and the boundaries are not blurred. Krissian et al. defined a new anisotropic diffusion filter for vessel segmentation which more accurately preserves small vessels. In their approach the diffusion filter is weighted based on the gradient direction and the maximal and the minimal principal curvatures [85]. Manniesing et al.applied a combination of Hessian and nonlinear diffusion filters to improve the segmentation of small vessel structure and the connectivity of vessels [113]. Weickert [183] employed a tensor based technique which improved the connectivity of the discontinuities in vessel structure [183]. Some scale space based papers are devoted to noise reduction and enhancement. See [116] for an example.

B Ridge-Based Methods

Ridge based methods utilize the intensity of the gray scale image as the third (or fourth) dimension, in addition to the two (or three) spatial dimensions. The

intensity of a two dimensional image is viewed as a height map. The problem in ridge-based approaches is cast as one of finding the ridges and valleys in the image height map. Intensity ridges are assumed to be a path along the mountain peaks and valleys. Some approaches use the scale space to extract the ridges [102].

Alyward and Bullit [15] used ridges as local maxima to extract the vessel centerlines, exploiting the size and location of the tubular vessel structure. At the first stage the image intensity is mapped to height to create intensity height surface. Then a manually seed is selected on the ridge as the initial point. Then based on the user-defined seed point a conjugate directions search with respect to Hessian matrix is applied to find the ridge points. Finally, the local widths of the vessel object is estimated using the points which are found on the ridges based on the conjugate directions search. For extracting the vascular tree structure more than one hundred mouse clicks is needed to get the seed points on the ridges. The clicks are the auxiliary points that are needed to be defined in different section of the tree structure.

Fridman et al. [47, 48] used Cores in order to extract the height map skeleton. To reduce the noise the image is filtered and enhanced, then it is segmented using marching technique utilizing the medial atoms and the medialness. The medial atom is a structure of four parameters, its components are (x,r,F,θ), where is the coordinate of the atom in 3D space, r is the radius of the object in the location x, F defines the orientation of the atom, and theta is the object angle which shows characteristics of an object such as its widening and narrowing. For a tube shape each atom with its four parameters imply a set of concentric vectors, named as spokes, that extends from the medial location x to the implied object boundary. Medialness of a medial atom is a scalar value that measures the fit of the medial atom to image data.

Spokes are concentric cores; each spoke is a vector and all the vectors have the same origin and the same radius from the center of a circle. Spokes define the boundary and segment the object correctly. The derivative of the Gaussian is measured in the direction of each spoke and the weight of each spoke is computed

and added resulting in the medialness value of the object [47, 48] . For contour detecting the method uses Gaussian derivatives as the edge detector and the constraints that as discussed on each medial atom is combined with it to find the vessel wall. The spokes are all supposed to be the same length which it assumes that the object is perfectly tubular. But the implementation results show that, the method can segment objects that are not completely circular in cross section, but if the cross section of the not a complete circle cross section is so long the method cannot segment it correctly.

The first core is manually specified. To determine subsequent cores, the marching algorithm in the tangent direction of the core or atom is applied. Additionally, an optimization method is used to optimize the location of the next core or atom. The location of the next core is found by further optimization over the spatial plane normal to the core tangent and passing through the predefined position of the core. The stopping criterion is either when the signal to noise ratio is found to be low or when the object traversing the cores has explicitly ended.

C Skeleton-Based Methods

Skeleton-based methods extract the centerline of the blood vessels. Subsequently, the centerlines are connected and the tree structure of the vessels is explored. The resulting centerline structure is utilized for 3D reconstruction. Since ridge based approaches that was discussed in the previous section detect skeleton of the desired object, it can be thought of as a specialized skeleton based methods.

The main reason for the preference for this group over alternative techniques is that computation is reduced to one dimension. The approach offers simplicity for different clinical measurement like stenosis and aneurysm quantification. Giving a brief description these methods apply thresholding and then object connectivity, or thresholding followed by a thinning procedure, or extracting based on a graph description. The extracted centerline is used for 3D reconstruction. Tracking based approaches start with a seed point given by the user. Subsequently, the position of the path is estimated and adjusted.

Tyrrell et al. [172] implements a new method on optical slice data imaged in vivo. The images have several artifacts like circulatory motion of the tissues, gaps and static red blood vessels. So the proposed method overwhelm these artifacts. The approach predicts the direction of the centerline utilizing the statistical estimator. A superelliposoid geometric model was used to find the vessel boundaries.

A superellipsoid is a geometric model with three parameters and is defined as:

$$|\frac{x}{a}|^n + |\frac{y}{b}|^n = 1 \tag{3}$$

where n, a, b are the parameters. In 3D the superellipsoid is defined as:

$$\left(|x|^{\frac{1}{\varepsilon_2}} + |y|^{\frac{2}{\varepsilon_2}}\right)^{\frac{\varepsilon_2}{\varepsilon_1}} + |z|^{\frac{2}{\varepsilon_1}} = 1 \tag{4}$$

Frangi et al. [44] and Wink et al. [185] use Hessian based filter to estimate the position of the centerline path. Fridman et al. [48, 47], Alyward et al. [14, 15], and Wink et al. [186] calculate the centers of the cross section separately and utilize the extracted information for changing and optimizing the vessel centerline model. Wink et al. [186] used multiscale vessel tracking based on the Hessian matrix.

Most of the centerline based techniques need a point as the seed point to start the extraction process or some of them need more points to be defined manually on the skeleton. Lacoste et al. [89] extracts the coronary vessel network imaged by X-ray angiography without defining any seed point. It is considered as a multiscale technique; for segmenting thick branches, it uses a coarse scale, while for smaller branches of the tree structure, it uses fine scale. And finally the optimization process via simulated annealing is done. For modelling the centerline a Markov object process is applied. Besides the points, lengths and orientations are considered as the components of the object. This is the reason that why it is called Markov object process. The Poisson distribution is assumed for modelling the data object. The centerline is extracted by Markov object process specified by a uniform Poisson process. The Markov segmentation process is based on some assumption: the gray level between the network and the background is large, the local average gray level inside the network is homogeneous. The centerline point is penalized or

reward based on which the segments are connected or disconnected. For completing the segmentation procedure, the segments that are defined by lines are connected and the initialize tree structure is extracted. Then the edges are defined for each segmented line which is the vessel wall. Then the optimization method completes the procedure and the final tree network is defined.

Subsequently, the tree branches are extracted using the centerline of the segmented regions where each segment has a center line. Instead of using piecewise curves, Frangi et al. [45]used B-spline curves to extract a smooth centerline and then the vessel walls were defined by cross section structures swept along the curve. The method will be discussed in detail under Hessian based filters methods in Active Contour section.

Another method was proposed by Wesarg et al. [184]. It can be categorized as a method that uses 2D cross section in order to find the centerline. A point is given in the vessel and the direction of the center is defined by cross section plane, and the vessel wall is along a circle which p is its center. Wesarg used intensity thresholding to select candidate contour points. Corkscrew method is utilized to select the points on the contours. The search direction is along the x,y and z axis; if the search finds a border; then this position is stored and the search continues along the next axis. The start and the end point are given manually. The center of mass of the search points measure the centerline. The border is extracted based on thresholding and applying a morphological filter and then the median filter. A polygan mesh using marching cubes is used to visualize the segmented vessel.

Some of the centerline methods use the path in 3D intensity images; the boundaries are extracted using surface evolution method which will be described later in this report in the level set approach section. The Euclidean distance function is calculated from the boundaries so the desired path is centered in 3D. The user needs to define some points or specify a particular path using auxiliary points. For instance the first and the last point and some points on the path are specified manually. Finally the propagation between the two start and end points is implemented by the fast marching algorithm [131].

Briefly, a 3D approximation of the skeleton is built and a weighted graph is made which the weight assignment is a function of the Euclidean distance from a user defined source and Euclidean distance from the boundary of the object. A minimum path finder like Dijkstras algorithm is applied to the weighted graph to find the centerline path [190].

Wink et al. [186] presented a technique to extract the centerline based on the Frangi Hessian matrix features in order to find a measure of vesselness and the centerline locations. A minimum path algorithm then connects user-selected points to recover entire centerlines. The centerline generates the skeleton of the vessel which is then combined with 3D visualization method in order to extract the vessel surface.

D Region Growing Approach

The principal steps of techniques in this group are similar but the functions utilizing for how to grow the region is varying. In the first step, a seed point is manually selected in the region of interest; a function for vessel region definition is specified and if the neighbours' pixels satisfy the constraints of this function then they are selected as vessels points and hence the region is expanded. The algorithm iterates until there are no unlabelled pixels in the image. Every pixel is tested only once. The test function is substantially based on the intensity similarity between the seed point and its neighbours. Hence, both the spatial coordinates and intensity values have a major role in making a decision about each pixel. Commonly the methods use edge detection in order to find the region boundaries. Besides the need for defining a manual seed point for the start point the algorithm could result in holes in the region or could potentially over-segment the region. Region growing approaches basically assume that that pixel that are close to each other and have similar intensity value belongs to the same region. Vessel images suffer from noise or artifacts like partial volume artifacts so region growing by itself cannot be a good choice for segmenting; however they may be combined with other methods.

A method proposed by Yim et al. [190] is based on the ordered region

growing (ORG) and represents the image as an acyclic graph. It used a gray-scale skeletonization method for finding vessel tree structure and was applied to MRA images. ORG produces a graph extracted from the image. Within this graph, the path between each two points in the image is defined. The region extending is by choosing any point on the boundary which has the highest intensity. So in 2D images the eight neighbors are considered and in 3D the 26 vicinity voxels are tested for finding the one with the highest intensity for iteratively extending the region. After defining the graph the skeleton of the vessel is found. Two different algorithms were proposed in the paper. Some auxiliary points are defined by the user, so significant paths are extracted from the graph. Therefore, the skeleton of the vessel is found based on the ORG graph, the seed point, as well as the end points are also defined by the user.

Another skeleton technique is based on pruning small branches removed from the main vessel. The pruning of the branches is based on the fact that no branch is to be retained if the distance from the end to the nearest bifurcation is less than a given minimum length.

E Active Contours

The original paper which proposed the use of active contours (also known as snakes) in image analysis was Kass et al.s [73]. An active contour is an elastic spline curve which in response to external forces derived from the image and internal smoothness forces derived from the curve geometry deforms. Fundamentally, external forces are designed to push the snake towards image features and the internal forces are designed to keep the snake smooth. The process is governed by iterative numerical optimization of an energy function which is a linear combination of the internal and external energies. The literature on active contour models may be classified into two groups: those employing parametric active contours (or snakes) and those employing implicit active contours (or more commonly referred to as the level set technique).

1 Parametric Active Contours

Eigensnakes is method proposed by Toldeo [169] for vessel segmentation. It is an automated segmentation method which learns the shape structure based on a statistical feature vector and the target shape structure is found by a likelihood criteria in the feature space. The eigensnake describes the optimal object structure. Vessel description is defined by filtering the image with a Gaussian filter in different scales with the maximum filter response being chosen. To avoid the large size of the data the algorithm uses principle component analysis to reduce the feature space measurement. The measure can be regarded as a likelihood function giving the probabilities of each pixel belonging to a vessel. The likelihood map is specified in the statistical feature space and incorporated within an energy minimization framework. A Mahalanobis distance map is account as a measure based on the distance between the clustered training dataset and the new object.

The feature space shows the vessels so if the Mahalanobis distance between the training and test data vector in the feature space is small it shows that the new data belongs to vessel. Because if the distance between the training data set and the test data set in the feature space is zero then the two vectors shows the same feature vector of the object and it resemble that the objects are the same. The intrinsic probabilistic nature of the eigensnake makes it possible to obtain the whole tree structure using a likelihood probabilistic map. The direction of the vessels are calculated based on the principle component analysis of the intensity distribution of the image.

Previously, Yim et al.s algorithm [190] was cited and noted that his approach was in the skeleton- based class of techniques. In fact, that algorithm belongs to the class of tubular deformable models - based on the centerline the surface mesh is extracted.

Frangi [43] proposed a tensor-product-based B-spline modeling scheme where the control points were optimized. This is one of the most well-known vessel segmentation methods. Earlier, the methods was discussed in the context of

multiscale-based methods (see corresponding section). Frangis method can be considered as a snake and/or deformable model in three-dimensions and is based on the Hessian matrix. The model extracts the central vessel axis and smoothes the center axis using a B-spline representation. The curvature of the central curve of the vessel is extracted by computing the Hessian matrix of each voxel. The eigenvalue of the Hessian matrix is calculated in different scales so the method can be classified as the multiscale method vessel segmentation too. The vesselness in different scales is calculated based on the eigen values of the Hessian matrix. The cross section circles are swept around the B-spline curve and the optimization method deforms the central vessel to the vessel wall.

Mille et al. [122] proposed a new parametric deformable model by extracting the centerline of a curve with varying radii. The method is a region-growing energy based method. An overview of the algorithm is that the user initializes two or more points on the surface then the minimum path or the geodesic path is deformed until the central B-spline curve with its control points is extracted then a circular cross sections are swept along the axis which is used as the initialization of the wall model and the vessel wall is optimized using an optimization method to fit the vessel wall. The image features for optimization are based on the Hessian matrix.

2 Level Set Approach

Level set technique is an evolution method which was introduced by Osher and Sethian in 1988 [128]. The level set evolution describes the movement of the surface by mapping it to a zero level set of a higher dimensional function. One important advantage is that it is easy to build accurate numerical schemes to approximate the equations of motion of the curve.

Level set method is a numerically useful technique for general image segmentation and has found numerous applications to vascular image analysis. Parameterization of complex curves and surfaces is in general not possible - the level set helps to do the computation of the curves and surfaces on a Cartesian grid and provides the capability for the topology to change. It is also a great tool for

modelling time-varying objects.

The level set function which refer to as has a zero level set corresponds to the boundary of the shapes in the upper row. Furthermore,ϕ takes on positive values in the interior of the shapes and negative values in the exterior of the shapes. The middle shape in the upper row shows when the topology of the shape is changing; i.e., the two parts are splitting. If looking at the second row, one can see that it is quite easy to track the shape with the level set function. Γ is the close curve and ϕ is the auxiliary function which is called the level set function. In order to represent the bounded curve with a level set function, it needs to have:

$$\Gamma = (x, y)|\varphi(x, y) = 0 \tag{5}$$

Γ is the zero level-set of φ; φ is assumed to be positive inside the Γ region and negative outside of it. If the curve Γ moves in the normal direction with speed F then the level set function φ satisfies the level set equation: $\frac{\delta\varphi}{\delta t} = F|\nabla\varphi|$ where t represents time. There are several numerical solutions for propagating the zero-level set front, most important of which is an upwind difference scheme with a narrow-band implementation to speed up computations.

This section studies some papers that focus on segmenting vessels utilizing the level set technique. Reference [105] defines a level set vascular segmentation method for finding vessel boundaries in CTA images. Due to presence of iodinated contrast, in CTA images, vessels are brighter than the background. An estimate of background and vessel intensity distributions is made utilizing the intensity histogram which is used to lead the level set to the vessel boundaries. The point of minimum classification error represents the boundary between the vessel and the background. For a certain intensity value the function g_b and g_v describe the intensity values whether belongs to the background or the vessel. A speed function is needed in order to have a smooth movement on the boundary.

The speed function is defined as:

$$F_{im} = \frac{g_v - g_b}{g_v + g_b} with g_v(x) = \frac{1}{\sigma_v\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_v}{\sigma_v})^2}, and g_b(x) = \frac{1}{\sigma_b\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_b}{\sigma_b})^2} \tag{6}$$

g_b and g_v are the distribution of the vessel and the background which both are assumed to be Gaussian. Minimum classification error is exactly where the histogram distribution of the two classes intersect, which the speed function is defined to be zero. The final level-set partial differential equation which has to be solved is: $\varphi_t + F|\nabla\varphi| = 0, F = F_{ext}(c-), F_{ext}$ is an external term based on image feature, c is the constant which is chosen 1 in this paper, for vessel structures it is preferred smoothness along the longitudinal direction which the curvature term is minimal, therefore k is chosen $k_{min}, k = k_{min}$. After substitution the final partial differential equation is:

$$\varphi_t + F_{im}(1 - \varepsilon k_{min})|\nabla_\varphi| = 0 \tag{7}$$

εk_{min} is the weighted curvature which is responsible for keeping the surface smooth during evolution. The speed function has zero value at the optimal threshold and takes on positive values within the vessel and negative values within the background.

The CURVES technique [105] proposed by Lorigo et al. is a curve evolution approach which uses level sets to segment vessels. The method models the object boundary as a manifold and the curve evolution is achieved through energy minimization. Basically, the method is capable of evolving the 1D curve on 3D domain. A new energy term is defined which specifies the lowest curvature of the surface which is the assumed to be the principal vessel direction. The method is an extension of geodesic active contour which is mainly based on [77] reference. In CURVES the dimension of the manifold is one and its codimension is two (the codimension of a manifold is the difference between the dimension of the evolving space and dimension of the manifold). A manifold is a mathematical space that on a small enough scale resembles the Euclidean space of the same dimension, which is called the dimension of the manifold. For instance, any point on a two dimensional surface of a sphere is surrounded by a circular region which can be changed to a circular plane. Circle is the simplest example of a manifold which on a small scale is homomorphic to a line; both a circle and the line are one-dimensional manifolds.

Deformable models are applied to 3D vascular segmentation too. In such methods, the initial boundary is deformed iteratively and the energy function

depends both on image information and surface smoothness. Such an algorithm is called minimal surface. For evolving the curve in 3D the traditional level set equations does not hold. If the curve is in the plane and surface is three-dimensional; so the surface has co-dimension one and the curves in three dimension has co-dimension two. To extend the method to higher dimesions, the CURVES algorithm defines an auxiliary function which maps the 1D curve to 3D.

Instead of finding the closest point on the zero level set; it is defined as the nearest point on the zero level set. The difference between the previous methods and CURVES and the steps involved in the curve evolution is stated.

Rochery [142] proposes a new quadratic energy function which works even if there is occlusion in the network. Hence, if one part of the vessel image is occluded with other tissues, creating a gap between an elongated line or vessel, the specified energy function can connects the gap. The energy minimization uses the level set technique to evolve the curve for higher order active contour energy. Traditional energies are expressed as a single integral over the contour, the higher order active contour energy utilized multiple integrals, so the interaction is between different sets of contour points. For likelihood term they describe multi point interactions between the contour and the data. The forces derived are non local.

The force definition penalizes local gaps that are found between the contours and links them.

Some methods are recently proposed which are focused on variational formulation of flux maximization instead of the curve or surface [95, 96]. The flux maximization aims at aligning the surface normal to the gradient vector field. All these methods mainly based on vectorial information and it is proved to be able to detect the even low contrast and narrow vessels.

F Conclusions

Vascular segmentation is a very challenging research area. The method that is utilizing for vessel segmentation depends on the dataset. Accuracy, level of automation and computational efficiency are different parameters that should be

considered when a segmentation method is chosen. The combination of methods is the best solution for vessel segmentation. The dataset and its features and the information that are needed to extract should be considered to make the best decision of what type of methods should be chosen. In order to get a more accurate output different algorithms can be combine in a sequential order. A preprocessing can be used to enhance the image and improve the quality of the image.

Most of the vessel segmentation approaches substantially depend on initialization. The techniques like the region growing, ridge based and skeleton based techniques needs a start and end point or some auxiliary points on the path defining manually. In addition, the active contour methods need an initialization contour in a right place and near the desired contour. The method might find the local minimal and cannot converge correctly. Therefore, a pre-segmentation process can be utilized in order to have a better segmentation result. The purpose of the vessel segmentation is very critical in order to choose the best combination of methods. The automation and accuracy are two main features considered in vessel segmentation.

CHAPTER III

CLASSIFICATION TECHNIQUE: A MULTI-SCALE NON-LINEAR VESSEL ENHANCEMENT TECHNIQUE

A new adaptive segmentation technique is developed on brain MR images and the technique is coupled statistical clustering technique with the non-linear diffusion filter. The method is published in [6]. An enhancement method is represented which is based on nonlinear diffusion filter and statistical intensity approaches for smoothing and extracting 3D vascular system from Magnetic Resonance Angiography (MRA) data. The method distinguishes and enhances the vessels from the other embedded tissues. The Expectation Maximization (EM) technique is employed with non-linear diffusion in order to find the optimal contrast for enhancing vessels; therefore, smoothing while dimming the embedded tissues around the vessels and brightening the vessels. The non-linear diffusion filter smooths the homogeneous regions while preserving edges. The EM technique finds the optimal statistical parameters based on the probability distribution of the classes to discriminate the tissues in the image. The proposed enhancement technique has been applied to four 3D MRA-TOF datasets consisting of around 300 images and has been compared to the regularized Perona and Malik filter. The experimental results show that the proposed method enhances the image, keeping only the vessels while eliminating the signal from other tissues. In comparison, the conventional non-linear diffusion filter keeps unwanted tissues in addition to the vessels [6].

A Introduction

Vascular diseases are among the most significant causes of death in the world. An enhanced three dimensional visualization of blood vessels could help in

diagnosing the disease and choosing appropriate treatment. The parameters that make extraction of blood vessels challenging include anatomical variability of the vasculature, surrounding tissues, image contrast, resolution, and noise. In many approaches to analysis, the preprocessing step enhances the vessels and improves its visualization; assisting the task of segmentation and centerline extraction. Scale space theory can be utilized for smoothing and enhancing medical images. In scale space theory a set of smooth images are generated by employing the diffusion equation; the original image is the initial condition of the function. The diffusion function is specified either as scalar or tensor based. The original idea of image diffusion for image filtering was proposed by Perona and Malik and was based on a scalar function; it was proposed as a solution to edge detection [134]. Weickert added orientation to enhance small vessels and coherence structure [183, 182].

Subsequent methods to conventional diffusion filtering replaced the diffusion tensor by the Hessian. Multi-scale vessel enhancing methods based on eigenvalues of the Hessian matrix typically determine the vesselness of a pixel. Different geometric interpretation extracted by the eigenvalue system of the Hessian matrix is used to measure the vesselness [81, 104, 148, 46, 24]. For review of anisotropic diffusion, please refer to [164, 181, 39]. For implementation review of tensor based diffusion filters in ITK refer to [38]. Catte [22] and Yu and Accton [191] applied a new filter as an edge detection method on Ultrasound images, only considering speckle noise in the image. Frangi [46] proposed the multi-scale enhancing method based on Hessian and tensor structure. Weickert [182] defined the coherence-enhancing diffusion which improves the tensor based diffusion to find divided regions and to connect them. Krissian [86] and Manniesing [113] proposed anisotropic diffusion filters to segment vessels in 3D, based on a tensor structure filter. Fischl proposed a new method to indicate the best kernel function that matches the image [42]. For further review on vessel analysis the reader is referred to [97]. The proposed method uses the scalar diffusion function, and is mainly based on the conventional Perona and Malik nonlinear diffusion filter. Vessels constitute a small area within each slice. They are surrounded by other tissues and are thin and small. Nonlinear diffusion

filtering enhances the regions while preserving the edges, but it cannot distinguish the homogenous vessel region from other tissues. A new method is proposed to enhance the vessel structure which employs the conventional Perona and Malik non-linear diffusion filter while making use of the Expectation Maximization (EM) algorithm [32]. EM is an optimization method estimating statistical parameters.

It is an iterative method and discriminates the classes which are defined based on their probability density function (PDF). Vessel class is discriminated further in each iteration, and the difference between the contrast of the vessel and the other tissues is increased in every iteration of the smoothing process. The smoothing changes are adaptive because the contrast of the image is changed adaptively. The experimental results demonstrate that the proposed method improves vessel enhancement when compared to the conventional non-linear diffusion filter. In addition, the proposed method is a new technique that finds the constant gradient threshold of the diffusion function adaptively. In next sections the Perona and Malik diffusion filter and the EM algorithm are introduced. Then the proposed method is described and shows experimental results. Finally, conclusions are given.

B Non-Linear Diffusion Filter

Foremost, Perona and Malik [134] proposed the idea of the non-linear diffusion filter. Their proposed diffusion functions are mainly based on the gradient operator to limit smoothing across edges and regions. The generic definition of the diffusion function is indicated as:

$$\frac{\delta I}{\delta t} = div\left[c\left(|\nabla I|\right) \cdot \nabla I\right] \tag{8}$$

I is the image, ∇I is the Gradient of the image, $c\left(|\nabla I|\right.$ is the diffusion function which controls the smoothness of homogeneous regions and preserves the edges of the regions. c is one in the interior of the regions and zero on the boundaries. $c \cdot \nabla I$ is called the flow function, the greatest flow happens when the gradient magnitude

is close to the threshold. Two diffusion equations were proposed in [134]:

$$c\left(|\nabla I|\right) = \frac{1}{1 + (\frac{|\nabla I|}{k})^2} \tag{9}$$

$$c\left(|\nabla I|\right) = exp\left[-\left(\frac{|\nabla I|}{k^2}\right)\right] \tag{10}$$

If $|\nabla I| \gg k$ then $c(|\nabla I|) = 0$, if $|\nabla I| \ll k$ then $c(|\nabla I|) = 1$. The parameter k is a constant and is a threshold for choosing the smoothing value. If k is chosen to be a large number then the homogeneous regions are smoothed to a greater extent. There is still no automatic solution to finding the k, on the other hand choosing an appropriate k is critical to implementation results. The effect of different $k\,'s$ is shown on one slice of the MRA images in Figure 2.

As shown in Figure 2 the chosen should not be very small or very large in given dataset, when the k is chosen 10; the regions are smoothed and the edges are kept; however, Figure 2 shows a noisy image which is the result for k=30. Therefore, the gradient threshold is vital to implementation results; in the proposed approach, it is substituted with the best threshold that is calculated from the EM algorithm. It is not needed to test different thresholds on a dataset in order to find the best setting.

C Probabilistic Model for MRA-TOF

Statistical approaches play an important role in extracting regions of the image. Three classes are defined: vessels, the background and the other tissues. The total probability distribution of the Gaussian mixture model is:

$$p = w_1 \cdot P(q|vessel) + w_2 \cdot P(q|background) + w_3 \cdot P(q|othertissue) \tag{11}$$

where q is the given data or intensity level and $P(q|anyclass)$ is the probability density function for each class. Figure 4 and Figure 5 show the Gaussian mixture model for the classes. w_s are the proportion of each class in the image and their summation should be one.

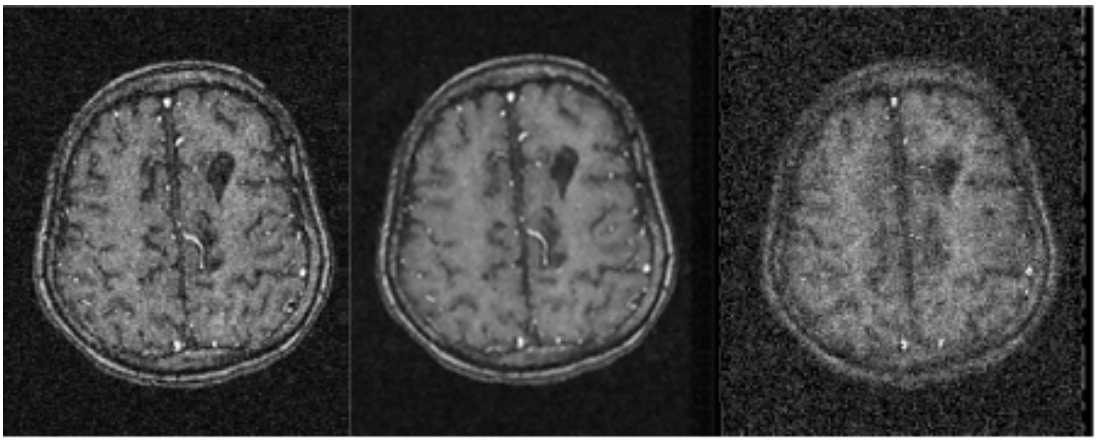

Figure 2. The results of Perona and Malik non-linear diffusion filter after 10 iterations from left to right a) Original Image b) k =10 c) k=30 [6].

