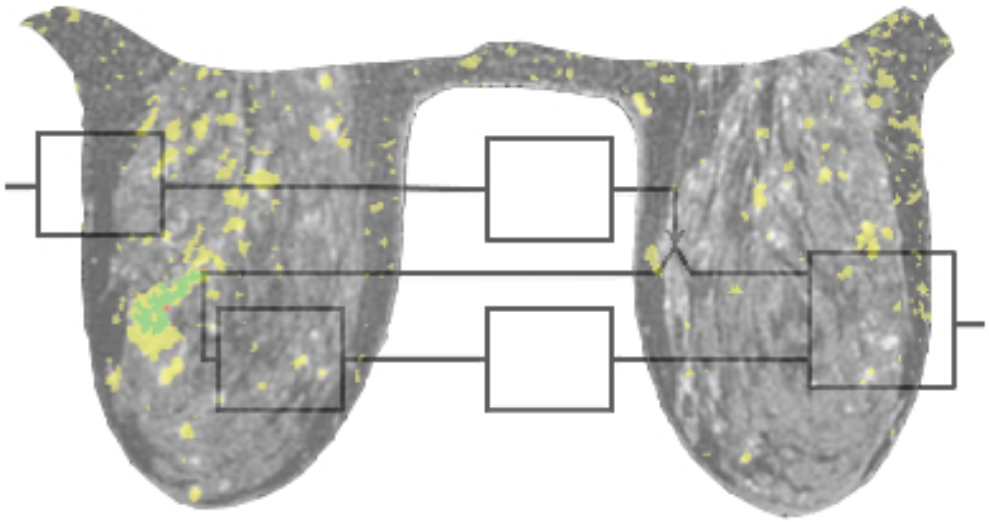
CHALMERS



Automatic Detection of 3D Breast Lesions in Dynamic Contrast-Enhanced MR Images

Master of Science Thesis in Communication Engineering

JONATHAN ARVIDSSON

Master of Science Thesis in Biomedical Engineering

FREDRIK JOHANSSON

Department of Signals and Systems
Division of Biomedical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2011
EX097/2011

MedTech West

MASTER OF SCIENCE THESIS 2011

Automatic Detection of 3D Breast Lesions in Dynamic Contrast-Enhanced MR Images

Master of Science Thesis in Communication Engineering
JONATHAN ARVIDSSON

Master of Science Thesis in Biomedical Engineering
FREDRIK JOHANSSON

Thesis Supervisor and Examiner:
Assistant Prof. Andrew Mehnert
Image Analysis
Chalmers University of Technology
Department of Signals and Systems
SE-412 96 Goteborg Sweden

Department of Signals and Systems
Division of Biomedical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2011

Automatic Detection of 3D Breast Lesions in Dynamic Contrast-Enhanced MR Images

JONATHAN ARVIDSSON
FREDRIK JOHANSSON

Master of Science Thesis 2011
Department of Signals and Systems
Division of Biomedical Engineering

Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: + 46 (0)31-772 1000

Cover:
Suspiciously enhancing breast tissue in an MR image slice superimposed on a flow chart representing the proposed algorithm.

Göteborg, Sweden 2011

Automatic Detection of 3D Breast Lesions in Dynamic Contrast-Enhanced MR Images

Master of Science Thesis in Communication Engineering
JONATHAN ARVIDSSON
Master of Science Thesis in Biomedical Engineering
FREDRIK JOHANSSON
Department of Signals and Systems
Division of Biomedical Engineering
Chalmers University of Technology

Abstract

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is being increasingly used in the clinic as an adjunct to x-ray mammography and ultrasound for the detection and characterisation of breast cancer. The technique involves acquiring T1-weighted volume images of one or both breasts before and at several time points after the injection of a contrast agent. Interpretation of this 4D data is a complex task for the radiologist and is becoming more so with developments of higher field-strength MRI scanners and associated increases in spatial, temporal and contrast information. Several commercial computer-assisted detection/diagnosis (CAD) systems for breast MRI have been developed in recent years to help radiologists with this task. However, these systems at present fall short of automatically locating and classifying malignant lesions. It is not surprising, therefore, that the efficacy of breast MRI CAD remains an open question. A recent review concluded that breast MRI CAD needs to be based on "quantitative features extracted preferably from the automatically segmented 3D lesion".

This thesis deals specifically with the problem of automatically segmenting (i.e. delineating) 3D lesions in breast DCE-MRI data. In particular it reviews existing approaches and proffers a novel approach based on voxel-wise classification. The new approach involves assigning a "suspiciousness" score to each voxel using features extracted from its time series, and then computing the spatial co-occurrence of this score in a neighbourhood including the voxel. The thesis also presents an empirical evaluation of the efficacy of this technique versus a competing method based on multi-spectral co-occurrence. The evaluation was performed on real clinical breast MRI data for 32 subjects. The results demonstrate that the proposed method achieves a comparable level of performance (AUC of 0.8989 ± 0.0021 versus AUC of 0.9330 ± 0.0018). However the advantage of the proposed method over the competing method is that it does not require subjective specification of feature ranges for computing co-occurrence.