EM is an optimization method for estimating parameters. The probabilistic model of the incomplete data is available. In the given domain the observed data is the intensity of the image. Labels, indicating whether the pixels belong to the vessel or the background, are unobserved information. Mean and variance of the Gaussian distribution are the statistical parameters EM estimates. The EM algorithm definition is as follows: Let X be the incomplete data, Y the complete data, and the parameter vector. At the initialization stage, it is assumed to have a value not very far from the final answer; the algorithm updates vector until changes are very small; $argmaxQ$ can be any mathematical method that finds the maximum of Q. In the paper the maximum likelihood is utilized to maximize the parameters. The Expectation and Maximization steps make the changes and update the feature vector.

$$Expectation - Step : Q(\theta/\theta(k)) = E[lnP((Y/\theta)/Y, \theta(k)] \tag{12}$$

$$Maximization - Step : \theta(k+1) = argmaxQ(\theta/\theta(k)) \tag{13}$$

$\theta(k)$ is the parameter vector in k-th iteration. $P(Y/\theta)$ is the probability density function of classes. The chosen parameter vector is the mean and variance of each class in the image. The proposed application has three classes, so the mixture model was used and wit the assumption that all the classes are independent variables. $\theta(k)$ is the parameter vector in k-th iteration. $P(Y/\theta)$ is the probability density function of classes. The chosen parameter vector is the mean and variance of each class in the image. The application has three classes, so the mixture model is used with the assumption that all the classes are independent variables.

D Proposed Method-Combination of Non-Linear Diffusion Filter and EM Algorithm

Vessels are thin, have weak edges, and are surrounded by the other tissues. Conventional non-linear diffusion filter smooths different regions; however cannot enhance vessels from other organs. Hence, a combination of the non-linear diffusion

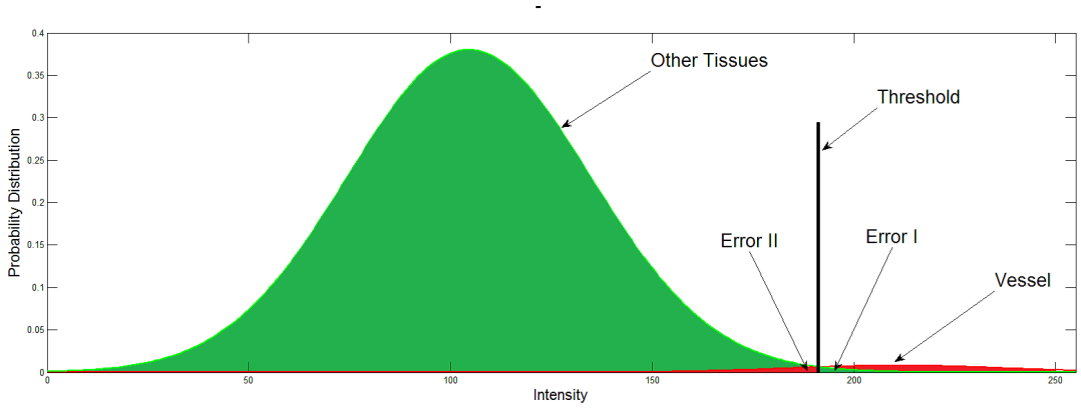

Figure 3. The probability distribution of two classes, ErrorI and ErrorII are the error regions between two classes[6].

filter and the EM algorithm is utilized. Three classes for the given dataset are defined: vessel, background, and other tissues. The iterative framework smooths every image slice while EM pulls out the optimized separated distribution of the three classes. The error area from the overlapping regions of the vessel class and the other classes is found.

The inverse of this value is the adaptive contrast, which is added to the vessel regions and subtracted from other regions, to intensify vessels in the background in each iteration. The risk function between the vessel class and the other two classes is calculated in order to find the adaptive contrast. Figure4 shows the risk area between two different classes. The area of ErrorI and ErrorII is the overlap error area between the two classes. An adaptive contrast based on risk function is indicated:

$$\tau = \frac{1}{error_1 + error_2} \tag{14}$$

$error_1$ and $error_2$ are the error areas between the vessels and the other classes distribution in 3D dataset. tau is the adaptive contrast which is extracted based on the risk function, and it is utilized to enhance the images. The adaptive contrast is extracted from all the slices of the dataset; and the nonlinear diffusion filter is applied on each slice individually while using the 3D neighborhood information to smooth the regions. The non-linear diffusion filter and image enhancement are applied simultaneously on each slice. After each iteration, the intensity level of all the images are changed. In the next iteration, the EM has to find the optimized threshold for updated images.

Figure 4 shows the probability distribution function of the three classes in the first iteration for one 3D MRA-TOF dataset. The right most marginal Gaussian distribution belongs to the vessel class whose intensity level is high and it constitutes only a small part of the whole image. Adding a threshold to the vessels in each iteration causes increased contrast and further difference between the vessels and the other tissues. The distribution and its parameters are computed in 3D so only one figure for all the slices is generated. The initialization is based on manual

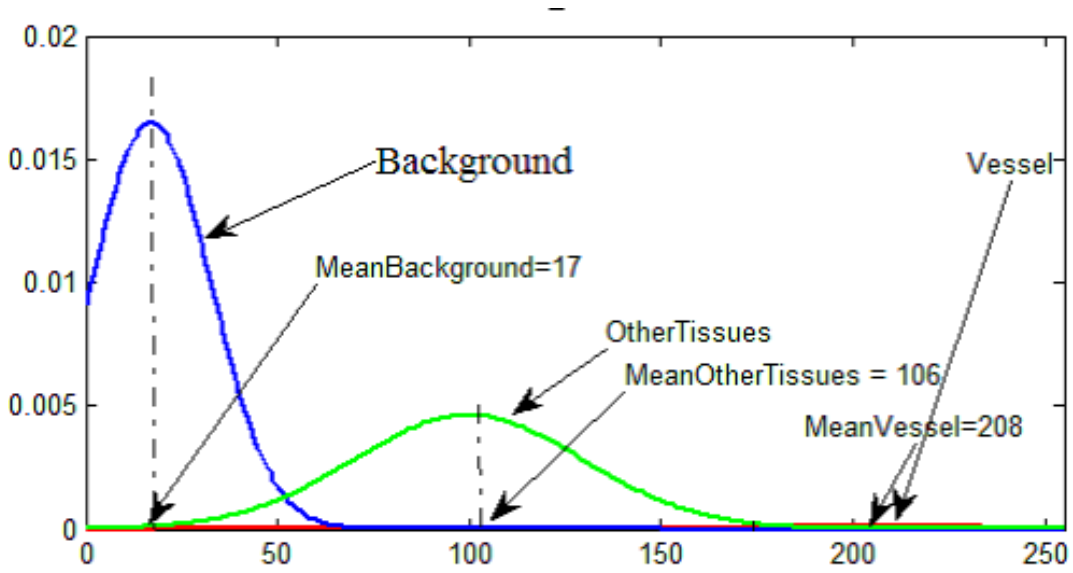

Figure 4. The initial Probability Density Function of the three classes (Background, Vessel and Other Tissues) [6].

sampling of each class. The background is so dark and the vessels are the most bright regions in each slice and other tissues have the intensity between the two above classes. The initial probability for each class is chosen equal at the initialization step. Figures 4, 5 shows the distribution of the classes after applying the proposed technique; the intensity level of the surrounded tissues is decreased. Consequently, the whole image is darkens while the vessels are distinguished and enlightened in the background.

The diffusion function is calculated for each pixel utilizing the 3D neighbors.

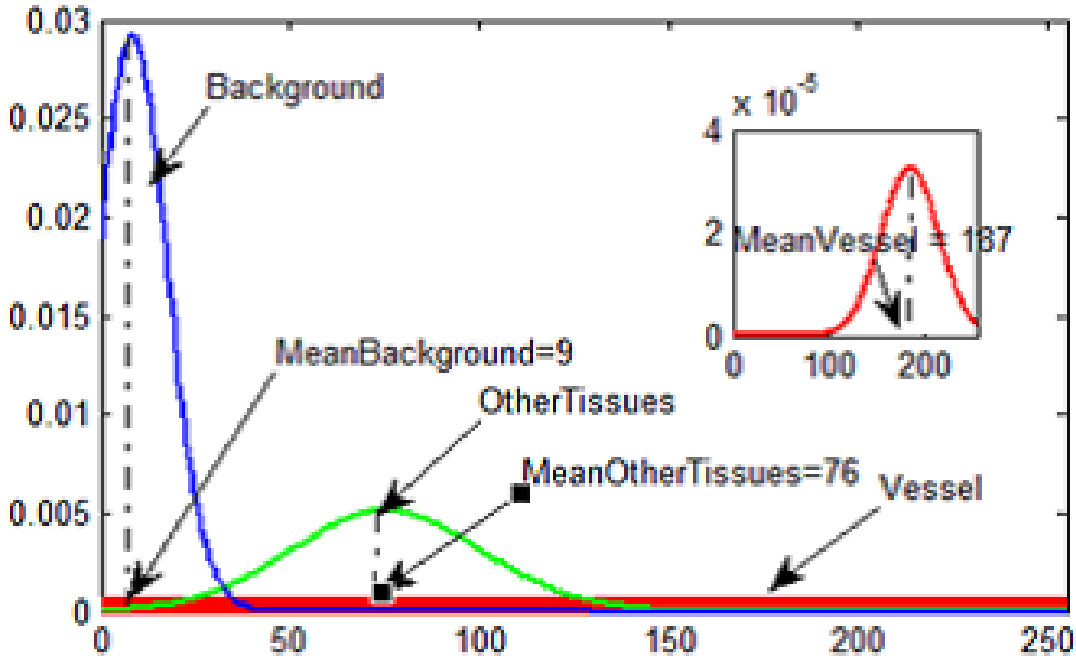

Figure 5. The final Probability Density Function of the three classes, vessels are brightened while the mean of the other two tissues in the left are darkened [6].

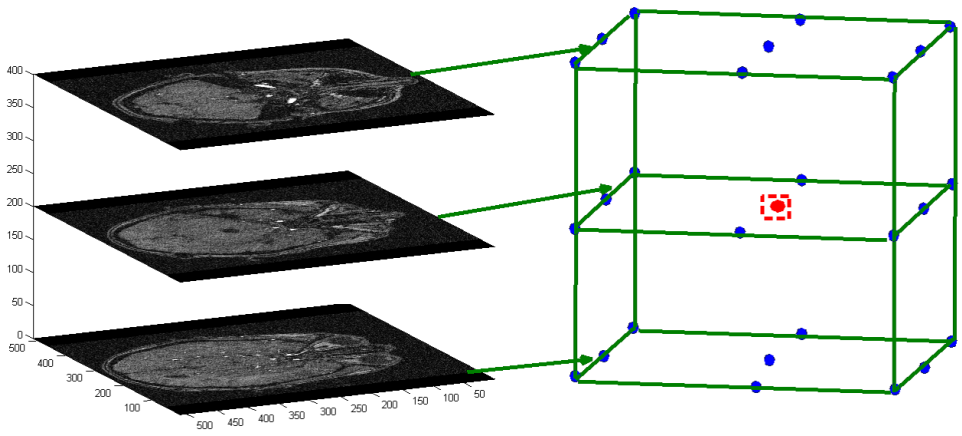

Figure 6. The 3D neighbourhood of each pixel has 26 pixels which are shown for one pixel.[6]

Hence, a cubic grid for the pixels neighbours is defined. 3D weighted neighbourhood pixels intensity in utilized in the non-linear diffusion function. The new intensity level for each pixel is indicated as:

The implementation process has the following steps:

1. 3D filtering utilizing the non-linear diffusion filter.

2. Applying the EM classifier and computing three dimensional Probability Density Function.

3. Finding the adaptive contrast threshold.

4. The result of step 3 is added to the vessels regions and subtracted from other regions in all the slices [5].

5. If the changes between the updated image and the previous image is smaller than a threshold then stops, else repeat the process from step 1.

E Experimental Results

The proposed method is applied on 4-datasets consisting of around 300 MRA-TOF image slices. The classifier uses all the slices and computes the adaptive contrast threshold based on all the image slices. The smoothness and intensity of the entire dataset affects each slice. The results are compared with conventional non-linear diffusion filter. Figure 7 shows the results of the proposed method and conventional non-linear diffusion function. The zoom and scaled version of only one slice of the image which contains vessel shows that the proposed method successfully excludes the vessel from its neighbors region. Figure 8 shows the output of the proposed method and the non-linear diffusion filter and the binary image of ground truth on one slice. The manual segmentation for all the slices is also available to us. Comparing to the binarized image of the ground truth, it is clear that the final enhanced image for the proposed is very similar to the ground truth; the non-linear diffusion filter however keeps more tissues and the vessels are not enhanced and segmented clearly. Figure 9 shows the implementation results on 3 slices from a dataset with 93 slices. The first row contains three slices of one of the datasets before applying the proposed method which the vessels are embedded in surrounded tissues. The second row shows the enhanced images, so the proposed method preserves the vessels and darkens other tissues.

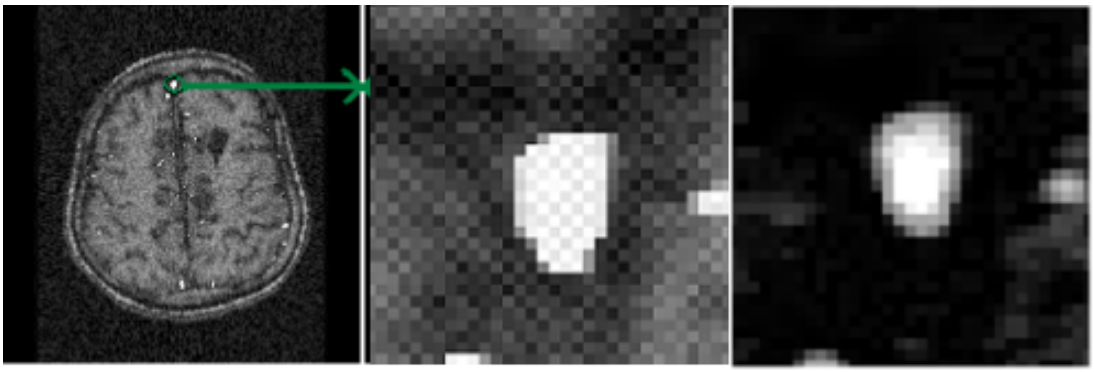

Figure 7. Scaled image of one part of the vessels in one slice, comparison of conventional non-linear diffusion filter and the proposed method. The method completely enhances the vessel with respect to the surrounding tissues. a) Original image, shows the vessel in a green circle. b) Scaled image, conventional non-linear diffusion filter result. c) Scaled image, the proposed methods result [5].

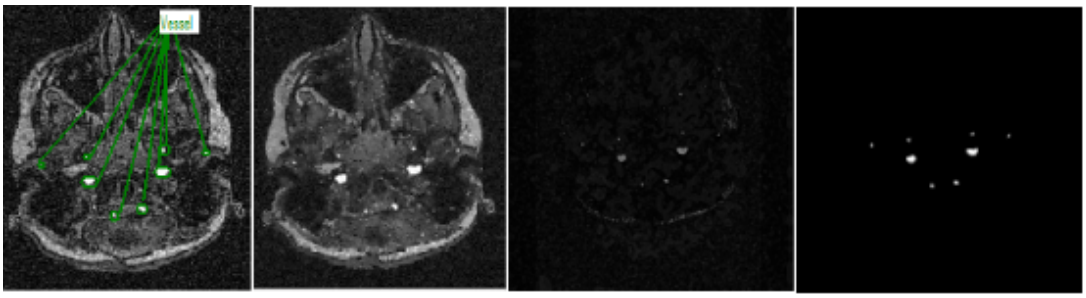

Figure 8. The final result of vessel regions on one slice. a) Original Image b) Non-linear diffusion filter c) The proposed method d) Binary ground truth [6].

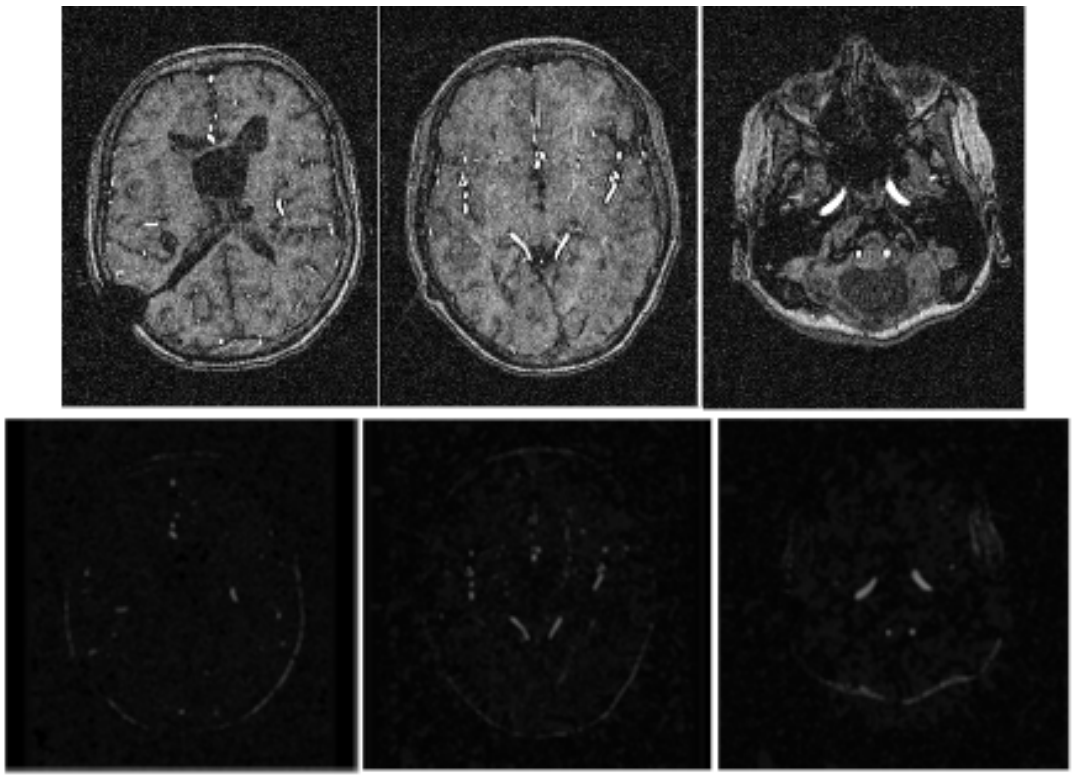

Figure 9. Original and enhanced images of a dataset containing 93 slices. First row shows images before applying the proposed method. Second row shows enhanced images based on the proposed method. From left to right are: slice 22, slice 44 and slice 90.[6].

TABLE 1. Evaluation criterion [6]

Measure	Nonlinear Diffusion Filter	Proposed Method
τ	0.4432	0.7217

F Evaluation Metric

The metric that is used for evaluation is a criterion defined in [24]:

$$\tau = (V_{Ground} \cap V_{ourmethod})^2 = (V_{Ground} \times V_{ourmethod}) \tag{15}$$

V_{Ground} is the volume of the vessel from the ground truth and $V_{ourmethod}$ is the volume of vessel obtained with the proposed method. τ is between zero and one. A τ of one indicates perfect segmentation. Table 1 shows the that is calculated for the proposed method and conventional non-linear diffusion filter. τ for the proposed method is higher than the for the conventional non-linear diffusion filter. Finally, each slice of the image is binarized and then the final result of all the slices in 3D is visualized. Figure 10 shows the final result of 3D visualization of the method. Figure10a is the 3D visualization of non-linear diffusion filter segmentation output; and it detects other objects as vessel. Figure 10b is the 3D segmentation of the method and it shows that the proposed technique distinguishes the vessels and extracts them from the surrounded tissues.

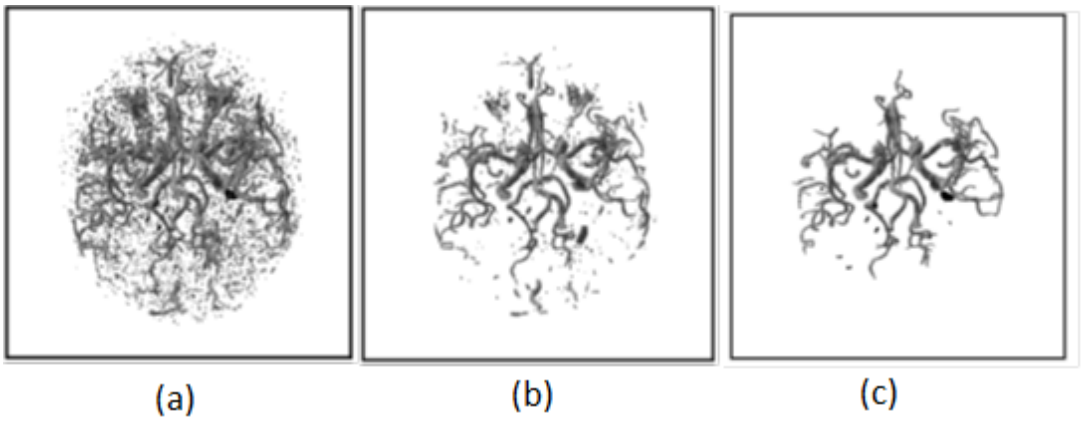

Figure 10. The 3D final result of nonlinear diffusion filter, the proposed method and ground truth. a) 3D visualization of Nonlinear-diffusion filter b)3D visualization of the proposed method c) is the ground truth 3D result [6].

CHAPTER IV

COUPLING TUMOR MODELLING WITH IMAGE ANALYSIS

In this section the goal was to use machine learning techniques in order to find a way to use multi-modal information such as medical images and mathematical tumor growing models and using machine learning to predict the parameters of the mathematical tumor growing models. Finding a large number of training set was not possible, so the idea is published as a survey in the book chapter [5]. It is difficult to extract values for the tumor model parameters from the sparse data available for any particular patient. Medical image analysis can measure the shape, size, volume and placement of tumors from MR and CT images for individual patients, yet these techniques are limited. For instance, the threshold for cell detection is a density of 8000 cells/mm3 in MRI, which may miss a significant number of active tumor cells and thus potentially lead to inaccurate prognoses [117, 61]. A set of methods by Konukoglu et al. [61, 83] is reviewed which integrates mathematical modelling of tumor growth with patient-specific medical images, with the goal to offer disease development modelling.

Typically, reaction-diffusion equations model tumor growth at the tissue-scale contain terms that describe the change in cells in space and time, and their collective proliferation rate. The local diffusion of the cells is defined as a tensor in the calculations. A typical differential equation may take the form [61]:

$$\frac{\delta u}{\delta t} = \nabla \cdot (D(x)\nabla u) + \rho u(1-u), where D\nabla u \cdot n_{\delta\Omega} \tag{16}$$

where u is the tumor cell density, D is the local diffusion tensor, ρ is the proliferation rate, and Ω is the boundary of the domain tissue, which in most models that have incorporated imaging data has been the brain. The diffusion term of the tumor cells

is $\nabla \cdot (D(x)u)$ and the reaction term is $\rho u(1-u)$ [61]. The tumor cell density observed clinically is linked to the reaction-diffusion model by defining a density function based on the image intensity of the lesion [160]. Although the parameter estimation has focused mainly at brain tissue because of image availability and easier tumor identification, the modelling concepts apply generally to solid tumors.

The main challenge of integrating this type of reaction-diffusion model with imaging data is that the model describes the evolution of tumor cell densities in time, while in the image sequences only the shape of the tumor in space is observed

The diffusion tensor and the reaction parameters are estimated from the medical images, meaning that the evolution of the tumor equation can be specified for each individual patient. To illustrate this process, the case of tumors in the brain is considered in more detail, for which extensive modelling work has been done (e.g., Hogea et al. [61], and Swanson and coworkers [160, 161]). These methods usually assume that the velocity of tumor growth differs in different types of tissue (e.g., white and grey matter), so different diffusion tensors are defined based on the location of the tumor. The diffusion tensor for brain tissue is defined as:

$$D(x) = d_g I, x \in graymatter \tag{17}$$

$$D(x) = d_w D_{water}, x \in whitematter \tag{18}$$

whereas tumor cells are modelled to diffuse isotropically in the gray matter, the diffusion in the white matter is proportional to the diffusion tensor of water. Tumor cells diffuse isotropically in grey matter with rate d_g, d_w is the diffusion rate in white matter, and D_water is the diffusion tensor of water molecules [69].Medical images provide data to estimate the tumor growth parameters for individual patients: the velocity of the tumor growth (v), the diffusion of the tumor (D), and proliferation rate (ρ). Here, different mathematical relation between these three parameters are summarized. The calculated parameters are based on an assumption that the tumor margin evolves linearly in time [111]. One possible linear relation is defined as: $v^2/4\rho$, which uses Fishers Thr. The diffusion coefficients in white and gray matters are, respectively, $D_g = v_g^2/4\rho$ and $D_w = v_w^2/4\rho$ [60]. The tumor margin in image

sequences approximates the velocity rate.

Another mathematical estimation is stated as $v = 2\sqrt{\rho D}$ [83, 162]. The tumor margin advances as a travelling wave, which expands radially and linearly, and the diffusion coefficient changes centrifugally. If T1 and T2 weighted images are available, then the gradient between these two can be defined as the ratio of diffusion over proliferation [162], where the tumor margin is detected from T1 weighted mages and the edema is detected from T2 weighted images [162]. The gradient has also been defined as $v = 4\sqrt{D\rho}$ [111, 162] delineates the kinetics of the tumor growth; simulations have shown that D/ρ can indicate the spatial extent of nonvisible tumor tissue [111, 162]. The results show that utilizing $\sqrt{D/\rho}$ instead of $D\rho$ may reflect the tumor growth rate more accurately [111, 162].

Another method defines a bio-physical reaction diffusion function while adding a mechanical advection term [88]. For individual patients the parameters of the tumor growth are estimated from available image sequences. The mechanical advection term translates the elasticity of the tissue through which the tumor cells diffuse. This model employs different velocities depending on the tumor location; however, the unavailability of serial scans of the lesion precludes the measurement of precise parameter values. The model constraints can be defined in such a way that the problem becomes an optimization exercise with new parameters. The very first scan where the tumor is observed is defined at $t = 0$, and the diffusivity and elastic material coefficients are the new model parameters.

Parameters (e.g., diffusion, velocity and tumor proliferation) extracted from images through these techniques have been used in modelling the tumor evolution in time and space (spatial-time models). Jbadi et al. [69] modeled the diffusion of tumor cells in anisotropic tissue. They proposed a new definition rate for the diffusion tensor in water, based on calculating the highest eigenvalue of the tensor of water molecules at each point. Another method considers a probabilistic approach [160]. The tumor growth evolution $(\rho\,(u(t)|\theta_x, \theta_t, \theta_p))$ is a conditional probability where tumor growth parameters describing time, location, diffusion and proliferation rate are approximated. θ_x is the tumor location parameter, θ_t is the

parameter change in time and θ_p is the personalized parameter: diffusivity and proliferation rate. These parameters are defined based on image sequences.

Some of the modeling work focuses on matching the spatial-time evolution predicted by the model with the known tumor cell density from series of scans that have been prepared independently. The object is to minimize the difference between the estimated tumor cell density calculated from the model with the given tumor cell density from a particular subject [161]. A recent method proposes a modified anisotropic model which models the tumor delineation considering the curved front and the effect of time in its speed [61].

Spatial-time tumor growth models have mainly considered avascularized tumors whereas it is vascularized tumors that are the most dangerous. Further, the extent of tumor vascularization may affect the chosen treatment. Yet informing the model parameters from vascular imaging information is challenging due to the problem of vessel segmentation. Vessels can be visible in MR and CT images; they usually appear brighter in CTA (Computed Tomography Angiography) and MRA (Magnetic Resonance Angiography) images taken with contrast agents. In general, automatic segmentation vessel trees entail two main steps: extracting features from image slices, and then reconstructing the 3D model of the vessels. Even if the appearance of vessel features is accurately extracted from the images, the 3D reconstruction of curvature is complex: number of vessel branches, curvature shape of the vessel, and numerous other factors affect the accuracy of the segmentation in 3D [6]. Vessels connect to tumors with infinite possibilities: the appearance of vessel branches is different for each individual patient, so one cannot define a predefined model to be able to quantify this information.

CHAPTER V

FEATURE EXTRACTION TECHNIQUE-ANALYZING PERFUSION CURVES FOR AUTOMATIC DETECTION OF TUMOR VASCULARIZATION AND LUNG TRANSPLANT PREDICTION

In this chapter a new feature extraction technique is developed and the technique is published in [175, 75]. Perfusion imaging measures blood volume in tissues; a relation between volume and histological features may be assumed. Perfusion imaging is typically used as a method for determining prognosis in the clinic [100]. In research, imaging of perfusion through a tissue has been used to measure vascular geometry and histological features of tumor angiogenesis, and also to estimate micro-vascular flow through capillaries and venules [93]. Measurements of perfusion flow can provide intravascular blood volume (reflecting the MVD) and mean transit time of blood through the tissue. Some studies show a potential correlation between perfusion imaging and MVD [163, 74, 135, 175], but others did not observe such a correlation [159]. In a clinical study of lung carcinoma angiogenesis using contrast-enhanced dynamic CT images, VEGF and MVD were correlated with maximum values of time attenuation curves instead of perfusion images [135].

Heterogeneities in the perfusion of solid tumors prevent optimal delivery of nano-therapeutics. Clinical imaging protocols to obtain patient-specific data have proven difficult to implement. It is challenging to determine which perfusion features hold greater prognostic value and to relate measurements to vessel structure and function.

Tumor vasculature is characterized by structural abnormalities that produce

spatial and temporal heterogeneities in blood flow. The vasculature lacks a regular hierarchical network of large proximal vessels feeding into successively smaller vessels; instead, vessel interconnections are irregular in size and spacing [159, 66]. Endothelial cells lining these vessels have altered morphology, pericytes (cells that support endothelial cells) are poorly attached or absent, and the basement membrane is often abnormal. The resultant vessels are dilated, tortuous, vascular, and vulnerable to collapse [66, 174]. The presence of fenestrations [65, 141, 54] combined with incomplete vascular walls [54], can yield localized regions of blood plasma leakage that alter macromolecule transport [114, 110], and increase interstitial pressure [40]. Collectively, these vascular abnormalities lead to regions of tumor tissue that are perfused poorly, intermittently, or not at all[68, 137].

Tumor perfusion is still poorly understood, particularly with respect to what conditions lead to effective or poor treatment. Attempts to characterize tumor perfusion using static data, such as the measurement of microvessel density from patient biopsies, have shown mixed prognostic capacity[59, 34]. Clinical imaging modalities capable of monitoring perfusion dynamically, such as MRI [51], CT[176], PET [132] and Doppler sonography [120], have been used to produce time-series images that enable pixel-by-pixel analysis of contrast kinetics within tumors. Parameters measured from the resultant time-signal curves are placed into pharmacokinetic (PK) models in order to extrapolate information regarding vascular anatomy and physiology. Principal features derived using PK models include the blood flow velocity, blood volume, and mean transit time. Numerous methods have been proposed to extract these features in human tissues [173, 58, 152, 64, 90, 28, 41, 79, 171]. While the prognostic capacity of such an approach remains to be determined, MRI [121], CT [82], and PET [87] have demonstrated that tumor transport plays a role in treatment response, and that persistence of unfavorable perfusion characteristics (high blood volume fraction, rapid transit time, focal hyperpermeability, and/or high FDG metabolism) following chemotherapy correlates with a poor treatment response. Due to the difficulties of relating clinical perfusion imaging with underlying tumor structure and function,

intravital microscopy (IVM) studies in live animals are becoming increasingly popular [57, 78]. Using video-rate laser-scanning microscopy, blood flow velocity, flux, and hematocrit can be measured by tracking trajectories of fluorescent red blood cells (RBCs) [140]. Concomitant injection of a fluorescent tracer allows measurement of shear rate [71], blood volume fraction [147], and tissue permeability [147, 35, 170, 67, 71]. These physiological parameters can be related to local variations in gene expression, enzyme activity, pH, metabolites, and other parameters of interest (reviewed in [78]) by simultaneously imaging multiple fluorescent reporters. A major advantage of IVM is that tumor perfusion can be characterized on a vessel-by-vessel basis, potentially leading to insights into how local variations in perfusion can affect nanotherapeutics delivery and treatment response [76].

In the current research, a theoretical framework for automated evaluation of IVM perfusion curves is described in order to model the delivery of nanotherapeutics. The hypothesis is that tumor-specific perfusion features may be used to model nanotherapeutics accumulation; thus, this framework aims to transcend the challenges posed by the typically abnormal tumor vasculature. Primary tumor fragments, collected from triple-negative breast cancer patients and grown as xenografts in mice, were injected with a bolus of 40kDa FITC-dextran tracer and monitored at 30 fps using IVM. The fluorescence intensity of each vessel was measured over time to yield a heterogeneous set of arterial and venous perfusion curves on a tumor-by-tumor basis. Two features were considered: the time to arterial peak and the venous delay, which acted as inputs for a Fuzzy C-Mean (FCM) classifier. The data was classified into three defined groups (poorly vascularized, well vascularized, and in between vascularized), which were correlated to experimental nanoparticle accumulation measurements. This approach enables an automated ranking of tumor vascular perfusion in order to model the delivery of nanotherapeutics. Using an independent validation set, it is demonstrated that new samples can be mapped into the feature space to determine their perfusion ranking and hence estimate their nanoparticle retention. A major strength of this approach

is that it enables the ranking of tumors and evaluation of their behavior in an automated manner without requiring PK models.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a non-invasive imaging technique that has been explored in perfusion-related concerns in many clinical applications, e.g., in evaluation of the kidney, brain and heart. At short times (up to about 2 minutes) after administration at DCE-MRI, parameters can be derived that reflect the agent delivery to the tissue bed. In this study a novel and automated comprehensive framework for the non-invasive classification from 2D DCE-MRI of non-rejection and acute rejection transplants is evaluated. Recently, a method is proposed for the automatic classification of normal and acute rejection transplants from 2D DCE-MRI, consisting of four steps: kidney segmentation, non-rigid registration to align the object, cortex segmentation, and classification of normal and acute rejection transplants by evaluation of perfusion curves [19].

A function-based model is used to analytically classify the perfusion or TICs of renal transplant patients in order to determine their prognosis in terms of transplant acceptance or rejection. These curves quantify the average intensity of renal perfusion for up to four minutes. First identifying a model function that can be consistently fitted to all the TICs. Then, the output of the model is classified into two groups, namely, non-rejection (successful) or rejection (unsuccessful) transplants. After selecting the features of the function classes and training the data classifier, this classifier is used to classify new (unknown transplant outcome) curves. The pioneering work of Larson, Tofts and Brix enabled the modeling of tracer kinetics using DCE-MRI [93, 93, 106, 92, 167, 155]. Advances in MRI technology in later years enabled models to estimate perfusion and capillary permeability more accurately. Recent studies have focused on revealing physiological characteristics of the tissues. The main idea is to extract a relation between the perfusion and vascular functionality of the tissue, which enables measuring blood volume and capillary permeability. Newer models have focused on extracting these critical tissue features [106].

Mathematical tumor modeling describing the evolution of tumor mass in time

(recent reviews [49, 21, 13, 53, 11, 31, 179, 150] and are coupled with biological data to represent tumor growth and treatment responses. Tumor vascularization is an important stage in tumor evolution. Tumor vascularization, metastasis and mathematical modelling developed to simulate these processes are reviewed in the published chapter. Here the research is mainly based on weakness and strength of mathematical tumor modelling and medical image analysis and how to combine these two groups.

Discrete models are useful for studying cell-cell and cell-microenvironment interactions, natural selection carcinogenesis, and genetic instability. Continuum models simulates as a collection of tissues, employing principles from continuum mechanics to describe cancer-related variables as continuous fields using partial differential and integro-differential equations [49]. In contrast to discrete and continuum models, hybrid approaches utilize both the continuum and discrete representations of tumor cells and microenvironment components. Discrete models are representing tumor evolution in cell-scale while continuum models illustrate the evolution in tissue-scale, the hybrid model upscale the cell-scale to inform the phenomenological parameters of models at the tissue scale. Recent reviews on mathematical tumor modelling include [49, 13, 53, 11, 31, 179, 150, 144, 129]. Besides the avascular growth, models have focused on tumor-induced angiogenesis [192, 107, 139], metastasis [126], intra-cellular pathway, intra-cellular pathways [119], stem cells [37] and treatment [98, 177].

Medical image analysis is mostly used as a tool for patient screening. Cancer screening has the potential to allow for early detection. Medical image contain lots of information that radiologist need to consider many parameters to extract the disease information and also making the treatment decision [12, 136]. Spatial reconstruction of a specific lesion depends on many factors, including image resolution, contrast level of tissue, appearance of very small cells in images, the thickness and available number of slices, the accuracy of the applied method for lesion reconstruction [38]. These factors prevent defining a uniform and robust technique for lesion extraction. Imaging techniques cannot accurately localize tumor

cells, underestimation or overestimation is another problem even if the image analysis technique is chosen correctly and it is based on the ground.

The weakness and strength of mathematical tumor growth models and medical image analysis, one of the ideal idea is to combine these two methods. The biological parameters like the diffusion term or velocity of the tumor growth could be defined based on image series. The main advantage of this combination is personalization of tumor evolution. Tumor growth factors are extracted for each patient based on the location of the tumor and its surrounded tissues and also the tumor growth rate might be different for each patient. Image series are one the input information that is given for each individual patient.

The organ under consideration in medical images has a critical role in choosing the method to analysis medical images.

Medical image analysis is useful in providing observable clinical information for all organs in the body. The research was focused on reconstruction of blood vessels. The parameters that make the extraction of blood vessels challenging are anatomical variability of the vasculature, location of the blood vessel, image contrast, resolution an also the imaging modality. Scale space smoothing is employed, which smooth images in different scale by employing diffusion equation. Peronal and Malik [134] proposed a new scale space edge detection method based on diffusion equation. Weickert added orientation to the diffusion filter to be able to enhance small vessels and coherence structure in images [183, 182]. Subsequent methods replaced the diffusion scalar by diffusion tensor employing the Hessian matrix configuration and analyzing eigen values of the Hessian matrix. Different geometric interpretation extracted from Hessian matrix and its eigen values configurations [81, 104, 148, 46, 24, 98]. For review of anisotropic diffusion, please refer to [181, 39, 86]. Krissian [86] and Manniesing [113] proposed anisotropic diffusion filters to segment vessels in 3D, based on a tensor structure filter. Fischl proposed a new method to indicate the best kernel function that matches the image [42]. For further review on vessel analysis the reader is referred to [97]. The proposed method uses the scalar diffusion function, and is mainly based on the

conventional Perona and Malik nonlinear diffusion filter. The reason is its more simplicity and it needs less computational calculation considering 3D vessel analysis.

The main idea of the research is providing and utilizing prediction and classification techniques on biomedical dataset.

A Analysis of Tumor Perfusion

Four samples of breast cancer cells are injected to mice and the perfusion curves of arteries and veins near the tumor cells are given as an input. Figure 4-1 highlights sample arterial and venous curves generated by 2147 and 4195 tumors implanted into mice. Several key features can be observed: The arterial curves (red) are generally characterized by a rapid increase in fluorescence intensity which plateaus within the first minute, drops off, and levels out. The venous curves (blue) demonstrate a more gradual increase in fluorescence intensity, resulting in a delayed plateau. A large number of such curves (30-60) were generated for each implant, allowing characterization of intratumoral heterogeneity. Differences were observed in peak intensity, time to arterial peak, and venous delay within each tumor, suggesting that vascularization can be heterogeneous within a given tumor. Independent stability analysis was performed for each video to confirm that ROI placement had no significant impact on the rate of tracer influx and that variations in signal intensity at any given time-point fell within the overall signal noise. Data was considered robust and included for classification when these conditions were met.

B Identification of Tissue

Heterogeneities in intratumoral perfusion make it difficult to apply standard curve-fitting models for perfusion classification. The first-pass perfusion signal in tumor arteries, for example, does not necessarily rise very quickly to a maximum as would be expected in normal tissue. This was particularly evident for the 4195 and 3887 biopsy implants, in which a slow rise in arterial fluorescence was followed by little or no drop, indicative that the tracer was already fully mixed in the blood by

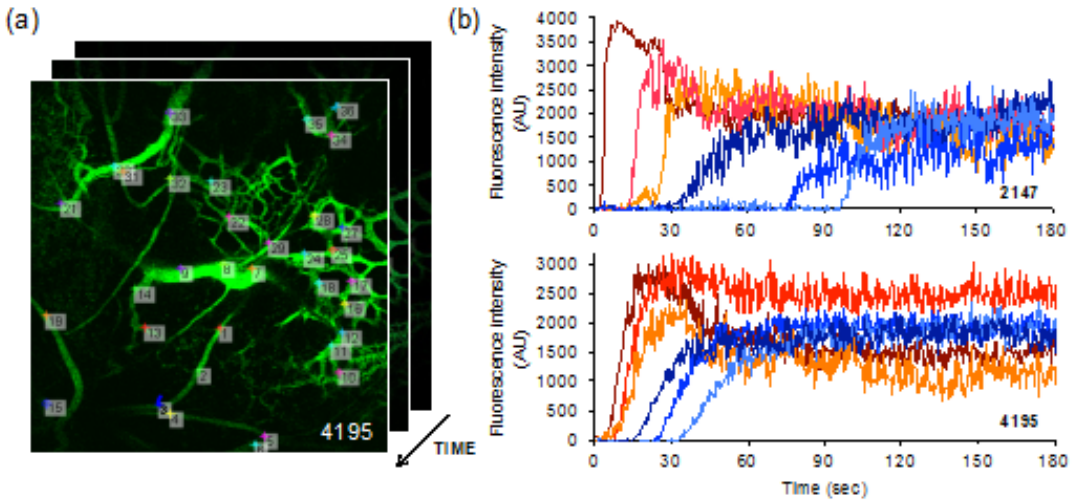

Figure 11. Quantification of tumor first-pass perfusion following injection of a 40kDa FITC dextran tracer. (a) Representative regions-of-interest (ROIs) selected for time measurements of fluorescence intensity. Circular ROIs were randomly defined inside arterioles, venules, and capillaries, between branching points, yielding approximately 30-60 ROIs per video. (b) Representative arterial (red) and venous (blue) perfusion-time curves of individual vessels, measured for single 2147 and 4195 tumors. Variations in time to arterial peak and venous delay are observed within each tumor. The videos from which data was derived may be found in the Supplementary Information [175]

the time of arrival. In contrast, 2147 and 2665 tumors behaved more like normal arteries in that their fluorescence intensity rapidly peaked and then declined immediately after the peak.

These observations suggest that the time it takes for arterial flow to reach its peak is a critical feature of tumor perfusion. Accordingly, a distribution density function is generated in which the highest peak has the highest probability of signal, and the time to reach the peak following tracer entry was defined as the time to arterial peak. The perfusion of the tumor venous system showed distinct abnormalities as well. In normal tissue, the difference from the time blood enters an artery until it reaches the corresponding vein is a few seconds. In the tumors in this study, however, this process varied from a few seconds to over a minute. The 2147 tumor in Figure11, for example, showed a delay of 30-90 seconds between when the tracer appeared in the arteries and appeared in the veins, whereas the 4195 showed a venous delay of 10-30 seconds. Thus, it is postulated that venous delay is also an important feature of tumor perfusion. Accordingly, statistical rules are applied to measure the probability of this delay. The time to arterial peak and venous delay was calculated for each biopsy implant and mapped to a single point within a two-dimensional feature space. The final value of a given feature is the expectation of all the probabilities of the vessels imaged over a 10 mm2 field-of-view. Table 2 shows the results of the two features calculated for each implant. When plotted in a two-dimensional feature space, the complexity of the data becomes apparent (Figure 12). The scattering of these points reflects both intra- and inter-tumoral heterogeneities. If no intra-tumoral heterogeneities were present, for example, it is expected to see four tight clusters of data, one for each set of tumor replicates. Since the data does not fall in a single line, it cannot be linearly classified.

The data was separated into three different classes based on the pattern of scatter: the first includes cases that are poorly vascularized (i.e. short time to peak and short venous delay), the second includes cases that are well-vascularized (i.e. long time to peak and long venous delay), and the third class is neither poorly nor well-vascularized (in between vascularized). These in between cases can represent

TABLE 2. The calculated expectation of individual tumor features [175]

	Tumor type	Feature 1: Venous delay	Feature 2: Arterial peak
1	2147	0.0415	0.0342
2	2147	0.1608	0.1307
3	2147	0.1468	0.1381
4	2665	0.0300	0.0270
5	3887	0.0187	0.0952
6	3887	0.0961	0.2361
7	3887	0.2766	0.2896
8	3887	0.0540	0.0912
9	4195	0.0830	0.0609
10	4195	0.1822	0.1047
11	4195	0.1334	0.0751
12	4195	0.3526	0.2309

tumors of homogeneous vasculature with intermediate perfusion properties or tumors of heterogeneous vasculature with regions with differing perfusion properties. No restrictions were placed on the classification of very well or very poorly vascularized cases. Figure12 shows the data set in the two-dimensional feature space after applying the FCM classifier.

The data was ranked by taking into account both the distance of each data point to the center of its cluster and a weighted term for each of the measured features:

$$rank(p) = W_1 \times feature_1(venousdelay) + W_2 \times feature_2(arterialpeak) \qquad (19)$$

where p is the tumor replicate in each class and w_1 and w_2 are the weighted terms for the venous delay and arterial peak, respectively. There exist numerous optimization algorithms to determine such weighted terms [163, 135, 194, 195]. Here, the values for w_1 and w_2 were chosen by separately calculating the first moment of each feature in each class. Since each feature is described as a separate distribution function, w_1 and w_2 were selected so that $w_1 + w_2 = 1$. The other constraint, which is based on the observation that normal vasculature is usually associated with a higher probability of venous delay than rapidly perfused tumors, assigns a higher weight to w_1. The ratio r of the mean of the two features in each class thus defines the second constraint as $w_1/w_2 = r$. Using the experimental data, $w_1 = 0.6$ and $w_2 = 0.4$ are calculated. Note that the Euclidean distance is not chosen from the samples to the center of the classes as the ranking criteria since the accuracy of such

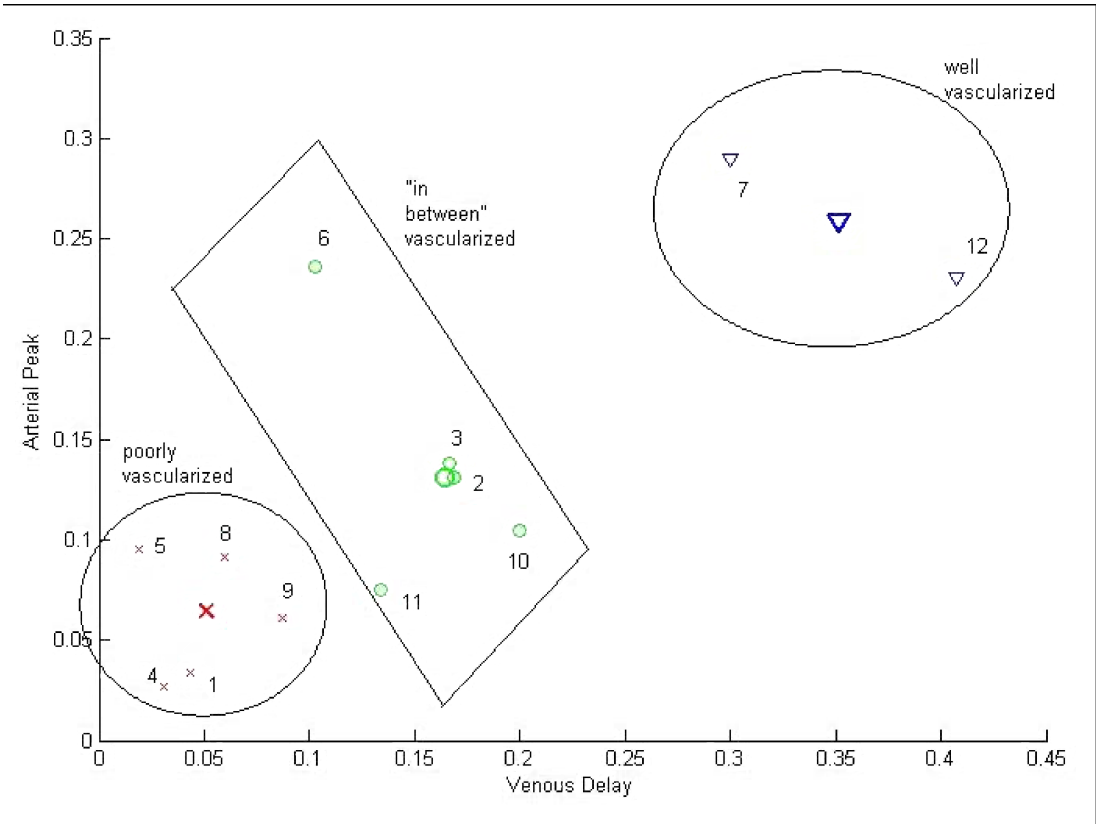

Figure 12. Application of the FCM classifier partitions the data set into 3 distinct categories. The numbers denote the individual tumor replicates as listed in Table2. Three classes are marked and the samples are shown with the same color as the center of the class. The red crosses represent poorly vascularized cases, the blue triangles belong to the well-vascularized class, and the green circles denote the in between cases [175].

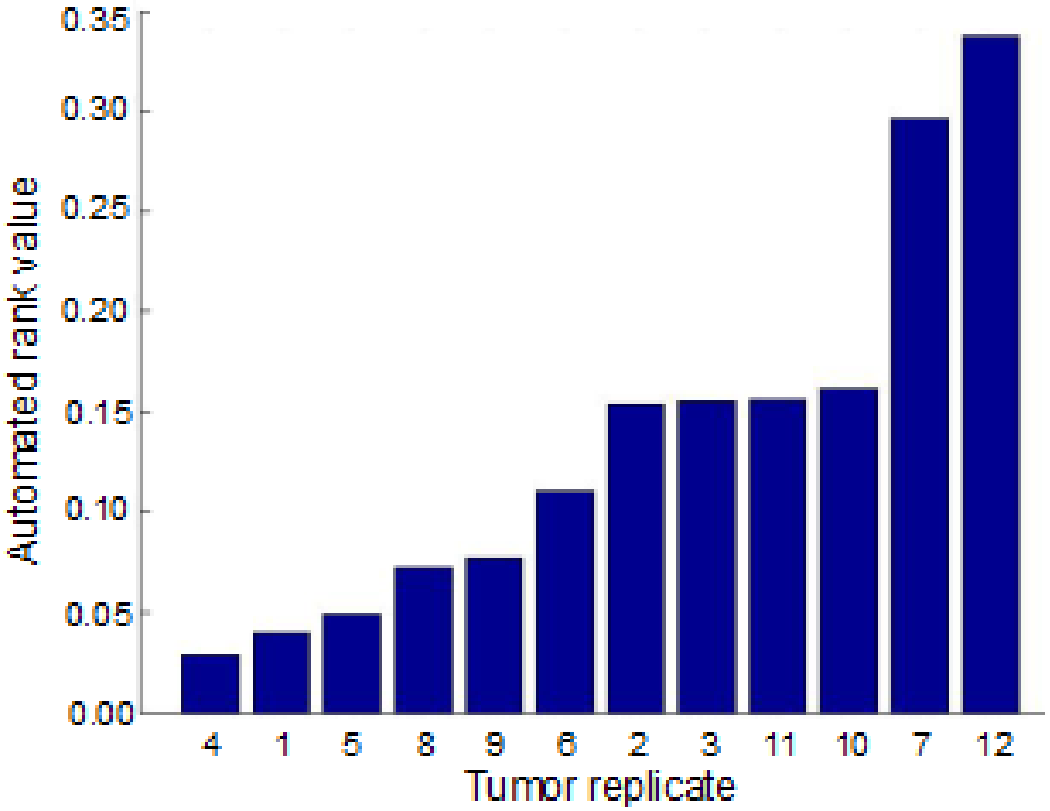

Figure 13. Automated ranking of the implants based on weighted feature probabilities.

an approach is proportional to the sample number. Figure 13 shows the automated tumor ranking from poorly vascularized tumors (left) to well-vascularized tumors (right). Despite the small sample size, it can be observed that there is a large dynamic range (9.5 fold) in the measured values. Interestingly, replicates from the same patient (1-3 (2147), 5-8 (3887), and 9-12 (4195)) do not group together, which reflects intra-tumoral heterogeneities in growth and vascularization.

Particle accumulation was measured for all tumor replicates evaluated here. Table3 shows a comparison between tumor classification, automated ranking, and

TABLE 3. Comparison between classification, automated ranking, and experimentally derived nanoparticle accumulation measurements. Average values for each tumor type are followed by the individual measurements. Averages are presented as 1 standard deviation. Tumor vascularity is indicated by (poorly vascularized), +/- (in-between), or + (well-vascularized).[175]

	Patient	Vascularity classification	Automated rank value	Particle accumulation (/mm3)
1	2147	-	(0.0397)	(28,690)
2	2147	+/-	(0.1534)	(25,880)
3	2147	+/-	(0.1550)	(27,930)
4	2665	-	(0.0293)	(30,620)
5	3887	-	(0.0493)	(8,080)
6	3887	+/-	(0.1560)	(18,800)
7	3887	+	(0.2959)	(9,410)
8	3887	-	(0.0722)	(6,960)
9	4195	+/-	(0.0768)	(3,410)
10	4195	+/-	(0.1617)	(3,620)
11	4195	-	(0.1103)	(2,920)
12	4195	+	(0.3369)	(3,850)

particle accumulation on a tumor-by-tumor basis. Since the classification and ranking schemes do not show a consistent ranking, they cannot get compared directly with the experimental results. Therefore the average value of these features for each patient is calculated. Comparison of the averaged patient ranking with the averaged experimental measurement suggests that tumors classified as poorly vascularized would uptake the highest number of circulating particles, whereas tumors classified as well vascularized would uptake the lowest number of particles. This trend is shown in Figure 14 where the automated ranking appears inversely proportional to particle accumulation. Data was fit to a 2nd order polynomial based on the observation that 1000400nm plateloid particle accumulation is constrained by tumor-specific physiological transport phenomena (manuscript submitted). The 2665 tumor, though a single replicate, had relatively little impact on the classification scheme and was therefore considered robust and included in the fit. Thus relationship between tumor rank and particle accumulation appears to be non-linear, with small changes in the upper ranks yielding large changes in particle accumulation.

Triple-negative MDA-MB-231 xenografts were generated for model validation. Grown simultaneously in littermates for 30 days under identical conditions, these tumors nevertheless demonstrated significant differences in tumor

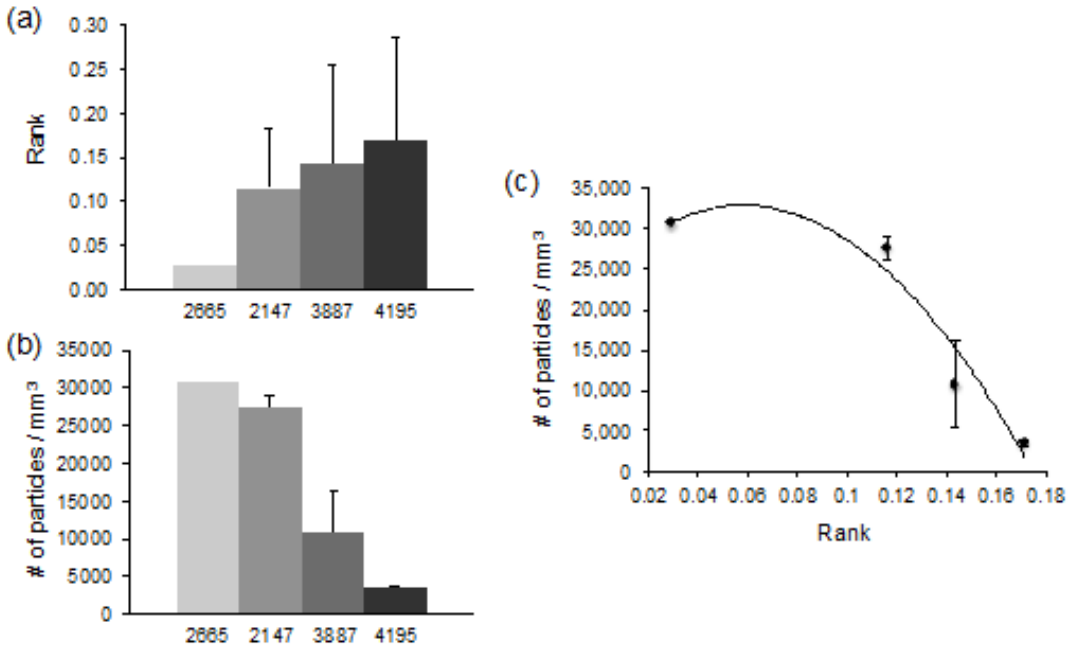

Figure 14. Comparison of tumor rank to experimentally observed particle accumulation. (a) Average tumor rank, grouped by patient number. Note that sample 2665 is a single replicate. (b) Average particle accumulation, grouped by patient. Particle accumulation is observed to be inversely proportional to tumor rank. (c) Plot of nonlinear relationship between the average particle accumulation and the average tumor rank. [175].