Keywords: Breast Magnetic Resonance Imaging, Automatic segmentation, Computer Aided Detection, Mammography, 3-D grey level co-occurrence, Lesion detection, Pattern recognition, Dynamic Contrast Enhanced MRI, GLCM

Contents

Acknowledgements

We wish to acknowledge and thank those who contributed to this thesis:

Andrew Mehnert for his endless support and valuable guidance throughout the project.

Stefan Candefjord for valuable input and discussions.

Darryl McClymont for his contributions on breast-air boundary segmentation.

Dominic Kennedy and Queensland X-Ray for providing MR data for us to use.

Qaiser Mahmoud for his endless optimism and encouragement.

MedTech West for providing an inspiring environment for us during our stay.

Göteborg September 2011
Jonathan Arvidsson, Fredrik Johansson

Acronyms

3-D/3D	Three Dimensional
ACR	American College of Radiology
ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
BCa	Breast Cancer
BIC	Bayesian Information Criterion
BIRADS	Breast Imaging-Reporting and Data System
BMRI	Breast Magnetic Resonance Imaging
CAD	Computer Aided Detection
CCR	Correct Classification Rate
DCE	Dynamic Contrast Enhanced
FN	False Negative
FP	False Positive
GLCM	Grey-Level Co-occurrence
k-NN	k-Nearest Neighbour
LR	Logistic Regression
LS	Least Squares
MRI	Magnetic Resonance Imaging
MSC	Multispectral Co-occurrence
NLS	Non-linear Least Squares
NMR	Nuclear Magnetic Resonance
Pixel	Picture Element
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SFS	Sequential Forward Selection
SLR	Stepwise Logistic Regression
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TSC	Temporal Score Co-occurrence
VOI	Volume of Interest
Voxel	Volumetric Picture Element

1 Introduction

The research described herein was undertaken in the MedTech West[1] centre located at Sahlgrenska University Hospital in Gothenburg, Sweden between May and October 2011. It constitutes part of a larger research project seeking to develop image analysis tools to assist radiologists with the task of interpreting magnetic resonance (MR) images of the breast and prostate. Such images are used in the detection and characterisation of both breast cancer and prostate cancer. In particular they are used for staging the cancer (i.e. staging the degree of progression), determining the most appropriate treatment and for follow-up after cancer treatment.

The need for objectivity, together with the desire to simplify for the radiologist the increasingly complex task of interpreting MR data, has spawned research and development of computer-assisted detection/diagnosis (CAD) systems for breast MRI including several commercial systems such as CADstream®[2] and DynaCAD®[3]. These systems at present fall short of automatically locating and classifying malignant lesions. Instead they automate many of the image processing and analysis functions that would otherwise have to be performed manually and visualise the data to aid interpretation. It is perhaps not surprising, therefore, that a meta-study published in March of this year [1] concluded that "commercial CAD systems for breast MRI do not improve the accuracy of experienced radiologists" and so their interpretation remains essential. This is in stark contrast to CAD for x-ray mammography which has been shown to increase the number of detected cancers by approximately 10%. This corresponds to the accuracy obtained when two radiologists analyse the DCE-MRI data [2]. A 2009 review of breast MRI CAD concluded that such systems need to be based on quantitative features preferably extracted from the automatically segmented three dimensional (3-D) lesion [3]. For these reasons the research presented in this thesis specifically concerns the development and evaluation of algorithms and software for automatically segmenting (i.e. detecting and delineating) tissue in breast MR images that is suspicious for malignancy.

1.1 Background

The most frequently diagnosed cancer in women today, in both the world's developed and developing regions, is breast cancer. It is also the cause of most cancer-related deaths in women, closely followed by lung cancer [4].

The key strategy for improving survival rates is early detection of the disease. For this reason, screening programs have been introduced worldwide based on x-ray mammography. Nevertheless the technique has some well known limitations including low specificity (i.e. certainty that what was detected is cancer), problems with the detection of lesions in dense breast tissue as well as the fact that the technique yields two-dimensional projections of inherently 3D tissue [5]. This has, in part, prompted the exploration of alternative imaging modalities such as magnetic resonance imaging (MRI), sonography (ultrasound) and nuclear medicine (PET and SPECT) imaging [6]. These modalities are typically used in patients with known or suspected breast cancer [7] or in the screening of high risk patients [8]. Of these modalities MRI shows the most promise for improved screening of high risk women [9].

Modern breast MR imaging is based on dynamic contrast-enhanced (DCE) MRI. The

[1] MedTech West partners are Chalmers University of Technology, Gothenburg University, the University of Borås, Sahlgrenska University Hospital and the Region of Västra Götaland.

[2] CADstream website: http://www.gehealthcare.com/euen/mri/products/applications/breast/cadstream.html

[3] DynaCAD website: http://www.invivocorp.com/avs/dynacad.php

technique involves the acquisition of three-dimensional T1-weighted images of one or both breasts before and several times after the injection of a gadolinium-containing contrast agent. A typical clinical DCE-MRI scan comprises a precontrast volume followed by four to six contrast-enhanced volumes. Each volume, containing a large number of 2D slice images, is acquired in intervals about 60 to 120 seconds [10].