TABLE 4. Automated classification and ranking of MDA-MB-231 tumors. [175] Tumor vascularity is indicated by (poorly vascularized), +/- (in-between), or + (well-vascularized)

	Vascularity classification	Automated rank value	Particle accumulation (/mm3)	
			Predicted	Measured
1	-	0.3779	0	1,070
2	-	0.0977	28,940	28,310
3	-	0.0920	29,960	34,300
4	-	0.0352	31,610	26,070
5	+/-	0.2245	0	1,900

vascularization, particle accumulation, and ranking. Figure 15a highlights the morphological differences observed under bright field illumination and following FITC-dextran injection n. Cumulative particle accumulation, as measured by IVM, was found to vary by as much as 30-fold across the 5 tumors studied (Figure 15b). These tumors were individually classified and ranked in a blinded manner (Table 4).

The calculated tumor ranks were found to range from 0.035 (tumor 4, poorly vascularized) to 0.378 (tumor 1, well vascularized). Figure 15c shows the predicted and measured particle accumulation values, plotted by tumor rank. The three tumors predicted to show high particle accumulation (greater than 20,000 particles/mm3) correlated in a statistically significant manner (two-tailed test with =0.05) with the model prediction (R=0.99 as measured by Pearson Product Moment Correlation), while those with relatively high ranks (greater than 0.18) showed low particle accumulation as expected.

C Classification of DCE-MRI perfusion curves in renal transplant patients

An accurate classifier is obtained using the gamma variate function; this function is commonly used to model the first cycle (i.e., the first-pass transient phase) of the transit of contrast agents, where peak time is the time that the agent circulating in the blood takes to reach its highest level during this cycle [75].

A sample renal agent kinetic curve is shown in Figure 16. To fit this type of data the kinetic curves is divided into two sections in time, namely the wash-in (transient) and wash-out (tissue distribution) phases. The transient (wash-in) phase

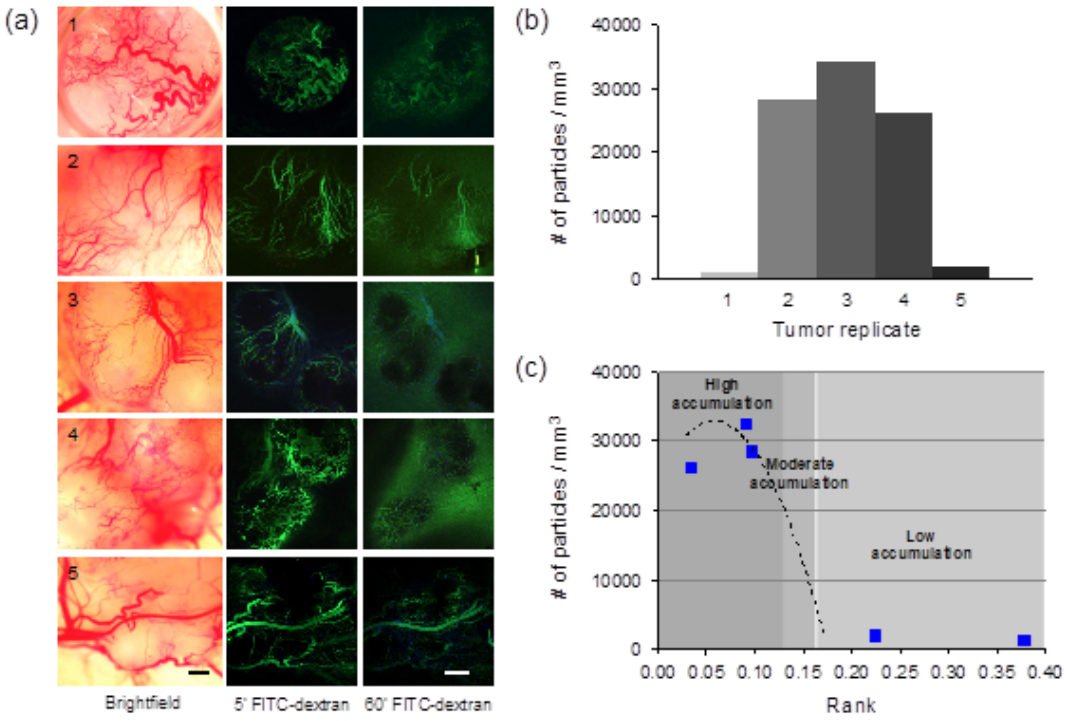

Figure 15. Vascularization, particle accumulation, and ranking of MDA-MB-231 xenografts. (a) Brightfield and fluorescence microscopy images of 5 individual tumors grown under identical conditions (Top-bottom: No. 1-5). Significant differences in vascular morphology were observed, as well as local differences in vessel permeability resulting in tracer extravasation. Scale bar = 200m. (b) Cumulative particle accumulation, as measured by IVM, ranged widely from 1,000 34,000 particles/mm3. (c) Predicted (—) and measured () particle accumulation values, plotted by tumor rank. The degree of particle accumulation was categorized by position along the predicted accumulation curve. High accumulation: >20,000 particles/mm3; Moderate accumulation: 5,000-20,000; Low accumulation: <5,000 particles/mm3.[175].

of the agent delivery is characterized by the time that it takes to reach the first peak, this transient phase is modelled with the gamma variate function. The agent delivery is also characterized during the more slowly varying phase (tissue distribution, or plateau phase) by its average signal change starting from the first peak until the endpoint of the curve. The plateau is considered of critical importance in clinical evaluation of renal transplant patients because it incorporates a large number of data points over the signal intensity time series to characterize perfusion; therefore it is less dependent on temporal sampling. The general form of the gamma variate function is given by:

$$y(t) = A(t - t_0)^{\alpha} exp\left(-\frac{t - t_0}{\beta}\right) \tag{20}$$

where A, α, and β are the free parameters; t is the time and t_0 is the initial time, which is considered to be 0. Here, a simplified gamma variate function is employed, as proposed by Madsen [109], using a least-square linear algorithm to fit the data points. This changes the original formulation in Eq.20 to a linear equation, with only one unknown parameter (see the appendix in [23] for a mathematical proof). The linearized equation is defined as:

$$ln\left(y\left(t'\right)\right) = ln\left(y_{max}\right) + \alpha\left(1 + ln(t') - t'\right) \tag{21}$$

This equation has the form $y = C + \alpha x$, where $ln\left(y\left(t'\right)\right)$ and $x = ln\left(y_{max}\right) + \alpha\left(1 + ln(t') - t'\right)$. The parameters A and β are derived from α [23]:

$$A = y_{max} t_{max}^{-\alpha} exp(\alpha) \tag{22}$$

$$\beta = t_{max}/\alpha \tag{23}$$

where t_{max} and y_{max} are the time and the intensity value, respectively, of the first peak in the perfusion curve [23]. This formulation i used to calculate the function parameters for both non-rejection and acute rejection transplant cases.

Although earlier work proposed by Madsen et al. [109] and Chan et al. [23] using the simplified gamma variate function is not renal-specific, the mathematical proofs and the methods presented therein to obtain the signal intensity using

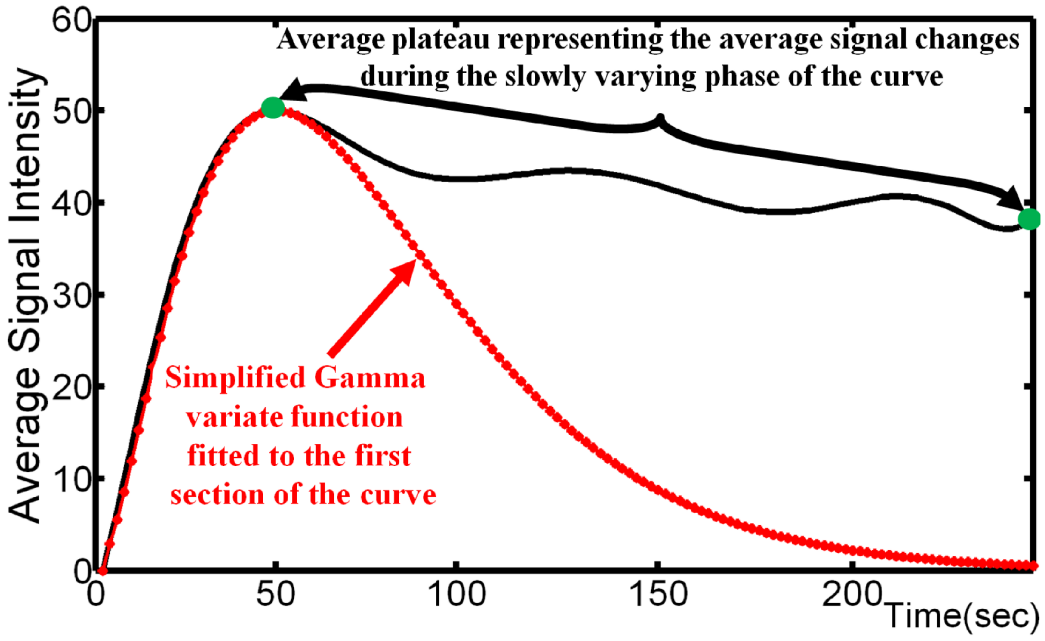

Figure 16. One sample of the given data and the fitted curve[75].

imaging techniques are tissue-independent; thus, the first cycle of the perfusion curves can usually be fitted to the gamma variate function (e.g., see recent review by Sourborn and Buckley [155]). Please note that the high accuracy obtained with the classifier results obviates the need to use the nonlinear version.

The ultimate goal of the overall framework is to provide a reproducible, non-invasive, diagnostic tool for the reliable detection of renal transplant rejection. Thus, to characterize the transplanted kidney, the final step of the proposed framework is to construct TICs of all subjects and to fit these TICs with the gamma variate function.

Following the perfusion modelling using the gamma variate function, five features were chosen for the classification of kidney status. Three features are derived from the functional model (parameter "α" estimated from Eq.21, parameter "β" defined by Eq.23 and coefficient "A" defined by Eq.22) and two features from the perfusion data (the time of the first peak and the tissue phase signal change index or average plateau (AP), see Figure11). These selected features map the data from the original data space to the feature space. These five features are calculated for all successful and unsuccessful transplants and are used for the classification of the kidney status.

D K-Nearest Neighbor (KNN) Classifier

To distinguish between non-rejection and rejection cases, a KNN classifier (K = 5) is used to learn statistical characteristics from the extracted features of training sets comprising both non-rejection and rejection groups. The perfusion data sets contain 23 cases of acute rejection and 27 cases of non-rejection transplant. 36% is used(10 non-rejection and 8 acute rejection cases) for training and the other 64% (17 non-rejection and 15 acute rejection cases) for testing. All five features were normalized by the maximum value of the respective feature of the training data. After this training step, the KNN classifier is used to classify the cases to be tested using only one of the normalized features at a time. Table5 presents the KNN diagnostic accuracy for each of the five features [75].

Then, the classifier was augmented by combining all normalized features: $\omega_1 \times \alpha + \omega_2 \times \left(\frac{1}{\beta}\right) + \omega_3 \times A + \omega_4 \times \left(\frac{1}{T}\right) + \omega_5 \times (AP)$ with appropriate weights (1.50, 0.48, 0.54, 0.69, and 1.7 for the parameters α, β, A, T, and AP, respectively) estimated by the genetic optimization using the training data sets. The weights were estimated by maximizing the Euclidean distance between the weighted-combined features of non-rejection and acute rejection groups in order to better classify the training data, based on the biopsy ground truth. For both testing and training data sets the proposed approach classifies all cases correctly (100%). Moreover, to identify kidney status a Bayes classifier is employed based on using the

Parazen window with the Gaussian kernel as density estimator, and based on using a Bayes classifier with the Gaussian distribution as density model to estimate the density distributions for each of the five features. The classification results for each feature are summarized in Table 5(b and c). For the augmented features the accuracy was 100% (18 out of 18) and 93.8% (30 out of 32) for the training and testing data sets, respectively, using the Parazen window; and 94.4% (17 out of 18) and 96.9% (31 out of 32) using the Gaussian distribution as density model [75].

Finally, to demonstrate the superiority of the perfusion analysis, the diagnostic accuracy is compared to the current clinical approach used by radiologists. Using clinical software, radiologists manually define multiple ROIs inside the kidney cortex. Then, these ROIs are used as a mask applied to all images without any segmentation or motion correction. Finally, the perfusion curve is obtained from the average intensity of these ROIs over all the time series images and three features, namely wash-in slope, time-to-peak, and wash-out slope, are extracted for the classification of the kidney status. The diagnostic accuracy of this method is reduced to 61.1% (11 out of 18) and 62.5% (20 out of 32) for the training and testing data sets, respectively. The reduced accuracy is due to the high frame-to-frame signal intensity variability related primarily to uncorrected motion effects, which eventually lead to noisy estimated parameters. These results highlight the advantage of the CAD system for perfusion analysis using the entire cortical area after correction of the global and local kidney motions compared to clinical software currently available [75].

E Performance Analysis of Selected Features

In order to evaluate the performance of the diagnostic system using the selected features for the classification of transplanted kidney status, two methods are used. First, the discriminatory ability of the selected features obtained from the perfusion curves is compared, averaged over the cortex using the well-established 95% confidence interval (CI) statistics. Based on the 95 CI separations of the groups, it is concluded that the parameters "α" and "AP" are the superior

TABLE 5. Diagnostic results using each of the selected features using the KNN (a), and a Bayes classifier based on the Parazen window with the Gaussian kernel as density estimator (b), and based on using a Bayes classifiers and the Gaussian distribution as density model (c). Note that α, β and A are the Gamma variate model parameters, T is the time-to-peak and AP is the average of the plateau phase of the perfusion curves [75]

	KNN Classifier		Bayes Classifier with Parzen Window		Bayes Classifier with Gaussian distribution	
Selected Feature	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
α	17/18	31/32	17/18	31/32	12/18	20/32
β	10/18	20/32	13/18	19/32	11/18	17/32
A	12/18	23/32	12/18	25/32	14/18	22/32
T	11/18	19/32	8/18	15/32	11/18	17/32
AP	17/18	31/32	17/18	31/32	17/18	31/32
(a)			(b)		(c)	

discriminators compared to other features.

Second, the receiver operating characteristic (ROC) is calculated as an additional metric to test the performance of the proposed diagnostic system [118]. The ROC curve tests the sensitivity of the proposed CAD system against the selection of the operating point (i.e., classification threshold) by showing the relationship between the TP rate (sensitivity) and FP rates at different operating points. Figure 17 shows the ROC curves for individual KNN classifiers of each of the normalized features and the weighted-combined classifier. For optimum performance, the area under the curve (Az) approaches unity. Visual inspection of the ROC curves in Figure17 shows that classification using the parameter "*beta*" has the worst performance, while the full combination of the features has essentially the best performance, as evidenced by Az = 1.0.

The bootstrapping method [153] computes the 95% CI for each Az. To carry out the bootstrapping method, randomly a sample is drew (n = 50) with replacement from the original data sets, and then performed the KNN-classification based on individual features as well as the augmented features using this bootstrapping sample. The procedure was repeated 1,000 times and the Az was computed each time. Next, calculated the 95% CI is calculated, defined as the 2.5% percentile to the 97.5% percentile, for the 1,000 bootstrapped AZ values for each feature. The 95% CIs were [0.965, 1.000], [0.421, 0.953], [0.423, 0.926], [0.224, 0.928], [0.911, 1.000], and [0.987, 1.000] for individual classifiers corresponding to *alpha*,

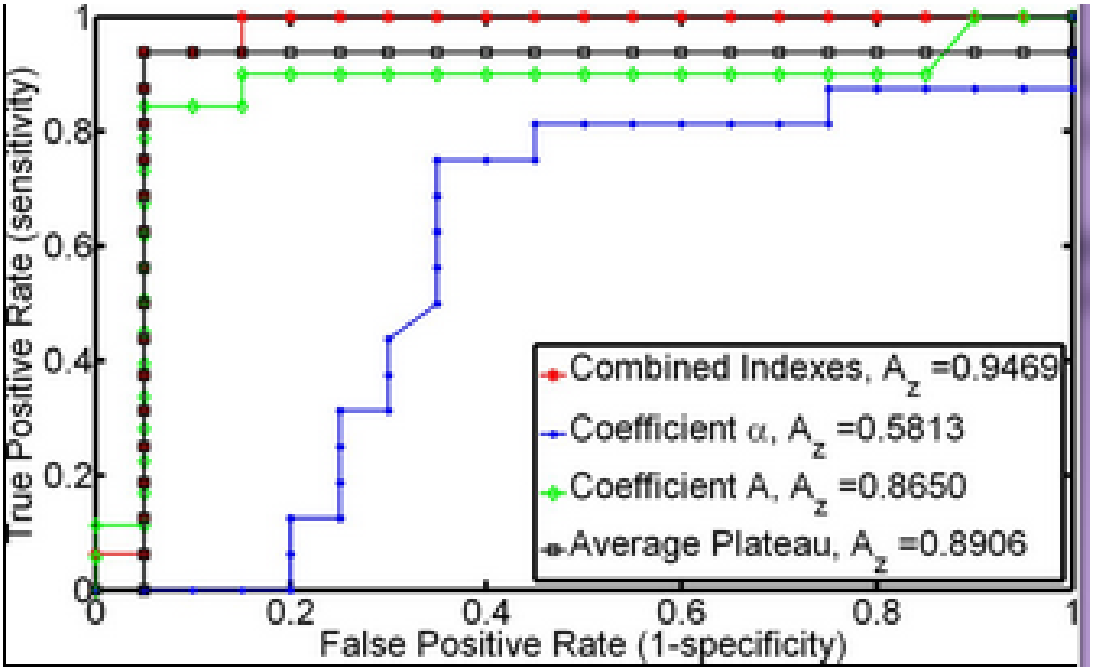

Figure 17. ROC curve for different features[75].

beta, *A*, T, AP, and for the combined classifier, respectively. Based on the bootstrapping objective of 95% CIs, it does appear that parameter "*beta*" has the worst observable performance.

In conclusion, one have presented an automated framework that incorporates deformable model segmentation, non-rigid registration, cortex segmentation, function-based modeling of contrast agent kinetics, and KNN classification of the kidney status based on functional parameters extracted from function models. The framework has the documented ability to reliably distinguish rejection from non-rejection, in a biopsy-proven preliminary cohort of 50 total participants. The preliminary results presented in this study demonstrate that the proposed framework holds promise as a reliable non-invasive tool for early diagnosis and determination of appropriate therapy for detected rejection. Although the proposed framework has been tested on 2D DCE-MRI time series data, it is believed that, in principle, the method could be extended to 3D data once technological progress in rapid MR acquisition sequences allows for sufficient spatial and temporal resolution.

CHAPTER VI

OVERVIEW OF CLASSIFICATION AND FEATURE ANALYSIS-ANALYSIS OF SEER DATABASE ON LUNG CANCER PATIENTS

In this chapter a new analysis on selecting the best feature vector is developed and two papers are under preparation. Lung cancer ranks as the second most common cancer and is usually classified as either Small Cell Lung Cancer (SCLC) or Non-Small Cell Lung Cancer (NSCLC). The diagnosis depends on cellular physical appearance evaluated via visible microscopy [146]. Available lung cancer data are analyzed from the Surveillance, Epidemiology, and End Results (SEER) program [146, 147, 164] from the National Cancer Institute (NCI) at the National Institutes of Health (NIH). The SEER Program is an authoritative repository of cancer statistics in the United States [8]. It is a population-based cancer registry which covers approximately 26% of the US population across several geographic regions and is the largest publicly available domestic cancer dataset [8]. The SEER data attributes can be broadly classified as demographic (e.g., age, gender, location), diagnosis (e.g., surgical procedure, radiation therapy), and outcome (e.g., survival time, cause of death), which makes the SEER data ideal for performing outcome analysis [8]. Lung cancer survival rate is a measure of patients who live longer than 5 years.

Machine learning techniques are commonly used in different domains and applications such as advertisement, insurance, finance, social media, fraud detection, etc. Healthcare datasets with several measured factors are not readily available, so of cancer outcomes using large datasets is a challenge. Most of published work considers small groups of patients with a few measured factors and applying

statistical analysis, which can lead to biased results due to the small number of samples.

Previous results have analyzed the SEER database via statistical techniques [138, 130, 16, 55, 165, 50, 187, 180], as well as classification techniques [8, 154, 9, 7, 72]. Survival time analysis in medical applications is very important because sometimes the patients are not following up their treatment and the survival time is missed in future so Kaplen-Meier estimate using the previous probabilities of patients in time and multiply them to find the missed survival time information[52]. Several survival function has been defined and measured based on a Kaplan-Meier estimator by which is mostly used for survival rate analysis [52]. The aim is to evaluate several standard classification methods to help determine patient survival based on a set of features extracted from the SEER database. The survival function is defined and measured based on a Kaplan-Meier estimator, which is often used for survival rate analysis [52]. Lung cancer data available from the SEER database [146, 147] are analyzed using machine learning techniques with the aim of developing accurate survival time and finding an optimal feature vector that discriminate patients based on the survival time. Some of the most common attributes are used and different classifiers and clustering techniques are applied to compare the results. The research considers reviewing the accuracy of different classification techniques using Weka and R. Several classification techniques should be applied on the data along with various machine learning techniques in order to determine an optimal classification. Tumor grade, tumor size, gender, age, stage, histology, number of primaries, lung cancer as the primary one, histology code, primary site are the analyzed features, which there are around 25 features available. The results show that SVM has the highest accuracy for classification of the patients.

Next goal is to develop an automatic technique to find the most similar patients because physicians cannot accurately reassure the patients about the survival time using the common clinical features. The analysis extracts the optimized feature vector that can be used to improve clustering the SEER database patients.

The papers that utilized machine learning techniques in analyzing SEER database are reviewed.

Recent work by Kai et al. [189] employs the idea of agglomerative clustering to generate groups of patients. Their technique has two steps: first, patients are divided into groups that do not have the same survival and the p-value for each is calculated; then the groups are merged so that groups with smaller p-value are combined and the closest groups are merged with smaller p-value groups. In the next step they compare the survival function of generated groups from the first step. If these groups do not have the same survival experience, then they merge the closest pairs; this forms a new group of patients. Both previous steps are repeated until in the kth iteration each pair of groups shows different survival experience. The method is applied on breast cancer data in the SEER dataset with selected attributes: tumor size, tumor extension and lymph node status. Numeric attributes are changed to categorical attributes with arbitrary levels. Tumor size is considered as seven levels, tumor extension as five levels and factor lymph node as two levels. If one only considers the stage attributes, patients are assigned into six stages. More than 95% of instances belong to three stages: I, IIA and IIB. The authors show that stage only cannot be a good categorization method to predict survival. If the patients are categorized into 12 groups, then their survival rates are similar to each other in each group and this yields an accurate method for categorizing patients based on survival rate [189]. The threshold for parameters to merge the groups, however, can be a challenge to determine, and changing the threshold values will affect the results.

Agrawal et al. [8] applied association rule mining techniques to generate several rules for lung cancer, some of which are redundant and are manually removed based on domain knowledge. Association rule mining techniques finds interesting association or correlation relationship among a large set of items, different techniques are proposed to extract the rules and there are several standard criteria which show how to choose the best rules and keep the optimized ones based on the given dataset [10]. The study implemented an automated technique to make

the tree of rules. The Hotspot algorithm implemented in Weka is used for implementation [8]. Three factors are considered for the algorithm: the maximum branching factors, adding a new branch, and the factor to be used when adding a new branch. The study considers both numeric and nominal attributes. Numeric attributes are those that the value of the feature vector are numbers and nominal attributes are those that the value of the features are categorical.

Agrawal 2011 [8] proposed a tree-based algorithm which uses the entire dataset at the very beginning, and descends into the data in a depth-first fashion using a greedy approach. Each node of the tree represents a segment and hence and association rule. The attributes used include: age birth place, cancer grade, diagnostic confirmation, farthest extension of tumor, lymph node involvement, type of surgery performed, reason for no surgery, order of surgery and radiation, scope of regional lymph node surgery, cancer stage, number of malignant tumor, total regional lymph nodes examined.

Measuring the efficiency of treatments and surgery is a desired analysis on the SEER dataset. Although the dataset lacks the information of chemotherapy, Yan Wu et al.in [187] considered the effectiveness of radiation and surgery. The study answered the question whether lung cancer patients survive longer with surgery or radiation, or both. The paper uses a Propensity Score which is a conditional probability that a unit will receive a treatment given a set of observed covariates. Two methods are applied for estimating the Propensity Score: logistic regression and classification tree. The results show that patients who have not received radiation with or without surgery have the longest survival time [187].

A SEER database attributes

Patients diagnosed with lung cancer from 2004-2009 are chosen. In the US, Tumor-Node-Metastasis (TNM) staging was introduced in 1959 by the American Joint Committee for Cancer Staging and End Results Reporting, now the American Joint committee on Cancer (AJCC). Stage grouping based on TNM is as follows: Stage I (T1,N0,M0), Stage II (T2, N0, M0), Stage IIIA (T3, N0, M0), Stage IIIB

(T4, N0, M0), and Stage IV(Any T, Any N, M1).

The lung cancer outcome calculator uses several patient attributes. The dataset has more than 9000 instances. Different classifiers that use two types of features include are applied: numeric and nominal.

The staging system of cancer patients has the same definition after 2002, so for consistency patients with lung cancer from 2003-2009 are chosen. The survival time ranges between 0 and 71 months for lung cancer patients in the SEER database. The data show that patients sharing the same clinical features exhibit a heterogeneous variety of survival times. Finding an optimal feature vector is therefore a challenge, since the aim is to show that patients with similar feature vectors also have similar survival rates, and vice versa.

B A Review on the Dataset Attributes for Lung Cancer Patients

Patients diagnosed with lung cancer from 2004-2009 are chosen. The dataset indicates minimum and maximum survival (in the range of 0 to 71 months). The selected lung cancer features extracted from the original dataset are summarized in Table 6.

Information gain measures the level of impurity (heterogeneity) in a group of samples. Samples are instances grouped by similar feature vectors. A common way to measure impurity is through entropy, with higher entropy indicating higher information content [18]. Entropy can be calculated as follows (where pi is the probability that a group of instances has the same attributes and i ranges over the number of instances in the group) [29, 145, 123, 124]:

$$Entropy = \sum -p_i * \log_2 p_i \tag{24}$$

The aim is to determine which attribute in a given set of feature vectors is most useful for discriminating between the two classes. Information gain enables determining the importance of a feature in each attribute of the feature vector. The order of the attributes given by information gain is utilized to set the nodes of the

TABLE 6. Lung cancer dataset attributes, first column is the names of the attributes and the second column is a brief description of the attribute and the third column is the attribute type: numeric or nominal.

Feature name	Description	Type
Age	Age of the patient at time of diagnosis	Numeric
Grade	A descriptor of how the cancer cells growth	Nominal
Radiation	Whether patient received radiation	Nominal
Radiation sequence with surgery	The order of surgery and radiation therapy	Nominal
Number of primaries	Number of malignant tumors	Numeric
T	AJCC component describing tumor size.	Nominal
N	AJCC component indicating lymph node involvement.	Nominal
M	AJCC component describing tumor dissemination to other organs.	Nominal
Primary Site	Location of tumor within the lungs.	Nominal
Stage	Stage of tumor based on T, N and M.	Nominal
First Primary	First malignant primary indicator.	Nominal
Sequence Number	Order of lung cancer occurrence with respect to other cancers.	Numeric
Histologyrecode-boadgroupings	The microscopic composition of cells/tissues	Nominal
RXSummScopeRegLNSur(2003+)	procedure of removal, biopsy, or aspiration of regional lymph nodes.	Nominal
RXSummSurgPrimSite(1998+)	Surgical procedure to remove or destroy tissue of the primary site.	Nominal
CSlymphnodes	The number of lymph nodes involved.	Nominal
DerivedSS1977	Derived SEER Summary Stage 1977 , effective with 2004+ diagnosis.	Nominal
Survival time	Number of months that patient is alive from the date of diagnosis.	Numeric
Survival time class	The survival time attribute is changed into six classes.	Nominal

decision tree. The information gain can simply be defined as:

$$Informationgain = Entropy(parent) - AverageEntropy(Children)$$

(25)

An alternative definition of information gain is:

$$Informationgain(S, A) = Entropy(S) - [\Sigma_{v \in Value\{A\}} \frac{|S_v|}{|S|} Entropy(S_v)] \qquad (26)$$

where S is the total number of samples from all the classes in a specific node of the tree, and S_v is the number of samples of each class in that node of the tree.

Gain ratio is a modified version of information gain based on the ratio of the information gain and a term called intrinsic information [29, 145, 30]. Intrinsic information is the entropy of distribution of instances into branches, i.e., it states how much information is needed to rank an attribute in order to determine which instance belongs to which branch. The advantage of using the gain ratio technique

TABLE 7. Ranking of attributes calculated based on gain ratio. Column 1 is the attribute number, Column 2 is the ranking, and Column 3 is the value of the calculated gain ratio.

Attribute Number	Rank	Gain Ratio
M	1	0.1510082
Stage	2	0.0639196
Surg Prim Site	3	0.46587
Sequence Number	4	0.041907
Derived SS	5	0.0416574
Scope Reg LN Sur	6	0.028007
T	7	0.0275921
CSLymph Nodes	8	0.0224748
N	9	0.0188275
Number of Primaries	10	0.0154234
Primary Site-Labeled	11	0.0123318
Histology Code	12	0.0080805
Grade	13	0.0075374
Age	14	0.0061738
First Malignant Indicator	15	0.0000343

instead of information gain is that gain ratio reduces the bias of the ranking [29, 145, 30, 62].Information gain mostly finds the most relevant attributes so they can be put near the root of the tree. Some attributes have high information gain but they are for example unique for each sample and cant be used in future for unknown samples so in this case it is preferred to use gain ratio .

$$IntrinsicInformation(S,A) = \left[\Sigma_{v \in Value\{A\}} \frac{|S_v|}{|S|} Entropy(S_v) * \log_2 \frac{|S_v|}{|S|} Entropy(S_v)\right] \tag{27}$$

$$Gainratio = \frac{InformationGain}{IntrinsicInformation} \tag{28}$$

Note that the value of the attributes decreases as the intrinsic information increases. Table7 lists the ranking and the name of the features in the dataset determined using gain ratio. The lowest gain ratio is based on age, tumor grade, and whether the lung tumor is the first primary or not. Age and grade were found previously to not be good predictors of survival [25]. Features with the highest gain ratios are metastatic grade, stage, whether the primary site was surgically resected, sequence number, and derived SS (refer to Table7).

The attributes with values similar for more than 90% of the records are manually removed since these attributes provide minimal discriminating value.

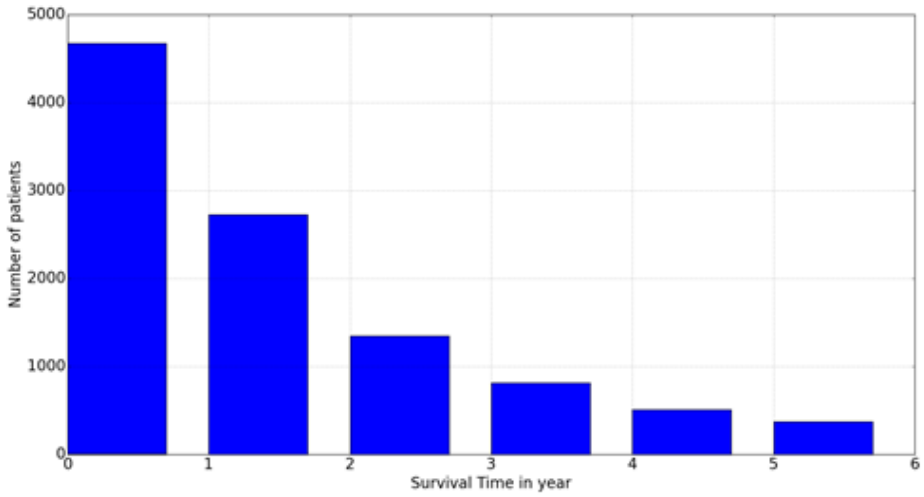

Figure 18. Survival time histogram for instances in 2004 extracted from the SEER database (0, 6 months, 6-12 months, 2 years, 3 years, 4 years, 5+ years).

Those attributes for which more than 65% of patients have the same value are deleted. Further, Table8 lists the attributes removed to avoid overfitting or skewness (for simplicity, the research does not pursue techniques to handle skew classifiers).

Figure 18 shows the range of living which each box shows a range of living for a number of patients and the number above each bar shows the number of patients that have lived or are living in that range.

It seems that grain ratio and entropy are not good criteria to evaluate and ranking the features.

TABLE 8. List of deleted features to avoid overfitting and skewness. First column is the name of the attribute and the second column is the number of instances which mainly have only one value.

Attribute Number Gain Ratio	Rank
Radiation Sequence with Surgery	1267/1549 receive Radiation After Surgery
Radiation	1502/1549 receive beam Radiation
Primary Site	1010/1549 have the same primary site(lung)
Cause of death	Except 371 who are still alive, 995/(1549-371) are dead because of lung and bronchus
Reason no directed cancer surgery	1140/1549 receive surgery

TABLE 9. Minimum, maximum, mean and standard deviation of survival time (months) for each year. Last column is normalized by the maximum number of survival months for each year.

Year of Diagnosis	Minimum Survival time	Maximum Survival time	Mean of survival time	Standard Deviation of survival time	Normalized survival time
2004	0	71	31.21	23.63	1.90
2005	0	59	28.86	18.96	1.73
2006	0	35	24.21	14.86	0.73
2007	0	47	20.27	10.03	2.70
2008	0	23	13.78	5.91	1.50
2009	0	11	5.18	3.31	1.81

Since the highest mean survival time corresponds to 2004, normalize the values of the other years by this value. The normalized values are calculated based on mean and standard deviation:

$$Normalizedvalue = \frac{maximumsurvivalmonths - meanofthedata}{standarddeviation} \tag{29}$$

The normalized results show that 2007 has the maximum rate of survival time; the ranking of the years in terms of survival time then are (from highest to lowest): 2006, 2008, 2005, 2009, 2004, 2007.

C SEER Dataset Classification

The first step of the classification involves removing corrupted or inaccurate instances from the dataset. Further, although techniques exist to handle missing attribute values [146], those instances for which one or more attributes are missing are eliminated in order to maintain a consistent dataset.

Previous results have analyzed the SEER database via statistical techniques [138, 130, 16, 55, 165, 50, 187, 180], as well as classification techniques

[8, 154, 9, 7, 72]. Survival time analysis in medical applications is very important because sometimes the patients are not following up their treatment and the survival time is missed in future so Kaplen-Meier estimate using the previous probabilities of patients in time and multiply them to find the missed survival time information [52].

In earlier work by Kai et al. (2007) [189] employs the idea of agglomerative clustering to generate groups of lung cancer patients. The technique has two steps: first, patients are divided into groups that do not have the same survival and the p-value for each is calculated; then the groups are merged so that groups with smaller p-value are combined and the closest groups are merged with smaller p-value groups. In the next step they compare the survival function of generated groups from the first step. If these groups do not have the same survival experience, then they merge the closest pairs; this forms a new group of patients. Both previous steps are repeated until in the kth iteration each pair of groups shows different survival experience. The method is applied on breast cancer data in the SEER dataset with selected attributes: tumor size, tumor extension and lymph node status. Numeric attributes are changed to categorical attributes with arbitrary levels. Tumor size is considered as seven levels, tumor extension as five levels and factor lymph node as two levels. If one only considers the stage attributes, patients are assigned into six stages. More than 95% of instances belong to three stages: I, IIA and IIB. The authors show that stage alone is an insufficient categorization method to predict survival. If the patients are categorized into 12 groups, then their survival rates are similar to each other in each group, and this yields an accurate method for categorizing patients based on survival rate [189]. The threshold for parameters to merge the groups, however, can be a challenge to determine, and changing the threshold values will affect the results.

Association rule mining techniques can determine interesting association or correlation relationships among a large set of items; different techniques have been proposed to extract the rules and there are several standard criteria which suggest how to choose the best rules and select the optimized ones based on the given dataset [10]. Agrawal et al.[8] implemented an automated technique to make a tree

of rules for lung cancer, some of which are redundant and are manually removed based on domain knowledge. Three factors were considered: the maximum branching factors, adding a new branch, and the factor to be used when adding a new branch. The study considers both numeric and nominal attributes (for which the value of the feature vector are either numerical or categorical, respectively).

Agrawal 2011 [8] proposed a tree-based algorithm using the entire dataset from the very beginning, and descending into the data in a depth-first fashion using a greedy approach. Each node of the tree represents a segment and hence an association rule. The attributes include: age birth place, cancer grade, diagnostic confirmation, farthest extension of tumor, lymph node involvement, type of surgery performed, reason for no surgery, order of surgery and radiation, scope of regional lymph node surgery, cancer stage, number of malignant tumor, and total regional lymph nodes examined.

Measuring the efficiency of treatments and surgery is a desired result from analyzing the SEER dataset, although the dataset lacks the information of chemotherapy. Yan Wu et al. [187] considered the effectiveness of radiation and surgery. The study explored the question whether lung cancer patients survive longer with surgery or radiation, or both. A Propensity Score was used, representing a conditional probability that a unit will receive a treatment given a set of observed covariates. Two methods are applied for estimating the score: logistic regression and classification tree. Since patients can receive surgery or radiation separately or together, the score is calculated for each group and then the attributes are ranked. Statistical information related to the combination of survival time and radiation are extracted, and a classification tree is generated for each group. The results show that patients who have not received radiation with or without surgery have the longest survival time [187].

Several classification techniques are applied and compares the accuracy and F1-score of each classifier separately. Six different classes are defined, so multi error classification is calculated. The codes are developed in Weka to compare the results of different classification techniques. This section will compare each applied

classifier.

Classification techniques can be grouped by type. In this paper supervised learning is used to classify patients, and one from each group of classification techniques is chosen.

Supervised learning can be applied if there are a sufficient number of labeled data. The process involves collecting and labelling a current dataset, and then developing or customizing classification techniques for this dataset. The classification model is then used to classify unknown instances. About 60% of the data is used for training, 20% for cross validation, and 20% for testing. An alternative is to use k-fold cross validation so for each experiment k folds are used for training and the remaining samples are used for testing. The advantage of k-cross validation is that all the samples in the dataset are used for both training and testing. The error rate of the classifier is the average of the calculated errors for the k-number of experiments.

Chosen classification technique should not be low variance or high biased. Low variance means that the model is not well fitted to the current training set so the error rate increases for the test sets. High variance occurs when the model is overfitted and the error rate for the training set is very low while the error rate for the test or cross validation sets are high.

The classification techniques are implemented using Weka and applied k-fold validation to indicate the error rate. The k is chosen as 10, which means that each algorithm is applied 10 times and the error accuracy is the average of the error rate of the 10 experiments.

The dataset was extracted from the SEER database has several features, and those with overlapped information are excluded. The goal is to apply the various classifiers and compare the accuracy results based on the survival rate of the patients. Since labeled data are needed to classify the information, the dataset is divided into six classes, and each class indicates patients with one year survival time difference.

Machine learning techniques can be mostly divided into three groups: 1-

Supervised Learning 2- Unsupervised Learning 3- Semi-supervised Learning. Supervised learning algorithms categorize the records of instances or feature vectors based on the labeled data. The classification finds the model to maximize the difference between classes and minimize the difference within each class. Unsupervised techniques do not have any labeled data in advance, so the learning technique is based on measuring the similarity of the intra classes and dissimilarity of inter instances. Semi supervised techniques use a small group of labeled data and the mathematical model changes as new unknown data is added to the system.

In general, classification techniques can be categorized as:

- Logic based techniques: Decision tree from this group is chosen.
- Statistical techniques: Nave Bayes from this group is chosen.
- Instance based techniques: KNN(K Nearest Neighbor) is chosen, the k8 shows the best results.
- Support Vector Machine: Polynomial kernel function is applied.
- Neural Networks: Non technique from this category is chosen.

K-NN uses neighbors information to classify the data. K is the parameter that shows the number of the neighbors that should be defined to make the decision about the datas class. K is the number of data points in the neighbors, every new data point are labeled based on similarity measure which is the distance function between the new datapoint and the other data points with k nearest neighbors. A vote for each instance or a weighted function is calculated which shows the value of the data point that belong to that class [123, 124].

Support vector machine (SVM) was originally proposed by 1992training, vapnik1998statistical, cortes1995support. SVM is a supervised technique which is mainly a linear model but with kernel mapping would change to nonlinear model. If two data points are linearly separable there are several hyperplanes that can separate them into two. However the one that has the largest margin and the

distance to the points on the nearside of the margin is maximized. SVM is also using marginal information by using two hyperplane, so if the data points are linearly separable then the two hypeplane that there is no point between them are used and the distance between the two hyperplane is maximized, meanwhile the distance between the points in each group to the hyperplane side is minimized. The points that are defined on the margins are called control points. The hyperplane is defined linearly but if it needed to design a non-linear hyperplane then it uses a kernel function which is a nonlinear function which maps the feature space into a new space in such a way to be able to define a linear hyperplane [123, 124].

Decision Tree classifier: decision tree is one of the famous approach for data classification. Features are leveled and the most efficient features are selected in the root and features are arranged based on their efficiency in the dataset. The level of tree for cutting is chosen, so in this way the clusters are generated. Nave Bayes: this classification technique is one of the probabilistic techniques which use Bayes rules to classify data. The data are labelled and then the conditional probabilities or bayes rules should be defined to find the probability of the given data and classify them:

$$P(Y = y_i/X) = \frac{P(X/Y y_i)}{\sum P(X/Y = y_i)P(Y)} \tag{30}$$

Nave Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. A given set of data points construct the posterior probability for each class, the posterior is calculated as:

$$p(class_j|X) = p(X|class_j) * p(class_j) \tag{31}$$

the posterior is the probability that X belongs to classj [123, 124].

The second step is to define the likelihood of the given data for each class, and the conditional probability of a class to the data points [123, 124] are found.

Support vector machine (SVM) is one of the best classification techniques in which instead of finding a hyperplane in the feature space to separate data into different classes would find a margin between datapoints in different classes. The main approach is to maximize the distance of the margin and minimize the distance

of the critical points with the border of the marginal plane. One of the important step of the SVM is finding the critical points in each class, the margins in the feature space is defined based on this critical points. SVM does not apply the approach on original feature space however it would use kernels to map the data points into another mathematical space, the kernel function depends on the given dataset [123, 124, 27, 56, 157].Some techniques like PCA is used (principle component analysis) to map data into a compressed feature space [4, 70]. Selecting the appropriate kernel function is critical and depends on the dataset. Polynomial kernel function is used as follows. Optimization techniques can be used to find the optimized margin for the classifier [27, 56, 157].

Random forest is a classification technique that is called ensemble learning that it constructs a multitude number of decision trees at the training phase and based on their classification results would make the final decision about the label of the datapoint. The method is a combination of bagging and random forest selection of features. Random forest for the unsupervised learning and clustering data is used.

Adaboost defines a weighted function to find a strong classifier. The weighted terms are features or weak classifiers. The weighs are updated in each iteration until the minimum weighted error is gotten. The weights are initialized the same for all the terms and the update term for the weights are defined as:

$$D_{t+1} = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{32}$$

Where Z_t is a normalization factor chosen so that the weight in t+1 is a distribution.

The output would be the strong classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{33}$$

The first step of the classification involves removing corrupted or inaccurate instances from the dataset. Further, although techniques exist to handle missing attribute values [146], those instances for which one or more attributes are missing are eliminated in order to maintain a consistent dataset. The extracted dataset from the SEER database has several features, and those with overlapped

TABLE 10. Classifier results based on attributes selected in Table 6-1. Instances are labelled with period of one year.

Classifier Name	Statistical analysis of results		
Nave Bayes	Correctly Classified Instances	9619	92.1095 %
	Incorrectly Classified Instances	824	7.8905 %
	Relative absolute error	15.1994 %	
	Average Precision		0.034
	Average Recall		0.919
	Average F1-Score		0.919
SVM-polynomial	Correctly Classified Instances	10335	98.9658 %
	Incorrectly Classified Instances	108	1.0342 %
	Relative absolute error	1.4666 %	
	Average Precision		0.003
	Average Recall		0.988
	Average F1-Score		0.988
J48 (logic-based algorithm)	Correctly Classified Instances	10443	100%
	Incorrectly Classified Instances	0	0%
	Relative absolute error		0%
	Average Precision		1
	Average Recall		1
	Average F1-Score		1
Random Tree(using tree and probabilistic data structures)	Correctly Classified Instances	8289	79.3737%
	Incorrectly Classified Instances	2154	20.6263%
	Relative absolute error		29.9294%
	Average Precision		0.064
	Average Recall		0.791
	Average F1-Score		0.794
Random Forest	Correctly Classified Instances	9718	93.0576%
	Incorrectly Classified Instances	725	6.9424%
	Relative absolute error		36.0728%
	Average Precision		0.016
	Average Recall		0.928
	Average F1-Score		0.929
Adaboost	Correctly Classified Instances	7402	70.88%
	Incorrectly Classified Instances	3041	29.12%
	Relative absolute error		101.1879%
	Average Precision		0.103
	Average Recall		0.571
	Average F1-Score		0.615
KNN (k =8 and higher)	Correctly Classified Instances	5618	53.7968%
	Incorrectly Classified Instances	4825	46.2032%
	Relative absolute error		84.4012%
	Average Precision		0.478
	Average Recall		0.478
	Average F1-Score		0.477

information are excluded. The goal is to apply the various classifiers and compare the accuracy results based on the survival rate of the patients. Since labeled data are needed to classify the information, the dataset is divided into six classes, and each class indicates patients with one year survival time difference.

The accuracy results and the statistical properties of each technique are summarized in Table 10, showing that SVM with the polynomial kernel is the best technique to classify the dataset in terms of patient survival time:

Classification techniques can be grouped by type. This paper uses supervised learning to classify patients, and one from each group of classification techniques is chosen. The accuracy results are summarized in Figure19 based on the ROC (receiver operating characteristic) analysis:

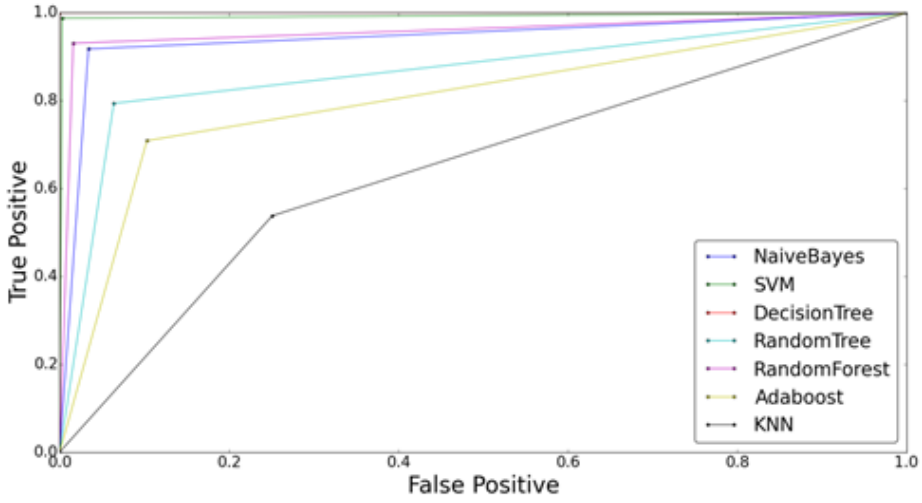

Figure 19. Comparison of ROC curve of different classification techniques. The order of the algorithms based on ranking of ROCs is: Decision Tree, SVM, Random Forest, Nave Bayes, Random Tree, Adaboost, KNN.

D Applying Clustering Techniques and Analyzing the Results on SEER Database

The features are selected from Table 10. The main goal is finding the most optimized feature vector in analyzing lung cancer patients survival time based on SEER database. Clinics mostly extract tumor grade and stage which indicate the main behavior of tumor growth in patients. Most of the papers in considering healthcare applications are biased and they do not report false negative results, so there are several patients that their final survival date is unknown. Although it seems reasonable that patients diagnosed with higher stage and grade, but an automatic technique that use machine leaning algorithm to evaluate the results is needed.

Figures 20 through 24 visualize the instances and attributes that are most clinically relevant; a common feature among them is survival time class which has seven values, namely patients who survive in a period of one year. In order to better evaluate the feature vector, some of the features that are clinically important are visualized in two dimensions. A rate of error is added to the data features to give a better sense of the density of instances in every row of each figure.

Each box and whisker plot in 20 through 24 indicates the first to the third quartile. The dataset is divided into four equal groups. The data is sorted in its ascending order then the lower half needs to be defined, median and the upper half of the dataset. Based on this categorization the quartiles is defined. The lower half of the dataset is all the values that are on the left side of the median values and the upper half is all the values that are on the right side of the median value. First quartile is the median of the lower half, second quartile is the median of the dataset and the third quartile is the median of the upper half of the dataset. The dark line in the boxes is the median, points lying outside the box are outliers.

Clustering techniques categorize data points and label the most similar points in one group. Clustering techniques can be classified into hierarchical, partitioning, density-based, multimodal clustering, grid-based, and soft-computing methods

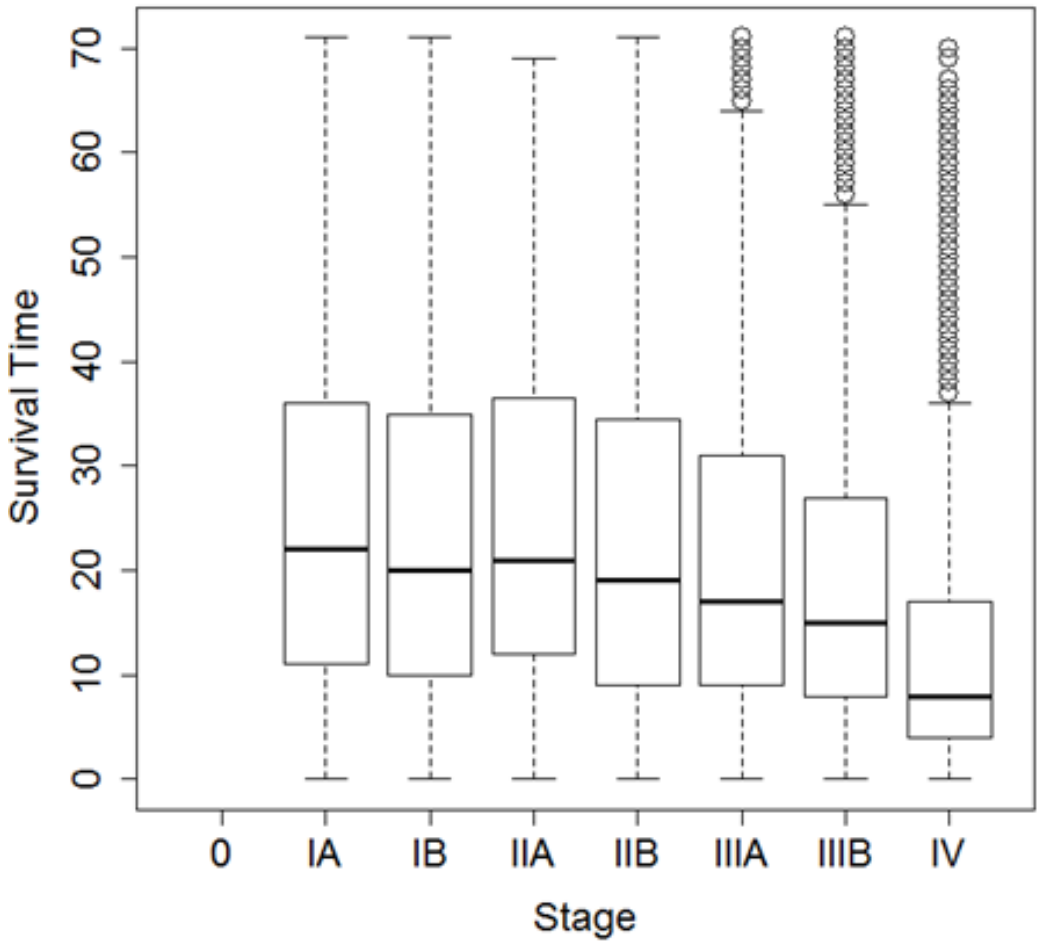

Figure 20. Survival Time vs Stage (there are only a few number of patients with stage 0 so no box is plotted.

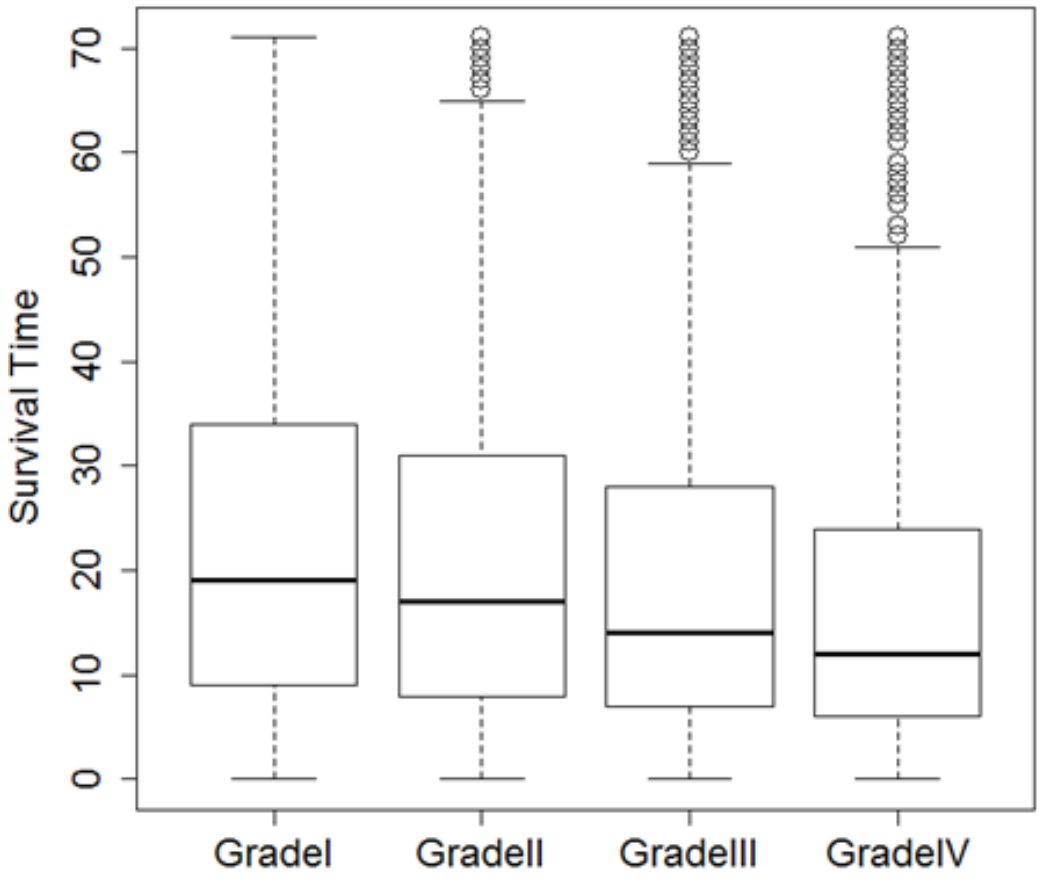

Figure 21. Survival Time vs Grade.

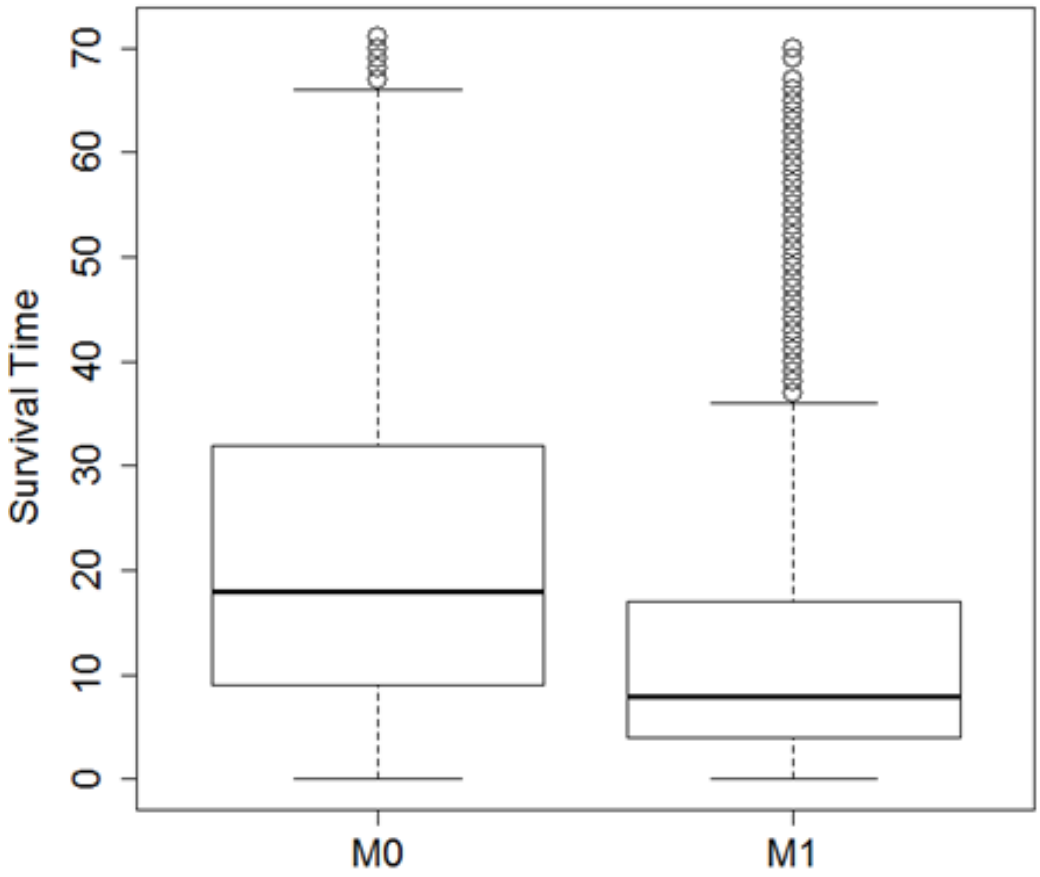

Figure 22. Survival Time vs M.

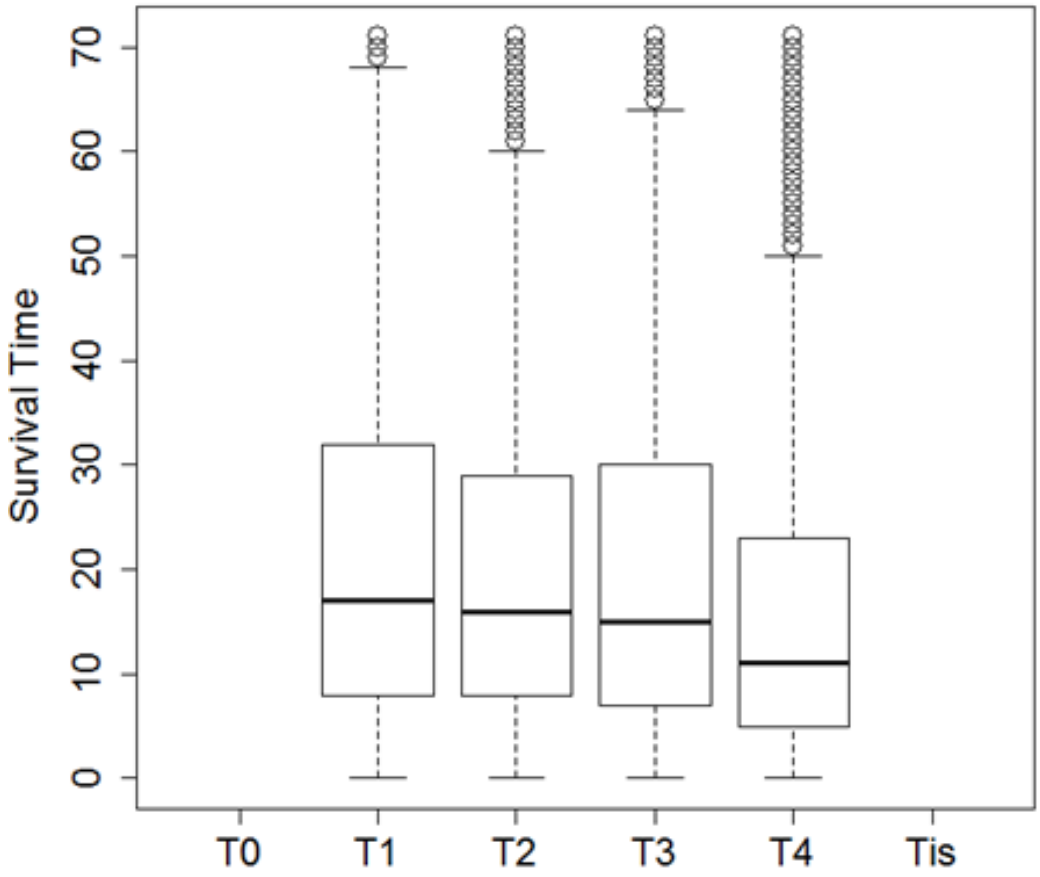

Figure 23. Survival Time vs T (there are only a few patients with T0 and Tis so no box is plotted)

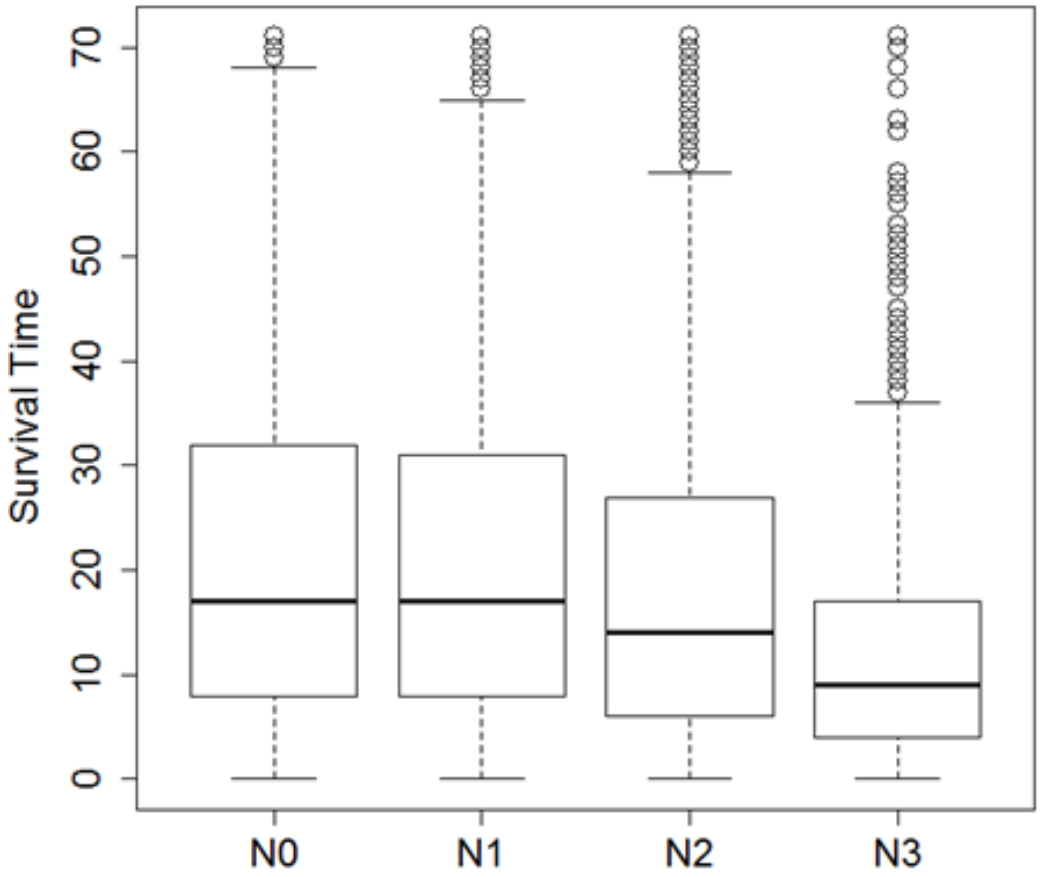

Figure 24. Survival Time vs N (Survival Time vs N)

[196, 108]. A different approach to group clustering techniques is considered in [196]: agglomerative vs divisive, monothetic vs polythetic, hard vs fuzzy, deterministic vs stochastic, incremental vs non-incremental. Such techniques are typically used to find a group of data points that have similarity. The similarity metric can be very different depending on the applied clustering technique. The clustering techniques that using distance metric might not be appropriate for qualitative features or with a combination of both quantitative and qualitative features.

Three groups of clustering techniques are compared: the first group uses distance metric to illustrate the membership of instances to each cluster. The second group uses probabilistic information of the data to find the most separable clusters, and the third group is called non-negative matrix factorization (NMF), which uses a mathematical technique to factorize the feature matrix. Although the three groups of clustering techniques use different methods to cluster the data, they generate the same results for the two groups of feature vector.

Hierarchical clustering creates a hierarchy of clusters, in which distance metrics are used as the similarity measurement to split or merge the clusters. Hierarchical clustering with density based clustering (EM) and matrix factorization are compared, both of which do not use distance measurement. One of the optimized adaptations of the density based clustering technique (an optimized EM) is considered.

The feature vector is split into two groups: one groups are the features that are used by clinics to predict the survival time and the second group is chosen from the remained feature vector, or first- and second-level feature vectors as defined above. The analysis shows that clusters generated from these vectors are able to determine very different patient survival times.

Several techniques are available to choose the appropriate number of clusters and they depend on the dataset and the values of the attributes. One of the most common methods for choosing the number of clusters is called elbow criterion [158, 156]. Another technique selects a threshold which defines the appropriate number of instances in each cluster. If the number of instances is less or greater

than a particular threshold, then the clusters are merged or split, respectively [70]. The elbow criterion is typically applied when the dataset instances are very large (e.g., 106 records). Since the dataset is not this large, zero is selected as the threshold to merge the clusters because the number of instances in each generated cluster is meaningful (i.e., the numbers are comparable to each other). The values of the features of many instances are similar to each other, so the clusters needs to merge based on a predefined threshold. Thus, if the number of instances in a cluster is zero then a new generated cluster is not expected.

E Multimodal clustering

A variety of mathematical models can be applied to fit a dataset. The approach is to cluster the dataset and find an optimized model for it. The multimodal library developed in R is used to select the optimized expectation maximization (EM) technique to analyze the data. EM is a density-based (statistical) technique which assumes that points belonging to each cluster are drawn from a specific probability distribution. The component densities could be multivariate Gaussian (in case of numeric data, e.g., tumor size) or multimodal (in case of categorical data, e.g., marital status). The overall distribution of the data is thus assumed to be a mixture of several distributions. The aim of the multimodal technique is to identify the clusters and their parameters. One of the prevalent means for this identification is using maximum likelihood. In this case, the parameters and the probability distribution of the data are chosen such that the parameters are maximized.

The application of EM involves two steps to find the best parameters. The first step is expectation and the second step is maximization. In the statistical domain the parameters can be defined as mean, variance or higher moments of the given dataset. The expectation step computes the parameters such as mean and variance of the distribution, thus calculating the conditional expectation of the complete data using the observed data and parameter estimations. The maximization step maximizes the complete data log likelihood from the expectation

step. The two steps are run iteratively until they converge. The EM algorithm is defined as follows:

X is the incomplete data, y is the complete data and is the parameter vector (e.g., mean, variance or higher moments). If at initialization is given, the algorithm updates until changes are small. The expectation and maximization steps perform the changes to update the feature vector. Expectation-Step:

$$Q(\theta|\theta(k)) = E[\ln P(y|\theta)|y, \theta(k)] \tag{34}$$

Maximization-Step:

$$\theta(k+1) = argmax Q(\theta|\theta(k)) \tag{35}$$

$\theta(k)$ is the parameter vector in the k-iteration. $P(y|\theta)$ is the probability density function of clusters.

The results of applying the multimodal clustering are summarized and compared with survival rates below. Figure 25 compares the density of clusters vs. survival rate. The left plot shows that by using the second-level features, the peaks of the density of the clusters mostly have the same range of survival time. The right plot shows that by using first-level features, the peaks of the clusters are very distinct, indicating that the instances can be well separated based on their survival time. Thus, it shows that the resulting clusters provide differentiation within the survival times, showing that the clusters are separable when choosing first-level features but not with second-level features.

F Hierarchical Clustering

Hierarchical clustering is based on finding a hierarchy of features and creating a tree of these features based on a distance matrix between instances. The method creates a hierarchical decomposition of the set of objects using information gain and entropy. The method iteratively merges two close groups until all the data are merged into a single cluster. Each level of the resulting tree represents a set of clusters of the data. The tree is cut (i.e., the level to be the best set of clusters is chosen) based on the threshold zero so that the number of instances in each cluster

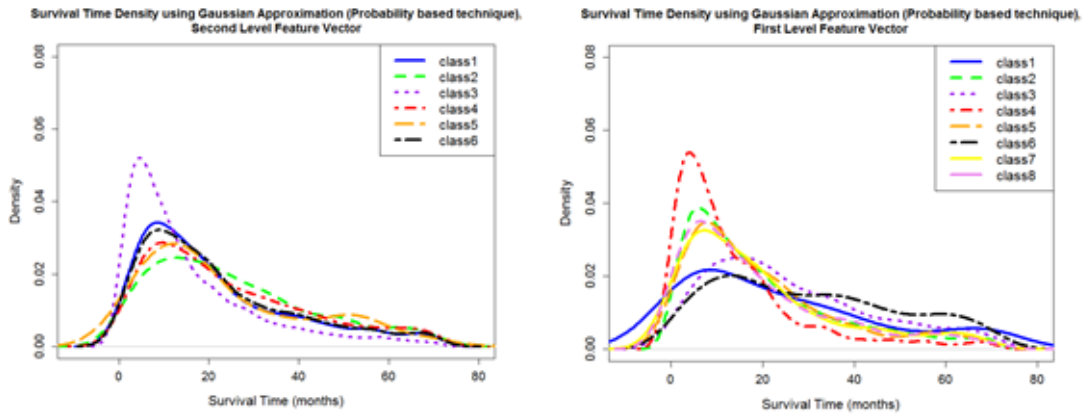

Figure 25. Survival Time determined using multimodal clustering technique. The left panel shows the clusters of first-level features, the right panel shows the clusters of second-level features.

of the set is meaningful. The similarity between instances within a cluster is determined based on the distance metric. Three popular similarity metrics include single-linkage, complete linkage and group average. The complete linkage is chosen to analyze the dataset, in which the distance between two clusters is the maximum of all pairwise distances between pairs in two clusters. This is in contrast to the single linkage method, in which the distance between two clusters is the minimum set of distances between all pairs of patterns drawn from the two clusters. Single linkage uses smallest dissimilarity between two points in opposite group and complete linkage uses largest dissimilarity between two points in the opposite group, finally average uses the average dissimilarity measure points in two opposite groups. The single linkage only need one pair of points to be close, however the complete linkages scoring is based on worst-case dissimilarity pairs so the clusters are impacts but some points are found that are closer to points in other clusters than its own cluster. The average clustering uses average pairwise dissimilarity. Average linkage is not used because the result of single and complete linkage clustering are unchanged under monotone transformation of dissimilarity [99].

The single-linkage finds the similarity of the closest point:

$$d_{SL}(G, H) = min_{i \in G, j \in H} d_{i,j} \tag{36}$$

The complete-linkage finds the similarity between the furthest pair:

$$d_{CL}(G, H) = max_{i \in G, j \in H} d_{i,j} \tag{37}$$

The group average finds the similarity between groups:

$$d_{GA} = \frac{1}{N_G, N_H} \sum_{i \in G} \sum_{j \in H} d_{i,j} \tag{38}$$

where i and j are instances selected from two clusters G and H. $d_{i,j}$ is the Euclidean distance between instances I and j. N_G and N_H are the number of the instance in the two clusters G and H.

A hierarchical clustering package developed in R-language is used. The tree is cut to obtain the highest number of separable clusters, which for this dataset is five.

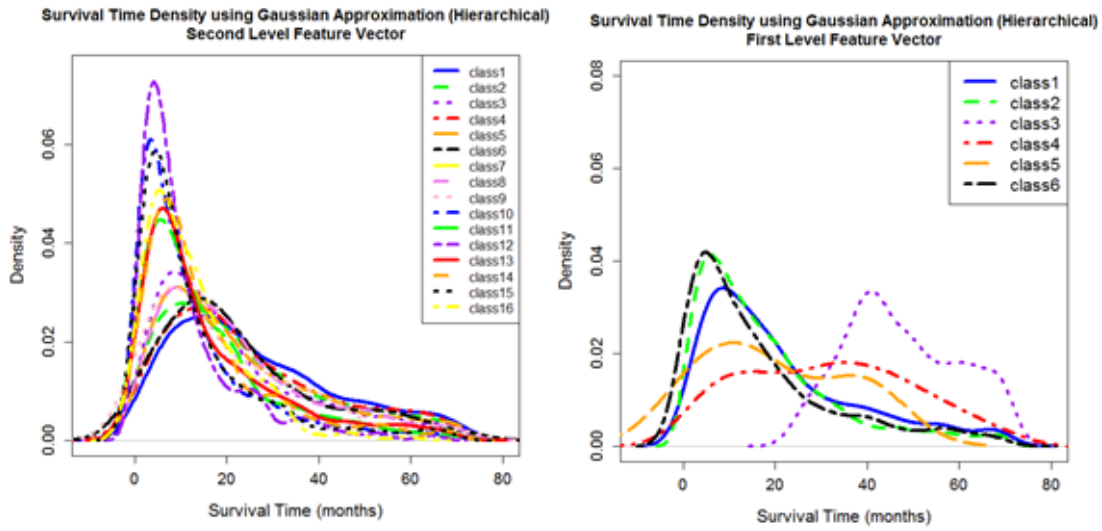

Figure 26. Survival Time determined using Hierarchical clustering technique. The left column show the clusters of second-level features, the right column show the clusters of first-level features.

Figure 26 (left) shows the results obtained by choosing the second-level feature vector. In this case, the density peaks of the clusters vs. survival time are within the same range, providing no discrimination between them. Figure26 (right) shows that by choosing the first-level feature vector, the density peak of the clusters are clearly separated, indicating that the survival time of patients can be well separated based on these features.

G Non-Negative Matrix Factorization (NMF)

NMF mathematically factorizes a matrix into two other matrices using linear algebra and other available techniques to find the basis vectors of each cluster. The resulting clustering depends on the method used to factorize the matrices. This clustering does not depend on the Euclidean distance of the samples in the feature space. One of the factorized matrices is the basis vector of the clusters and the other one is the weighted matrix of the basis vectors. Principal component analysis (PCA) is mostly used to analyze the data in a new space, data are mapped into the space which is generated based on the eigen vectors of the original database. The mapped data have n-dimension which n is the number of components which are chosen to map the dataset. In PCA domain each axis are orthogonal to each other, however NMF is using similar technique in the other way such that the axis of the new space are not necessarily orthogonal to each other. This new technique map data to a better space which has more direction information of the basis functions [4, 70].

Assume that the input data matrix is $X = (x_1, x_2, \ldots, x_n)$, where the columns indicate the feature vector and the rows represent the instances of the dataset. The input matrix is factorized into two matrices:

$$X \approx FG^T \tag{39}$$

where $X \in R^{p \times n}$, $F \in R^{p \times K}$ and $G \in R^{n \times k}$. Generally, the rank of matrices F and G is much lower than the rank of X, such that $k \ll min(p, n)$. F represents the centroids of clusters or the basis vectors that defines the clusters, while G represents the weighting matrix which could be defined as the ranking of the instances relative

to the basis vectors. Some of the matrix factorization techniques include:

1. SVD: Singular Value Decomposition. Principal Component Analysis (PCA) uses singular value decomposition.

$$X_{\pm} = U_{\pm}V_{\pm} \tag{40}$$

2. NMF: Non-negative matrix factorization. The input matrix is limited to only have non-negative signs.

$$X_{+} = F_{+}G_{+} \tag{41}$$

3. Semi NMF: The input data has both negative and positive signs.

$$X_{\pm} = F_{\pm}G_{+} \tag{42}$$

4. Convex-NMF: The elements of the F matrix can be any integer in a large space. In order to capture a better definition of the centroids of the clusters the matrix can be spanned by the columns of X, i.e:

$$f_l = w_{1l}x_1 + w_{2l} + \ldots + w_{nl}x_n or F = XW \tag{43}$$

where f_l is a convex combination $w_{ij} > 0$ of the data points.

5. Tri-Factorization: This technique clusters both the rows and columns of the clusters simultaneously.

$$X_{+} = F_{+}S_{+}G_{+}^{T} \tag{44}$$

The factorization is in such a way that F gives row clusters while G gives column clusters.

6. Kernel NMF: This method is based on a mapping function $\varphi(\cdot)$ The kernel is defined as: $k = \varphi^{T}(x)\varphi(x)$. The factorization equation is written as:

$$\varphi(X_{\pm}) \approx \varphi(X_{\pm})W_{+}G_{+}^{T} \tag{45}$$

TABLE 11. Different matrix factorization techniques.

	Technique	Matrix Factorization
1	SVD	$X_{\pm} = U_{\pm}V_{\pm}$
2	NMF	$X_{+} = F_{+}G_{+}$
3	Semi-NMF	$X_{\pm} = F_{\pm}G_{+}$
4	Convex-NMF	$f_l = w_{1l}x_1 + w_{2l} + \ldots + w_{nl}x_n or F = XW$
5	Kernel-NMF	$k = \varphi^T(x)\varphi(x)$
6	Tri-Factorization	$X_{+} = F_{+}S_{+}G_{+}^{T}$

In summary, the various matrix factorization matrix techniques include [99, 20]:

NMF package in R is used to apply two NMF-specific techniques to obtain the factorized matrix, namely alternating least squares and multinomial method. The results obtained through either technique were not significantly different. Multinomial method is chosen because it is a general version of probabilistic model, the nature of the dataset shows that probabilistic information can yield more accurate information. Figure27 (left) shows that the results obtained by choosing the second-level feature vector are unable to separate the density peaks of the clusters vs. survival time. Figure27(right) shows that by choosing the first-level feature vector, the density peak of the clusters are clearly separated.

The clustering results are compared not based on density of the survival rate but on the features of the instances within each cluster. Interestingly, the multimodal and hierarchical clustering show similar Principal Component Analysis (PCA) scores for instances in the same cluster, while NMF indicates that the first-level feature vector is the best choice for clustering patients.

H Analyzing the Results of Clustering Technique Based on PCA Scores

PCA is a technique mostly used for dimensional reduction, in which data in multiple dimensions are mapped into a lower dimensional space. The technique is based on linear combination of orthogonal variables, for which the basis shows the pattern of the data in orthogonal directions. Either spectral decomposition of the correlation matrix or singular value decomposition of the data matrix is performed to obtain linear combinations which are called principal components, where the

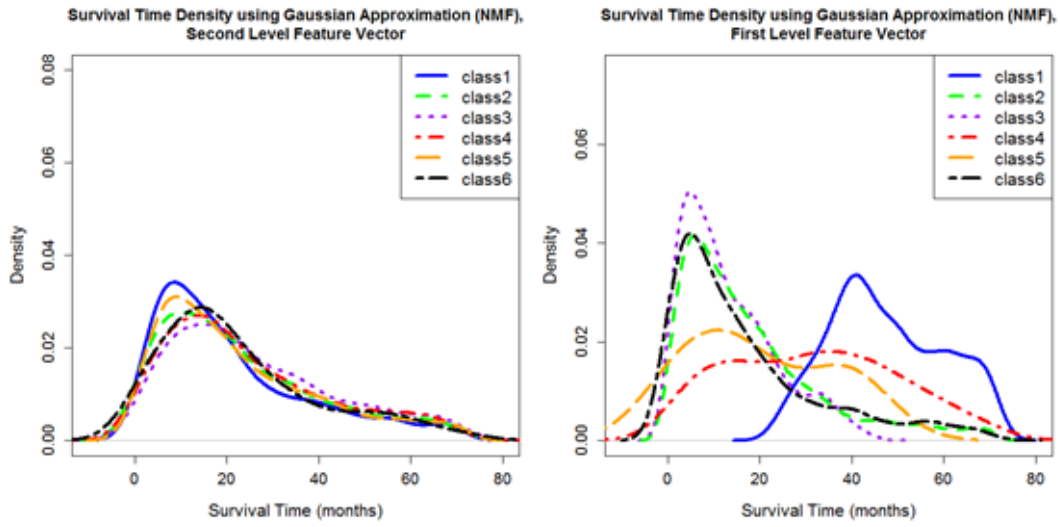

Figure 27. Survival time determined using Non-Negative Matrix Factorization (NMF) technique. The left column show the clusters of second-level features, the right column show the clusters of first-level features.

weights of each original variable in the principal component are called scores. Each sample of the dataset has a value for each feature, and for each feature the principal components are calculated. The first two largest components are chosen and the score for each sample is calculated. The two highest scores are thus calculated for each sample.

Figures 25, 26, 27 show that using multimodal and hierarchical clustering can create distinct clusters based on either first- or second-level feature vectors. This suggests that density- and distance-based methods are unable to distinguish between the two feature vectors based on separation of clusters. The NMF method, however, is able to create distinct clusters based on the first-level features but not based on the second-level features. NMF uses matrix factorization instead of probabilistic similarity metric (which multimodal clustering uses) or Euclidian similarity distance metric (which hierarchical clustering uses). The reason is that the values of both feature vectors have categorical variables and the two latter techniques cannot distinguish between patients with similar feature vector values when a similarity metric is applied.

The left panels in Figures 28, 29 and 30 plot the relation between the two highest PCA scores for first- and second-level features. For both multimodal (Figure 28) and hierarchical (Figure 29) clustering, the second-level feature vector (left panels) shows that the scores of the classes are in the same range to each other in principal component score space, and the scores are separated. In contrast, the first-level feature vector PCA scores (right panels) show that the instances in the same class have scores spanning the range of the horizontal dimension) and that the clusters are inter-mixed. Thus, although the scores obtained with the second-level feature vector show separation with both multimodal and hierarchical clustering, the clusters do not contain sufficient information to distinguish data points based on their survival time as shown in Figures 28, 29 and 30 (left panels).

For NMF clustering, the PCA scores for the second-level features (left) are completely mixed, while for the first-level features (right) the scores are separated in distinct groups. Since this technique enables distinguishing patients based on

TABLE 12. Columns 1 and 2 show the mean of each cluster (center of the clusters) for the three chosen clustering techniques. First column is based on second level feature vector and column 2 is based on first level feature vector.

	Technique	Matrix Factorization
Hierarchical	13.22	0
	17.93	23.10
	47.55	26.18
	31.48	24.5
	20.0	22.68
	16.83	22.71
		21.24
		21.75
		20.067
		13.04
		15.70
		11.43
		16.51
		15.54
		13.35
		12.65
NMF	14.38	10.86
	31.42	19.69
	10.94	24.57
	11.25	21.88
	56.01	21.89
	10.006	14.11
Probability Based Technique	24.03	24.05
	19.45	25.33
	19.81	15.58
	13.69	24.29
	19.53	23.14
	30.15	21.99
	19.33	
	18.98	

survival time (Figure 30), the separation of PCA scores suggests that NFM clustering enables distinguishing between patients based on similar PCA scores which reflect similar survival times.

The results in Figures 28, 29 and 30 are a visual summary of the density of survival time for each class. Table 12 summarizes the mean of each cluster for the two feature vectors. The same clustering techniques are applied to the two vectors. The first level feature vector shows more accurate results to cluster the patients records. The lack of separation between the clusters in the second level feature vector show that features such as stage and grade cannot discriminate between the survival time of lung cancer patients, while the first level feature vector can better distinguish between patient survival. Table 12 shows the mean of each cluster assumed to be at the center of each class. The variance of the center is higher for the second level feature vector when each clustering technique is applied.

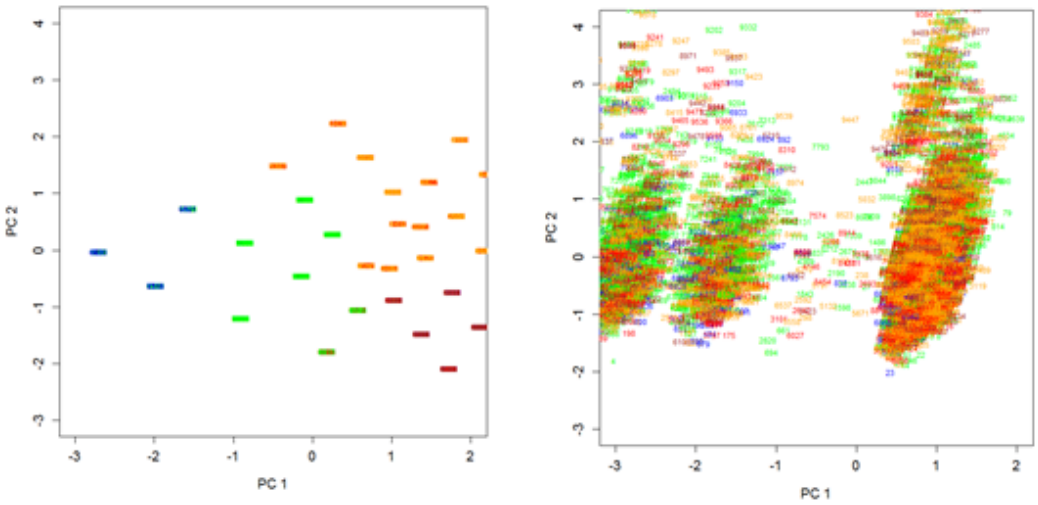

Figure 28. Relation between two highest PCA scores and classes from multimodal clustering. The left panel show the clusters of first-level features, while the right panel shows the clusters of second-level features. The clusters are shown with different colors. Scores sharing the same color belong to the same cluster.

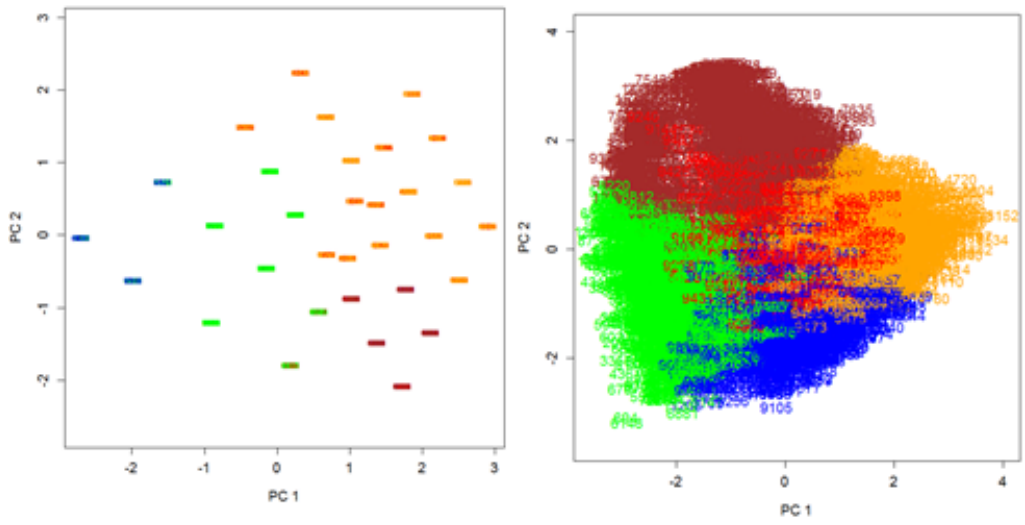

Figure 29. Relation between two highest PCA scores and classes from Hierarchical clustering. The left panel show the clusters of first-level features, while the right panel shows the clusters of second-level features. The clusters are shown with different colors. Scores sharing the same color belong to the same cluster.

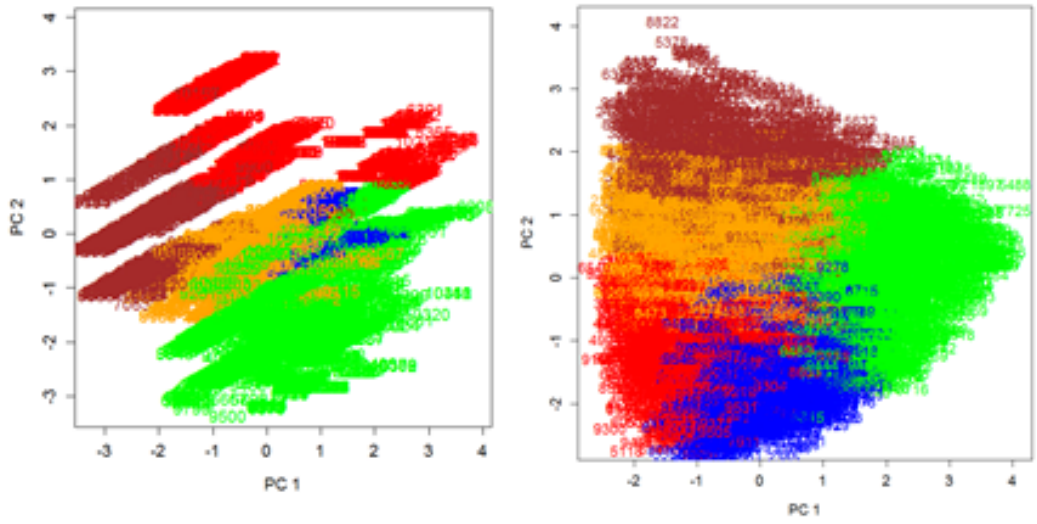

Figure 30. Relation between two highest PCA scores and classes from NMF clustering. The left panel show the clusters of first-level features, while the right panel shows the clusters of second-level features. Scores sharing the same color belong to the same cluster.

The results show that NMF clustering technique is the best choice for clustering this dataset and the second best technique is Hierarchical technique and finally the probability based technique which the results for two feature vectors are not very different.

I Conclusions

The results show that the current patient staging system based on Grade and Stage cannot adequately discriminate lung cancer patients to reliable predict their survival time. Three different types of clustering techniques are applied to show that second-level features such as cancer stage and grade are insufficient to establish a patient's survival time in lung cancer. Three groups of clustering are chosen: based on distance of instances in the feature space (hierarchical clustering), using the probability information of the dataset (multimodal clustering) which chooses the best modified model of expectation maximization, and matrix factorization which does not use similarity techniques. NMF shows that the center of the clusters have higher difference in survival time compared with the other two techniques, and the first level feature vector get more promising clusters than the second level feature vector in all three clustering techniques. The results show that the clinically common criterion of cancer stage is not correlated with patient survival time. The separation of patient features into first and second-level as defined in this study could be applied to analyze survival for other cancer types in the SEER database. [4]

CHAPTER VII

CONCLUSIONS AND FUTURE WORKS

Medical data analysis requires diverse applications. The input data and its complexity can be analyzed to determine the applicable framework. Two of the applications describe two different methodologies analyzing perfusion curves. One of them is a dataset of perfusion curves collected from DCE-MR images and the other dataset utilizes IVM collected from vessel perfusion. Features of interest were directly extracted from the measured datasets using a mathematical fitting technique and applying the most efficient and accurate classification algorithms. FCM and KNN were chosen because FCM is the best to rank each sample while KNN shows the best ROC value.

Personalization of healthcare datasets using machine learning techniques is a new idea which could help clinicians choose the best treatment and predict the best status of a particular patient based on the history of large number of patients. The techniques currently used and potential algorithms that are a combination of mathematical tumor models and medical images are reviewed. It is shown that the information of the location of the tumor in the body could be used to predict the speed of the tumor growth. In the future works a large dataset of this information from previous patients could be used to develop machine learning techniques to adapt the parameters of the tumor growth models.

The personalization project seemed to be hard to develop a large dataset of patients with lung cancer was used to be able to predict the survival time using machine learning techniques. The results are compared, and it finds that SVM has the best ROC value, and applied several clustering techniques. Comparison of the results and similarity of the records considering survival time indicate the best

feature vector to represent the patients and predict their survival time. The next step would be to develop a new Kernel function based on two analyzed feature vectors and apply the customized SVM which would make a connection between the clinical features that predict the survival time and also those that are available in the SEER database.

3D vessel segmentation and reconstruction is a difficult problem. A new vessel enhancement technique is proposed for MRA-TOF images. The proposed technique uses the strengths of the non-linear diffusion filter in finding homogenous regions in images while combining it with EM algorithm to brighten the vessels and glooming the background.

The experience with several practical healthcare datasets indicates that a novel solution could be found to combine all of the clinical information. The large number of factors requires new techniques such as machine learning algorithms to analyze large datasets in a way to find the pattern of similar patients and also to analyze them beyond just a statistical analysis. The relation between treatment and patients, survival time and treatment efficiency, could be considered more accurately with machine leaning. Personalization of treatment for each patient in this manner may thus help to cure the disease.

REFERENCES

[1] "Introduction to lung cancer.," URL http://www.cancer.gov/cancertopics/types/lung/, 2013, National Cancer Institute, SEER training modules.

[2] "Overview of the seer program. surveillance epidemiology and end results.," URL http://seer.cancer.gov/about/, 2013.

[3] "Seer training modules-introduction to lunc cancer.," URL http://training.seer.cancer.gov/lung/intro/, 2013, National Cancer Institute, SEER training modules.

[4] Herv Abdi and Lynne J Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[5] Behnaz Abdollahi, Neal Dunlap, and Hermann B Frieboes, *Bridging the Gap Between Modeling of Tumor Growth and Clinical Imaging*, pp. 463–487, Springer, 2014.

[6] Behnaz Abdollahi, Ayman El-Baz, and Amir A Amini, "A multi-scale non-linear vessel enhancement technique," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. 2011, pp. 3925–3929, IEEE.

[7] Ankit Agrawal and Alok Choudhary, "Association rule mining based hotspot analysis on seer lung cancer data," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, vol. 2, no. 2, pp. 34–54, 2011.

[8] Ankit Agrawal and Alok Choudhary, "Identifying hotspots in lung cancer data using association rule mining," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. pp. 995–1002, IEEE.

[9] Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary, "A lung cancer outcome calculator using ensemble data mining on seer data," in *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics*. p. 5, ACM.

[10] Rakesh Agrawal, Tomasz Imieliski, and Arun Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD Record*. vol. 22, pp. 207–216, ACM.

[11] Alexander RA Anderson and Vito Quaranta, "Integrative mathematical oncology," *Nature Reviews Cancer*, vol. 8, no. 3, pp. 227–234, 2008.

[12] Samuel G Armato, Feng Li, Maryellen L Giger, Heber MacMahon, Shusuke Sone, and Kunio Doi, "Lung cancer: Performance of automated lung nodule detection applied to cancers missed in a ct screening program 1," *Radiology*, vol. 225, no. 3, pp. 685–692, 2002.

[13] Sergey Astanin and Luigi Preziosi, *Multiphase models of tumour growth*, pp. 1–31, Springer, 2008.

[14] Stephen Aylward, Elizabeth Bullitt, S Pizer, and David Eberly, "Intensity ridge and widths for tubular object segmentation and description," in *Mathematical Methods in Biomedical Image Analysis, 1996., Proceedings of the Workshop on*. 1996, pp. 131–138, IEEE.

[15] Stephen R Aylward and Elizabeth Bullitt, "Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction," *Medical Imaging, IEEE Transactions on*, vol. 21, no. 2, pp. 61–75, 2002.

[16] Amrit Bhaskarla, Paul C Tang, Terry Mashtare, Chukwumere E Nwogu, Todd L Demmy, Alex A Adjei, Mary E Reid, and Sai Yendamuri, "Analysis of second primary lung cancers in the seer database," *Journal of Surgical Research*, vol. 162, no. 1, pp. 1–6, 2010.

[17] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. pp. 144–152, ACM.

[18] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, *Classification and regression trees*, CRC press, 1984.

[19] Gunnar Brix, Malte L Bahner, Ulf Hoffmann, Andrea Horvath, and Wolfgang Schreiber, "Regional blood flow, capillary permeability, and compartmental volumes: measurement with dynamic ctinitial experience," *Radiology*, vol. 210, no. 1, pp. 269–276, 1999.

[20] Wray Buntine, *Variational extensions to EM and multinomial PCA*, pp. 23–34, Springer, 2002.

[21] Helen M Byrne, "Dissecting cancer through mathematics: from the cell to the animal model," *Nature Reviews Cancer*, vol. 10, no. 3, pp. 221–230, 2010.

[22] Francine Catt, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll, "Image selective smoothing and edge detection by nonlinear diffusion," *SIAM Journal on Numerical analysis*, vol. 29, no. 1, pp. 182–193, 1992.

[23] Antoinette A Chan and Sarah J Nelson, "Simplified gamma-variate fitting of perfusion curves," in *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*. pp. 1067–1070, IEEE.

[24] Jian Chen and Amir A Amini, "Quantifying 3-d vascular structures in mra images using hybrid pde and geometric deformable models," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 10, pp. 1251–1262, 2004.

[25] Mary Cianfrocca and Lori J Goldstein, "Prognostic and predictive factors in early-stage breast cancer," *The Oncologist*, vol. 9, no. 6, pp. 606–616, 2004.

[26] Olivier Clatz, Maxime Sermesant, P-Y Bondiau, Herv Delingette, Simon K Warfield, Grgoire Malandain, and Nicholas Ayache, "Realistic simulation of

the 3-d growth of brain tumors in mr images coupling diffusion with biomechanical deformation," *Medical Imaging, IEEE Transactions on*, vol. 24, no. 10, pp. 1334–1346, 2005.

[27] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] David Cosgrove and Nathalie Lassau, "Imaging of perfusion using ultrasound," *European journal of nuclear medicine and molecular imaging*, vol. 37, no. 1, pp. 65–85, 2010.

[29] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[30] Danil de Kok and Harm Brouwer, "Natural language processing for the working programmer," 2011.

[31] Thomas S Deisboeck, Le Zhang, Jeongah Yoon, and Jose Costa, "In silico cancer modeling: is it ready for prime time?," *Nature Clinical Practice Oncology*, vol. 6, no. 1, pp. 34–42, 2008.

[32] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[33] Thomas Deschamps and Laurent D Cohen, "Fast extraction of minimal paths in 3d images and applications to virtual endoscopy," *Medical image analysis*, vol. 5, no. 4, pp. 281–299, 2001.

[34] Sean T Duggan and Gillian M Keating, "Pegylated liposomal doxorubicin," *Drugs*, vol. 71, no. 18, pp. 2531–2558, 2011.

[35] Lisa K Dunnwald, Julie R Gralow, Georgiana K Ellis, Robert B Livingston, Hannah M Linden, Jennifer M Specht, Robert K Doot, Thomas J Lawton,

William E Barlow, and Brenda F Kurland, "Tumor metabolism and blood flow changes by positron emission tomography: relation to survival in patients treated with neoadjuvant chemotherapy for locally advanced breast cancer," *Journal of Clinical Oncology*, vol. 26, no. 27, pp. 4449–4457, 2008.

[36] Lynne Eldridge, "Lung cancer survival rates by type and stage," 2013.

[37] Heiko Enderling, Mark AJ Chaplain, Alexander RA Anderson, and Jayant S Vaidya, "A mathematical model of breast cancer development, local treatment and recurrence," *Journal of theoretical biology*, vol. 246, no. 2, pp. 245–259, 2007.

[38] Andinet Enquobahrie, Luis Ibanez, Elizabeth Bullitt, and Stephen Aylward, "Vessel enhancing diffusion filter," *The Insight Journal*, 2007.

[39] Selim Esedolu, "An analysis of the peronamalik scheme," *Communications on Pure and Applied Mathematics*, vol. 54, no. 12, pp. 1442–1487, 2001.

[40] Jun Fang, Hideaki Nakamura, and Hiroshi Maeda, "The epr effect: unique features of tumor blood vessels for drug delivery, factors involved, and limitations and augmentation of the effect," *Advanced drug delivery reviews*, vol. 63, no. 3, pp. 136–151, 2011.

[41] Andreas Fieselmann, Markus Kowarschik, Arundhuti Ganguly, Joachim Hornegger, and Rebecca Fahrig, "Deconvolution-based ct and mr brain perfusion measurement: theoretical model revisited and practical implementation details," *Journal of Biomedical Imaging*, vol. 2011, pp. 14, 2011.

[42] B Fischl and EL Schwartz, "Learned adaptive nonlinear filtering for anisotropic diffusion approximation in image processing," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on.* 1996, vol. 4, pp. 276–280, IEEE.

[43] Alejandro F Frangi, Wiro J Niessen, Romhild M Hoogeveen, Theo Van Walsum, and Max A Viergever, "Model-based quantitation of 3-d magnetic resonance angiographic images," *Medical Imaging, IEEE Transactions on*, vol. 18, no. 10, pp. 946–956, 1999.

[44] Alejandro F Frangi, Wiro J Niessen, Romhild M Hoogeveen, Theo van Walsum, and Max A Viergever, "Quantitation of vessel morphology from 3d mra," in *Medical Image Computing and Computer-Assisted InterventionMICCAI99*. 1999, pp. 358–367, Springer.

[45] Alejandro F Frangi, Wiro J Niessen, Paul J Nederkoorn, Otto EH Elgersma, and Max A Viergever, "Three-dimensional model-based stenosis quantification of the carotid arteries from contrast-enhanced mr angiography," in *Mathematical Methods in Biomedical Image Analysis, 2000. Proceedings. IEEE Workshop on*. 2000, pp. 110–118, IEEE.

[46] Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever, *Multiscale vessel enhancement filtering*, pp. 130–137, Springer, 1998.

[47] Yonatan Fridman, *Extracting branching object geometry via cores*, Ph.D. thesis, University of North Carolina at Chapel Hill, 2004.

[48] Yonatan Fridman, Stephen M Pizer, Stephen Aylward, and Elizabeth Bullitt, *Segmenting 3D branching tubular structures using cores*, pp. 570–577, Springer, 2003.

[49] Hermann B Frieboes, Mark AJ Chaplain, Alastair M Thompson, Elaine L Bearer, John S Lowengrub, and Vittorio Cristini, "Physical oncology: a bench-to-bedside quantitative and predictive approach," *Cancer research*, vol. 71, no. 2, pp. 298–302, 2011.

[50] Jennifer B Fu, T Ying Kau, Richard K Severson, and Gregory P Kalemkerian, "Lung cancer in womenanalysis of the national surveillance, epidemiology, and end results database," *CHEST Journal*, vol. 127, no. 3, pp. 768–777, 2005.

[51] Biana Godin, Ennio Tasciotti, Xuewu Liu, Rita E Serda, and Mauro Ferrari, "Multistage nanovectors: from concept to novel imaging contrast agents and therapeutics," *Accounts of chemical research*, vol. 44, no. 10, pp. 979–989, 2011.

[52] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore, "Understanding survival analysis: Kaplan-meier estimate," *International journal of Ayurveda research*, vol. 1, no. 4, pp. 274, 2010.

[53] Hana LP Harpold, Ellsworth C Alvord Jr, and Kristin R Swanson, "The evolution of mathematical modeling of glioma proliferation and invasion," *Journal of Neuropathology Experimental Neurology*, vol. 66, no. 1, pp. 1–9, 2007.

[54] Hiroya Hashizume, Peter Baluk, Shunichi Morikawa, John W McLean, Gavin Thurston, Sylvie Roberge, Rakesh K Jain, and Donald M McDonald, "Openings between defective endothelial cells explain tumor vessel leakiness," *The American journal of pathology*, vol. 156, no. 4, pp. 1363–1380, 2000.

[55] Matthew J Hayat, Nadia Howlader, Marsha E Reichman, and Brenda K Edwards, "Cancer statistics, trends, and multiple primary cancer analyses from the surveillance, epidemiology, and end results (seer) program," *The Oncologist*, vol. 12, no. 1, pp. 20–37, 2007.

[56] Marti A. Hearst, ST Dumais, E Osman, John Platt, and Bernhard Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[57] Laurent Hermoye, Laurence Annet, PH Lemmerling, Frank Peeters, Franois Jamar, Pierre Gianello, Sabine Van Huffel, and BE Van Beers, "Calculation of the renal perfusion and glomerular filtration rate from the renal impulse response obtained with mri," *Magnetic resonance in medicine*, vol. 51, no. 5, pp. 1017–1025, 2004.

[58] Lynn Hlatky, Philip Hahnfeldt, and Judah Folkman, "Clinical application of antiangiogenic therapy: microvessel density, what it does and doesn't tell us," *Journal of the National Cancer Institute*, vol. 94, no. 12, pp. 883–893, 2002.

[59] Susan K Hobbs, Wayne L Monsky, Fan Yuan, W Gregory Roberts, Linda Griffith, Vladimir P Torchilin, and Rakesh K Jain, "Regulation of transport pathways in tumor vessels: role of tumor type and microenvironment," *Proceedings of the National Academy of Sciences*, vol. 95, no. 8, pp. 4607–4612, 1998.

[60] Cosmina Hogea, Christos Davatzikos, and George Biros, *Modeling glioma growth and mass effect in 3D MR images of the brain*, pp. 642–650, Springer, 2007.

[61] Cosmina Hogea, Christos Davatzikos, and George Biros, "An image-driven parameter estimation problem for a reactiondiffusion glioma growth model with mass effects," *Journal of mathematical biology*, vol. 56, no. 6, pp. 793–825, 2008.

[62] Heba Ezzat Ibrahim, Sherif M Badr, and Mohamed A Shaheen, "Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems," *arXiv preprint arXiv:1210.7650*, 2012.

[63] Anil K Jain, M Narasimha Murty, and Patrick J Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[64] R Jain, "Perfusion ct imaging of brain tumors: an overview," *American Journal of Neuroradiology*, vol. 32, no. 9, pp. 1570–1577, 2011.

[65] Rakesh K Jain, "Transport of molecules across tumor vasculature," *Cancer and Metastasis Reviews*, vol. 6, no. 4, pp. 559–593, 1987.

[66] Rakesh K Jain, "Determinants of tumor blood flow: a review," *Cancer research*, vol. 48, no. 10, pp. 2641–2658, 1988.

[67] Rakesh K Jain, Lance L Munn, and Dai Fukumura, "Dissecting tumour pathophysiology using intravital microscopy," *Nature Reviews Cancer*, vol. 2, no. 4, pp. 266–276, 2002.

[68] Rakesh K Jain and Triantafyllos Stylianopoulos, "Delivering nanomedicine to solid tumors," *Nature reviews clinical oncology*, vol. 7, no. 11, pp. 653–664, 2010.

[69] Sad Jbabdi, Emmanuel Mandonnet, Hugues Duffau, Laurent Capelle, Kristin Rae Swanson, Mlanie PlgriniIssac, Rmy Guillevin, and Habib Benali, "Simulation of anisotropic growth of lowgrade gliomas using diffusion tensor imaging," *Magnetic Resonance in Medicine*, vol. 54, no. 3, pp. 616–624, 2005.

[70] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.

[71] Walid S Kamoun, Sung-Suk Chae, Delphine A Lacorre, James A Tyrrell, Mariela Mitre, Marijn A Gillissen, Dai Fukumura, Rakesh K Jain, and Lance L Munn, "Simultaneous measurement of rbc velocity, flux, hematocrit and shear rate in vascular networks," *Nature methods*, vol. 7, no. 8, pp. 655–660, 2010.

[72] NS Kapadia, FD Vigneau, WO Quarshie, AG Schwartz, and FP Kong, "Patterns of practice and outcomes for stage i non-small cell lung cancer (nsclc): Analysis of seer-17 data, 1999-2008," *International Journal of Radiation Oncology* Biology* Physics*, vol. 84, no. 3, pp. S545, 2012.

[73] Michael Kass, Andrew Witkin, and Demetri Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[74] Stephen J Kennel, Ila A Davis, John Branning, Hongjun Pan, George W Kabalka, and Michael J Paulus, "High resolution computed tomography and mri for monitoring lung tumor growth in mice undergoing radioimmunotherapy: correlation with histology," *Medical physics*, vol. 27, no. 5, pp. 1101–1107, 2000.

[75] Fahmi Khalifa, Mohamed Abou ElGhar, Behnaz Abdollahi, Hermann B Frieboes, Tarek ElDiasty, and Ayman ElBaz, "A comprehensive noninvasive framework for automated evaluation of acute renal transplant rejection using dcemri," *NMR in Biomedicine*, vol. 26, no. 11, pp. 1460–1470, 2013.

[76] Fahmi Khalifa, Ayman El-Baz, Georgy Gimelfarb, and Mohammed Abu El-Ghar, *Non-invasive image-based approach for early detection of acute renal rejection*, pp. 10–18, Springer, 2010.

[77] Satyanad Kichenassamy, Arun Kumar, Peter Olver, Allen Tannenbaum, and Anthony Yezzi, "Gradient flows and geometric active contour models," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*. 1995, pp. 810–815, IEEE.

[78] Taco Kind, Ivo Houtzager, Theo JC Faes, and Mark BM Hofman, "Evaluation of model-independent deconvolution techniques to estimate blood perfusion," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. pp. 2602–2607, IEEE.

[79] Ernst Klotz and Matthias Knig, "Perfusion measurements of the brain: using dynamic ct for the quantitative assessment of cerebral ischemia in acute stroke," *European journal of radiology*, vol. 30, no. 3, pp. 170–184, 1999.

[80] Daphne Koller and Nir Friedman, *Probabilistic graphical models: principles and techniques*, MIT press, 2009.

[81] Thomas Markus Koller, Guido Gerig, Gabor Szekely, and Daniel Dettwiler, "Multiscale detection of curvilinear structures in 2-d and 3-d image data," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*. 1995, pp. 864–869, IEEE.

[82] AA Konstas, GV Goldmakher, T-Y Lee, and MH Lev, "Theoretic basis and technical implementations of ct perfusion in acute ischemic stroke, part 2:

technical implementations," *American Journal of Neuroradiology*, vol. 30, no. 5, pp. 885–892, 2009.

[83] Ender Konukoglu, Olivier Clatz, Bjoern H Menze, Bram Stieltjes, M A Weber, Emmanuel Mandonnet, Hervé Delingette, and Nicholas Ayache, "Image guided personalization of reaction-diffusion type tumor growth models using modified anisotropic eikonal equations," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 1, pp. 77–95, 2010.

[84] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.

[85] Karl Krissian, Grgoire Malandain, and Nicholas Ayache, *Directional anisotropic diffusion applied to segmentation of vessels in 3D images*, Springer, 1997.

[86] Karl Krissian, Grgoire Malandain, Nicholas Ayache, Rgis Vaillant, and Yves Trousset, "Model-based detection of tubular structures in 3d images," *Computer vision and image understanding*, vol. 80, no. 2, pp. 130–171, 2000.

[87] Kohsuke Kudo, Makoto Sasaki, Kei Yamada, Suketaka Momoshima, Hidetsuna Utsunomiya, Hiroki Shirato, and Kuniaki Ogasawara, "Differences in ct perfusion maps generated by different commercial software: Quantitative analysis by using identical source data of acute stroke patients 1," *Radiology*, vol. 254, no. 1, pp. 200–209, 2009.

[88] Philip Kunkel, Ulrike Ulbricht, Peter Bohlen, Marc A Brockmann, Regina Fillbrandt, Dimitrios Stavrou, Manfred Westphal, and Katrin Lamszus, "Inhibition of glioma angiogenesis and growth in vivo by systemic treatment with a monoclonal antibody against vascular endothelial growth factor receptor-2," *Cancer Research*, vol. 61, no. 18, pp. 6624–6628, 2001.

[89] Caroline Lacoste, Xavier Descombes, and Josiane Zerubia, "Point processes for unsupervised line network extraction in remote sensing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1568–1579, 2005.

[90] George Laking and Pat Price, "Radionuclide imaging of perfusion and hypoxia," *European journal of nuclear medicine and molecular imaging*, vol. 37, no. 1, pp. 20–29, 2010.

[91] Brian E Lally, Ann M Geiger, James J Urbanic, Jerome M Butler, Stacy Wentworth, Michael C Perry, Lynn D Wilson, Janet K Horton, Frank C Detterbeck, and Antonius A Miller, "Trends in the outcomes for patients with limited stage small cell lung cancer: An analysis of the surveillance, epidemiology, and end results database," *Lung Cancer*, vol. 64, no. 2, pp. 226–231, 2009.

[92] HBW Larsson, M Stubgaard, JL Frederiksen, M Jensen, O Henriksen, and OB Paulson, "Quantitation of bloodbrain barrier defect by magnetic resonance imaging and gadoliniumdtpa in patients with multiple sclerosis and brain tumors," *Magnetic Resonance in Medicine*, vol. 16, no. 1, pp. 117–131, 1990.

[93] HBW Larsson and PS Tofts, "Measurement of bloodbrain barrier permeability using dynamic gddtpa scanninga comparison of methods," *Magnetic Resonance in Medicine*, vol. 24, no. 1, pp. 174–176, 1992.

[94] Henrik BW Larsson, Frdric Courivaud, Egill Rostrup, and Adam E Hansen, "Measurement of brain perfusion, blood volume, and bloodbrain barrier permeability, using dynamic contrastenhanced t1weighted mri at 3 tesla," *Magnetic Resonance in Medicine*, vol. 62, no. 5, pp. 1270–1281, 2009.

[95] Max WK Law and Albert CS Chung, *Three dimensional curvilinear structure detection using optimally oriented flux*, pp. 368–382, Springer, 2008.

[96] Max WK Law and Albert CS Chung, "Efficient implementation for spherical flux computation and its application to vascular segmentation," *Image Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 596–612, 2009.

[97] David Lesage, Elsa D Angelini, Isabelle Bloch, and Gareth Funka-Lea, "A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes," *Medical image analysis*, vol. 13, no. 6, pp. 819–845, 2009.

[98] Feng Li, Hidetaka Arimura, Kenji Suzuki, Junji Shiraishi, Qiang Li, Hiroyuki Abe, Roger Engelmann, Shusuke Sone, Heber MacMahon, and Kunio Doi, "Computer-aided detection of peripheral lung cancers missed at ct: Roc analyses without and with localization 1," *Radiology*, vol. 237, no. 2, pp. 684–690, 2005.

[99] Tao Li and Chris Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. pp. 362–371, IEEE.

[100] Yuan Li, Zhi-gang Yang, Tian-wu Chen, Hui-jiao Chen, Jia-yu Sun, and Yan-rong Lu, "Peripheral lung carcinoma: correlation of angiogenesis and first-pass perfusion parameters of 64-detector row ct," *Lung cancer*, vol. 61, no. 1, pp. 44–53, 2008.

[101] Qingfen Lin, "Enhancement, detection, and visualization of 3d volume data," 2001.

[102] Tony Lindeberg, "Edge detection and ridge detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 117–156, 1998.

[103] Zhenqiu Liu, Dechang Chen, Guoliang Tian, Man-Lai Tang, Ming Tan, and Li Sheng, *Efficient support vector machine method for survival prediction with SEER data*, pp. 11–18, Springer, 2010.

[104] Cristian Lorenz, I-C Carlsen, Thorsten M Buzug, Carola Fassnacht, and Jrgen Weese, "Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2d and 3d medical images," in *CVRMed-MRCAS'97*. 1997, pp. 233–242, Springer.

[105] Liana M Lorigo, Olivier D Faugeras, W Eric L Grimson, Renaud Keriven, Ron Kikinis, Arya Nabavi, and C-F Westin, "Curves: Curve evolution for vessel segmentation," *Medical Image Analysis*, vol. 5, no. 3, pp. 195–206, 2001.

[106] John S Lowengrub, Hermann B Frieboes, F Jin, YL Chuang, X Li, Paul Macklin, SM Wise, and Vittorio Cristini, "Nonlinear modelling of cancer: bridging the gap between cells and tumours," *Nonlinearity*, vol. 23, no. 1, pp. R1, 2010.

[107] Paul Macklin, Steven McDougall, Alexander RA Anderson, Mark AJ Chaplain, Vittorio Cristini, and John Lowengrub, "Multiscale modelling and nonlinear simulation of vascular tumour growth," *Journal of mathematical biology*, vol. 58, no. 4-5, pp. 765–798, 2009.

[108] T Soni Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.

[109] Mark T Madsen, "A simplified formulation of the gamma variate function," *Physics in Medicine and Biology*, vol. 37, no. 7, pp. 1597, 1992.

[110] H Maeda, J Wu, T Sawa, Y Matsumura, and K Hori, "Tumor vascular permeability and the epr effect in macromolecular therapeutics: a review," *Journal of controlled release*, vol. 65, no. 1, pp. 271–284, 2000.

[111] Emmanuel Mandonnet, Johan Pallud, Olivier Clatz, Luc Taillandier, Ender Konukoglu, Hugues Duffau, and Laurent Capelle, "Computational modeling of the who grade ii glioma dynamics: principles and applications to management paradigm," *Neurosurgical review*, vol. 31, no. 3, pp. 263–269, 2008.

[112] Rashindra Manniesing, BK Velthuis, MS Van Leeuwen, IC Van der Schaaf, PJ Van Laar, and Wiro J Niessen, "Level set based cerebral vasculature segmentation and diameter quantification in ct angiography," *Medical image analysis*, vol. 10, no. 2, pp. 200–214, 2006.

[113] Rashindra Manniesing, Max A Viergever, and Wiro J Niessen, "Vessel enhancing diffusion: A scale space representation of vessel structures," *Medical Image Analysis*, vol. 10, no. 6, pp. 815–825, 2006.

[114] Yasuhiro Matsumura and Hiroshi Maeda, "A new concept for macromolecular therapeutics in cancer chemotherapy: mechanism of tumoritropic accumulation of proteins and the antitumor agent smancs," *Cancer research*, vol. 46, no. 12 Part 1, pp. 6387–6392, 1986.

[115] Andrew McCallum, Kamal Nigam, and Lyle H Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 169–178, ACM.

[116] Erik Meijering, Wiro Niessen, Joachim Weickert, and Max Viergever, "Evaluation of diffusion techniques for improved vessel visualization and quantification in three-dimensional rotational angiography," in *Medical Image Computing and Computer-Assisted InterventionMICCAI 2001*. 2001, pp. 177–185, Springer.

[117] Bjoern H Menze, Koen Van Leemput, Antti Honkela, Ender Konukoglu, Marc-Andr Weber, Nicholas Ayache, and Polina Golland, "A generative approach for image-based modeling of tumor growth," in *Information processing in medical imaging*. pp. 735–747, Springer.

[118] Charles E Metz, "Statistical analysis of roc data in evaluating diagnostic performance," *Multiple regression analysis: Applications in the health sciences*, , no. 13, pp. 365–384, 1986.

[119] Franziska Michor, "Mathematical models of cancer stem cells," *Journal of Clinical Oncology*, vol. 26, no. 17, pp. 2854–2861, 2008.

[120] Franziska Michor, Jan Liphardt, Mauro Ferrari, and Jonathan Widom, "What does physics have to do with cancer?," *Nature Reviews Cancer*, vol. 11, no. 9, pp. 657–670, 2011.

[121] KA Miles and MR Griffiths, "Perfusion ct: a worthwhile enhancement?," *Perfusion*, vol. 76, no. 904, 2003.

[122] Julien Mille, Romuald Bon, and Laurent D Cohen, *Region-based 2D deformable generalized cylinder for narrow structures segmentation*, pp. 392–404, Springer, 2008.

[123] Tom M Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, 1997.

[124] Tom M Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[125] Ngoc Thanh Nguyen, *New Frontiers in Applied Artificial Intelligence: 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2008 Wroclaw, Poland, June 18-20, 2008, Proceedings*, vol. 5027, Springer, 2008.

[126] Béla Novák and John J Tyson, "Design principles of biochemical oscillators," *Nature reviews Molecular cell biology*, vol. 9, no. 12, pp. 981–991, 2008.

[127] MER Obrien, N Wigler, MCBCSG Inbar, R Rosso, E Grischke, A Santoro, R Catane, DG Kieback, P Tomczak, and SP Ackland, "Reduced cardiotoxicity and comparable efficacy in a phase iii trial of pegylated liposomal doxorubicin hcl (caelyx/doxil) versus conventional doxorubicin for first-line treatment of metastatic breast cancer," *Annals of oncology*, vol. 15, no. 3, pp. 440–449, 2004.

[128] Stanley Osher and James A Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.

[129] Markus R Owen, Toms Alarcn, Philip K Maini, and Helen M Byrne, "Angiogenesis and vascular remodelling in normal and cancerous tissues," *Journal of mathematical biology*, vol. 58, no. 4-5, pp. 689–721, 2009.

[130] Taofeek K Owonikoko, Camille C Ragin, Chandra P Belani, Ana B Oton, William E Gooding, Emanuela Taioli, and Suresh S Ramalingam, "Lung cancer in elderly patients: an analysis of the surveillance, epidemiology, and end results database," *Journal of clinical oncology*, vol. 25, no. 35, pp. 5570–5577, 2007.

[131] David S Paik, Christopher F Beaulieu, R Brooke Jeffrey, Geoffrey D Rubin, and Sandy Napel, "Automated flight path planning for virtual endoscopy," *Medical Physics*, vol. 25, no. 5, pp. 629–637, 1998.

[132] A Parodi, N Quattrocchi, AL van de Ven, C Chiappini, JO Martinez, BS Brown, M Evangelopoulos, SZ Khaled, IK Yazdi, and MV Enzo, "Biomimetic camouflage imparts cell-like activity to synthetic particles," *Nature Nanotechnology*, vol. 8, pp. 61–68, 2013.

[133] Karl Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[134] Pietro Perona and Jitendra Malik, "Scale-space and edge detection using anisotropic diffusion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 7, pp. 629–639, 1990.

[135] G Petralia, L Bonello, S Viotti, L Preda, G d'Andrea, and M Bellomi, "Ct

perfusion in oncology: how to do it," *Cancer Imaging*, vol. 10, no. 1, pp. 8, 2010.

[136] Dzung L Pham, Chenyang Xu, and Jerry L Prince, "Current methods in medical image segmentation 1," *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[137] C Pozrikidis and DA Farrow, "A model of fluid flow in solid tumors," *Annals of biomedical engineering*, vol. 31, no. 2, pp. 181–194, 2003.

[138] S Ramalingam, K Pawlish, S Gadgeel, R Demers, and GP Kalemkerian, "Lung cancer in young patients: analysis of a surveillance, epidemiology, and end results database," *Journal of clinical oncology*, vol. 16, no. 2, pp. 651–657, 1998.

[139] Ignacio Ramis-Conde, Mark AJ Chaplain, Alexander RA Anderson, and Dirk Drasdo, "Multi-scale modelling of cancer cell intravasation: the role of cadherins in metastasis," *Physical biology*, vol. 6, no. 1, pp. 016008, 2009.

[140] Wilburn E Reddick, June S Taylor, and Barry D Fletcher, "Dynamic mr imaging (demri) of microcirculation in bone sarcoma," *Journal of Magnetic Resonance Imaging*, vol. 10, no. 3, pp. 277–285, 1999.

[141] W Gregory Roberts and George E Palade, "Neovasculature induced by vascular endothelial growth factor is fenestrated," *Cancer research*, vol. 57, no. 4, pp. 765–772, 1997.

[142] Marie Rochery, H Jermyn, and Josiane Zerubia, "New higher-order active contour energies for network extraction," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*. 2005, vol. 2, pp. II–822–5, IEEE.

[143] Lior Rokach and Oded Maimon, *Clustering methods*, pp. 321–352, Springer, 2005.

[144] Tiina Roose, S Jonathan Chapman, and Philip K Maini, "Mathematical models of avascular tumor growth," *Siam Review*, vol. 49, no. 2, pp. 179–208, 2007.

[145] Reuven Y Rubinstein and Dirk P Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*, Springer, 2004.

[146] Maytal Saar-Tsechansky and Foster Provost, "Handling missing values when applying classification models," 2007.

[147] Dushyant V Sahani, Sanjeeva P Kalva, Leena M Hamberg, Peter F Hahn, Christopher G Willett, Sanjay Saini, Peter R Mueller, and Ting Yim Lee, "Assessing tumor perfusion and treatment response in rectal cancer with multisection ct: Initial observations 1," *Radiology*, vol. 234, no. 3, pp. 785–792, 2005.

[148] Yoshinobu Sato, Shin Nakajima, Nobuyuki Shiraga, Hideki Atsumi, Shigeyuki Yoshida, Thomas Koller, Guido Gerig, and Ron Kikinis, "Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," *Medical image analysis*, vol. 2, no. 2, pp. 143–168, 1998.

[149] Hidenori Shikata, Eric A Hoffman, and Milan Sonka, "Automated segmentation of pulmonary vascular tree from 3d ct images," *Medical Imaging 2004: Physiology, Function, and Structure from Medical Images*, vol. 5369, pp. 107–116, 2004.

[150] John P Sinek, Sandeep Sanga, Xiaoming Zheng, Hermann B Frieboes, Mauro Ferrari, and Vittorio Cristini, "Predicting drug pharmacokinetics and effect in vascularized tumors using computer simulation," *Journal of mathematical biology*, vol. 58, no. 4-5, pp. 485–510, 2009.

[151] Ajit Singh, Demetri Terzopoulos, and Dmitry B Goldgof, *Deformable models in medical image analysis*, IEEE Computer Society Press, 1998.

[152] S Sinha and U Sinha, "Recent advances in breast mri and mrs," *NMR in Biomedicine*, vol. 22, no. 1, pp. 3–16, 2009.

[153] Hana Skalsk and Vclav Freylich, "Web-bootstrap estimate of area under roc curve," *Aust J Stat*, vol. 35, pp. 325–330, 2006.

[154] Iryna Skrypnyk, "Finding survival groups in seer lung cancer data," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. vol. 2, pp. 545–550, IEEE.

[155] SP Sourbron and DL Buckley, "Tracer kinetic modelling in mri: estimating perfusion and capillary permeability," *Physics in medicine and biology*, vol. 57, no. 2, pp. R1, 2012.

[156] Inc StatSoft, "Electronic statistics textbook," *StatSoft, Tulsa, OK*, 2007.

[157] Ingo Steinwart and Andreas Christmann, *Support vector machines*, Springer, 2008.

[158] Catherine A Sugar and Gareth M James, "Finding the number of clusters in a dataset," *Journal of the American Statistical Association*, vol. 98, no. 463, 2003.

[159] Agus Suryanto, Kuntio Herlambang, and Pasiyan Rachmatullah, "Comparison of tumor density by ct scan based on histologic type in lung cancer patients," *Acta medica Indonesiana*, vol. 37, no. 4, pp. 195–198, 2004.

[160] KR Swanson, EC Alvord, and JD Murray, "A quantitative model for differential motility of gliomas in grey and white matter," *Cell proliferation*, vol. 33, no. 5, pp. 317–330, 2000.

[161] KR Swanson, RC Rostomily, and EC Alvord, "A mathematical modelling tool for predicting survival of individual patients following resection of

glioblastoma: a proof of principle," *British journal of cancer*, vol. 98, no. 1, pp. 113–119, 2007.

[162] Kristin R Swanson, EC Alvord, and JD Murray, "Dynamics of a model for brain tumors reveals a small window for therapeutic intervention," *Discrete and Continuous Dynamical Systems Series B*, vol. 4, no. 1, pp. 289–296, 2004.

[163] Ukihide Tateishi, Masahiko Kusumoto, Hiroshi Nishihara, Kazuo Nagashima, Toshiaki Morikawa, and Noriyuki Moriyama, "Contrastenhanced dynamic computed tomography for the evaluation of tumor angiogenesis in patients with lung carcinoma," *cancer*, vol. 95, no. 4, pp. 835–842, 2002.

[164] Bart M ter Haar Romeny, *Geometry-driven diffusion in computer vision*, Kluwer academic Norwell, MA, 1994.

[165] Michael J Thun, Lindsay M Hannan, Lucile L Adams-Campbell, Paolo Boffetta, Julie E Buring, Diane Feskanich, W Dana Flanders, Sun Ha Jee, Kota Katanoda, and Laurence N Kolonel, "Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies," *PLoS medicine*, vol. 5, no. 9, pp. e185, 2008.

[166] PS Tofsts, G Brix, DL Buckley, JL Evelhoch, E Henderson, MV Knopp, HBW Larsson, T Lee, NA Mayr, and GJM Parker, "Estimating kinetics parameters from dynamic contrast-enhanced t1-weighted mri of a diffusible tracer: standardized quantities and symbols," *J Magn Reson Imaging*, vol. 10, pp. 223–232, 1999.

[167] Paul S Tofts, "Modeling tracer kinetics in dynamic gddtpa mr imaging," *Journal of Magnetic Resonance Imaging*, vol. 7, no. 1, pp. 91–101, 1997.

[168] Paul S Tofts and Allan G Kermode, "Measurement of the bloodbrain barrier permeability and leakage space using dynamic mr imaging. 1. fundamental concepts," *Magnetic Resonance in Medicine*, vol. 17, no. 2, pp. 357–367, 1991.

[169] Ricardo Toledo, Xavier Orriols, Petia Radeva, Xavier Binefa, Jordi Vitria, Cristina Canero, and JJ Villanuev, "Eigensnakes for vessel segmentation in angiography," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. 2000, vol. 4, pp. 340–343, IEEE.

[170] Gillian M Tozer, Simon M Ameer-Beg, Jennifer Baker, Paul R Barber, Sally A Hill, Richard J Hodgkiss, Rosalind Locke, Vivien E Prise, Ian Wilson, and Borivoj Vojnovic, "Intravital imaging of tumour vascular networks using multi-photon fluorescence microscopy," *Advanced drug delivery reviews*, vol. 57, no. 1, pp. 135–152, 2005.

[171] Tero Tuominen, "Perfusion deconvolution via em algorithm," *Helsiniki University of Technology, Department of Engineering Physics and Mathematics*, 2004.

[172] James Alexander Tyrrell, Emmanuelle di Tomaso, Daniel Fuja, Ricky Tong, Kevin Kozak, Rakesh K Jain, and Badrinath Roysam, "Robust 3-d modeling of vasculature imagery using superellipsoids," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 2, pp. 223–237, 2007.

[173] Bernard Uzzan, Patrick Nicolas, Michel Cucherat, and Grard-Yves Perret, "Microvessel density as a prognostic factor in women with breast cancer a systematic review of the literature and meta-analysis," *Cancer research*, vol. 64, no. 9, pp. 2941–2955, 2004.

[174] Benjamin J Vakoc, Ryan M Lanning, James A Tyrrell, Timothy P Padera, Lisa A Bartlett, Triantafyllos Stylianopoulos, Lance L Munn, Guillermo J Tearney, Dai Fukumura, and Rakesh K Jain, "Three-dimensional microscopy of the tumor microenvironment in vivo using optical frequency domain imaging," *Nature medicine*, vol. 15, no. 10, pp. 1219–1223, 2009.

[175] Anne L Van de Ven, Behnaz Abdollahi, Carlos J Martinez, Lacey A Burey, Melissa D Landis, Jenny C Chang, Mauro Ferrari, and Hermann B Frieboes,

"Modeling of nanotherapeutics delivery based on tumor perfusion," *New journal of physics*, vol. 15, no. 5, pp. 055004, 2013.

[176] Anne L van de Ven, Pilhan Kim, O'Hara Haley, Jean R Fakhoury, Giulia Adriani, Jeffrey Schmulen, Padraig Moloney, Fazle Hussain, Mauro Ferrari, and Xuewu Liu, "Rapid tumoritropic accumulation of systemically injected plateloid particles and their biodistribution," *Journal of Controlled Release*, vol. 158, no. 1, pp. 148–155, 2012.

[177] Ingeborg MM Van Leeuwen, Carina M Edwards, Mohammad Ilyas, and Helen M Byrne, "Towards a multiscale model of colorectal cancer," *World Journal of Gastroenterology*, vol. 13, no. 9, pp. 1399, 2007.

[178] Vladimir Naumovich Vapnik and Vlamimir Vapnik, *Statistical learning theory*, vol. 2, Wiley New York, 1998.

[179] Alejandra C Ventura, Trachette L Jackson, and Sofia D Merajver, "On the role of cell signaling models in cancer research," *Cancer research*, vol. 69, no. 2, pp. 400–402, 2009.

[180] Samuel J Wang, Clifton D Fuller, Rachel Emery, and Charles R Thomas Jr, "Conditional survival in rectal cancer: a seer database analysis," *Gastrointestinal cancer research: GCR*, vol. 1, no. 3, pp. 84, 2007.

[181] Joachim Weickert, *A review of nonlinear diffusion filtering*, pp. 1–28, Springer, 1997.

[182] Joachim Weickert, *Anisotropic diffusion in image processing*, vol. 1, Teubner Stuttgart, 1998.

[183] Joachim Weickert, "Coherence-enhancing diffusion filtering," *International Journal of Computer Vision*, vol. 31, no. 2-3, pp. 111–127, 1999.

[184] Stefan Wesarg and Evelyn A Firle, "Segmentation of vessels: the corkscrew algorithm," in *Medical Imaging 2004*. 2004, pp. 1609–1620, International Society for Optics and Photonics.

[185] Onno Wink, Wiro J Niessen, and Max A Viergever, "Fast delineation and visualization of vessels in 3-d angiographic images," *Medical Imaging, IEEE Transactions on*, vol. 19, no. 4, pp. 337–346, 2000.

[186] Onno Wink, Wiro J Niessen, and Max A Viergever, "Multiscale vessel tracking," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 1, pp. 130–133, 2004.

[187] X Wu, VW Chen, J Martin, S Roffers, FD Groves, CN Correa, E Hamilton-Byrd, and A Jemal, "Comparative analysis of incidence rates subcommittee, data evaluation and publication committee, north american association of central cancer registries. subsite-specific colorectal cancer incidence rates and stage distributions among asians and pacific islanders in the united states, 1995 to 1999," *Cancer Epidemiol Biomarkers Prev*, vol. 13, no. 7, pp. 1215–1222, 2004.

[188] Yan Wu, "Propensity score analysis to compare effects of radiation and surgery on survival time of lung cancer patients from national cancer registry (seer)," Tech. Rep., School of Public Health, SUNY-Albany, 2006.

[189] Kai Xing, Dechang Chen, Donald Henson, and Li Sheng, "A clustering-based approach to predict outcome in cancer patients," in *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on.* pp. 541–546, IEEE.

[190] Peter J Yim, Peter L Choyke, and Ronald M Summers, "Gray-scale skeletonization of small vessels in magnetic resonance angiography," *Medical Imaging, IEEE Transactions on*, vol. 19, no. 6, pp. 568–576, 2000.

[191] Yongjian Yu and Scott T Acton, "Speckle reducing anisotropic diffusion," *Image Processing, IEEE Transactions on*, vol. 11, no. 11, pp. 1260–1270, 2002.

[192] X Zheng, SM Wise, and V Cristini, "Nonlinear simulation of tumor necrosis,

neo-vascularization and tissue invasion via an adaptive finite-element/level-set method," *Bulletin of mathematical biology*, vol. 67, no. 2, pp. 211–259, 2005.

[193] Jon Sporring, Luc Florack, Mads Nielsen, and Peter Johansen, *Gaussian scale-space theory*, Kluwer Academic Publishers, 1997.

[194] David E Golberg, "Genetic algorithms in search, optimization, and machine learning," *Addion wesley*, vol. 1989, 1989.

[195] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright, *Optimization for machine learning*, Mit Press, 2012.

[196] Lior Rokach and Oded Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*, pp. 321–352. Springer, 2005.

APPENDIX

2D	Two Dimensional
3D	Three Dimensional
CAD	Computer Aided Diagnostic
CI	Confidence Interval
CT	Computed Tomography
DCE-MRI	Dynamic Contrast Enhanced Magnetic Resonance Imaging
EM	Expectation Maximization
FCM	Fuzzy C-Mean Clustering
IVM	Intravital Microscopy
KNN	K Nearest Neighbor Classifier
MRA	Magnetic Resonance Angiogram
MT	Magnetic Resonance
NCI	National Cancer Institute
NIH	National Institutes of Health
NSCLC	Non Small Cell Lung Cancer
ORG	Ordered Region Growing
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PK	Pharmacokinetic
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SCLC	Small Cell Lung Cancer
SEER	Surveillance Epidemiology End Results
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TOF	Time of Flight
VEGF	Vascular Endothelial Growth Factor

CURRICULUM VITAE

Behnaz Abdollahi: PhD 2015

email: behnaz.abdolahi@gmail.com

Education:

PhD - Electrical and Computer Engineering - 2015

University of Louisville

MS - Computer Engineering - 2006

Sharif University of Technology

BS - Azad University - 2000

Azad University Tehran North Branch

Professional Experience:

Software Developer and Machine Learning, 2015, Genscape

Research Assistant, 2009-2015, University of Louisville

Research Assistant, 2003-2006, Sharif University of Technology

Research Assistant, 2001-2003, Telecommunication Research Center

Software Developer, 2006, Information Center of Tehran University

Software Developer, 2006, Private Programming Company in Tehran

Project Analyst, 2005-2006, A Company funding private IT projects

Software Developer, 2000-2001, Tehran Traffic Control Company

Teaching Experience:

Lecturer, 2007-2008, Tehran Payam Noor University

Lecturer, 2004-2005, Tehran Jame Elmi Karbordi University

Skill:

Programming Languages: Python, Java, R, Matlab, Octave, C++

Tools: Scikit-Learn, Weka, ImageJ, Latex

Mathematical Knowledge: Machine Learning,Linear Algebra,Statistics

Scholarship/Fellowship and Awards:

Spring Semester Tuition, 2014, Awarded from University of Louisville

Adobe Grace Hopper Scholarship, 2013, Awarded to attend GHC

BPDM Scholarship, 2013, Awarded by BPDM to attend KDD

Grosscurth Fellowship, 2009-2011, Awarded from University of Louisville

Student Government Association, 2012 , Awarded from University of Louisville

Women in Machine Learning, 2009, Awarded by WiML to present my poster

Anita Borg Scholarship, 2010 Awarded to attend Grad Cohort workshop

Iran Telecommunication Research Center, 2006 Awarded for top MS thesis

Publication:

Peer Reviewed Articles and Proceedings

Abdollahi, Behnaz, and Hermann Frieboes, "Analysis of lung cancer staging systems through a clustering-based patients survival time comparison.", Under Preparation

Van de Ven, Anne L*, Behnaz, Abdollahi*, Carlos J. Martinez, Lacey A. Burey, Mellisa D. Landis, Jenny C. Chang, Mauro Ferrari, and Hermann Frieboes,

"Modelling of nanotherapeutics delivery based on tumor perfusion."
New Journal of Physics 15, no. 5 (2013): 055004. *joint first authorship

Khalifa Fahmi, Mohamad Abou El-Ghar, Behnaz Abdollahi, Hermann Frieboes, Tarek El-Diasty and Ayman El-Baz, "A comprehensive non-invasive framework for automated evaluation of acute renal transplant rejection using DCE-MRI" NMR in Biomedicine 26, no. 11 (2013): 1460-1470

Behnaz Abdollahi, Ayman El-Baz, Amir A. Amini, "A mulit-scale non-linear vessel enhancement technique." Engineering in Medicine and Biology Society, 2011 Annual International Conference of IEEE, pp. 3952-3929. 2011

Book Chapter

Behnaz Abdollahi, Neal Dunlap, Hermann Frieboes, "Bridging the gap between modelling of tumor growth and clinical imaging." Abdomen and Thoractic Imaging, p. 463-487. Springer US, 2014

Selected Presentations and Conference Abstracts:

Van de Ven, Anne L*, Behnaz, Abdollahi*, Carlos J. Martinez, Lacey A. Burey, Mellisa D. Landis, Jenny C. Chang, Mauro Ferrari, and Hermann Frieboes, "Modelling of nanotherapeutics delivery based on tumor perfusion." J.B. Speed School of Engineering E-Expo, University of Louisville, 2013, joint first authorship

Van de Ven, Anne L*, Behnaz, Abdollahi*, Carlos J. Martinez,Mauro Ferrari, and Hermann Frieboes, "Analysis of tumor perfusion using on engineering approach" James Graham Brown Cancer Center Annual Retreat, University of Louisville, 2012 joint first authorship

Abdollahi, Behnaz, Ayman El-Baz, Amir A. Amini, "A multi-scale non-linear vessel enhancement technique.", Grace Hopper Celebration of Women in Computing, Portland, OR, 2011