The shape of the signal-intensity time curve is an important criterion for discriminating benign and malignant lesions. This curve typically demonstrates early rapid uptake after contrast injection for cancers. This yields high sensitivity (i.e. ability to detect cancer) as reported by Kuhl et al. [11]. Moreover the nature of the post-initial enhancement can be useful in determining whether a tumour is malignant or benign: the curve may plateau, decline, or continue to increase more slowly with a delayed washout [11]. However given that many benign lesions demonstrate similar enhancement to malignant lesions, and that some malignant cancers are characterised by shallow or absent enhancement, the specificity of the technique is at best moderate [10].

In addition to the time intensity characteristics of a lesion, its morphology is an important criterion for classification. For example malignant lesions often exhibit spiculated margins [12]. The Breast Imaging-Reporting and Data System (BIRADS), published by the American Collage of Radiology (ACR) [12], is a lexicon devised to help radiologists report breast MRI findings in a consistent and standardised way.

The advantages of DCE-MRI as compared to x-ray mammography are among others its high sensitivity, the possibility to obtain both morphological and functional features related to tumour malignancy for analysis, the non-ionising[4] character of the modality and its ability to provide high resolution images [13]. Volume estimates of tumours have also been shown to be more accurate when performed with MRI than when performed with x-ray mammography or sonography [14, 15].

Integrating and evaluating all the information from the multi-temporal image sequence provided by a DCE-MRI scan is a labour intensive task for radiologists [16]. It is also subjective. For example random variations up to 29% of the tumour size between manual delineation and true volume have been reported in inter- and intraobserver analyses [14]. Automatic segmentation has been shown to be far less time consuming and can provide objective and reproducible results because of its operator-independency [13]. Computerised techniques may thus be a way to improve the objectivity, consistency as well as the efficiency of breast lesion segmentation in line with what has been achieved for x-ray mammography [17, 18]. Nevertheless DCE-MRI presents a number of challenges for automatic segmentation including the variation in both temporal and spatial contrast agent distributions in suspicious tissue for a single patient, as well as between patients [19], and that the observed MR signal cannot be easily calibrated (in contrast to CT imaging).

A few automatic methods for segmenting lesions in DCE-MRI data have been proposed. These are reviewed in Chapter 3 wherein it is concluded that textural analysis by co-occurrence in a voxel-wise fashion, as a basis for automatic segmentation of lesions, is one of the more promising approaches to the problem of characterising spatial and temporal intensity variation.

[4]Ionising radiation is composed of particles that individually have high enough energy to remove an electron from an atom or molecule. Exposure may cause cell damage, depending on dose. As opposed to modalities such as PET, x-ray mammography and nuclear medicine, MRI does not utilise ionising radiation.

1.2 Hypothesis

The hypothesis underlying this research is that it is possible to automatically and accurately segment 3D lesions in DCE-MRI breast data by means of voxel-wise classification based on quantitative features that describe both spatial and temporal change in contrast enhancement.

1.3 Aim and Objectives

The aim of this research was to explore the validity of this hypothesis by:

1. Developing a spatio-temporal segmentation method for 3D breast lesions based on characterising the spatial co-occurrence of enhancement; and

2. Evaluating the performance of the new method using real clinical breast MRI data.

1.4 Scope

The focus of this research was only on the development of 3D lesion segmentation in breast DCE-MRI data. A complete CAD system would additionally include steps for extracting quantitative features for detected lesions and automatic classification of these lesions; e.g. into benign and malignant. These steps are not considered herein.

2 Background and theory

This chapter presents the necessary background and theory needed for the remainder of the thesis. The first four sections cover breast cancer, magnetic resonance imaging (MRI), dynamic contrast-enhanced MRI, and the analysis of contrast enhancement in DCE-MRI. The remaining sections cover aspects of statistical pattern recognition, types of classifiers, and texture analysis by co-occurrence.

Throughout the thesis scalars are denoted by italics (x) and vectors are denoted by bold face ($\mathbf{x}$). Sets are designated by capital letters (X). Matrices and higher dimensional arrays are denoted by bold face and capital letters ($\mathbf{X}$). A bar ($\bar{x}$) denotes an average and estimates are denoted by a hat ($\hat{x}$). Spaces are written in calligraphy ($\mathcal{X}$). This thesis concerns several disciplines of engineering, thus variable notations may occur several times in different contexts.

2.1 Breast cancer

Cancer originating from breast tissue is termed breast cancer. The most common breast cancers originate from either the milk ducts or the lobules that supply them with milk. The former are termed ductal carcinomas and the latter lobular carcinomas. If the tumour has not spread from the originating tissue it is termed *in-situ*. If it has it is termed *invasive*. More commonly breast lesions are benign in nature. Benign lesions include fibroadenoma, composed of epithelial and stromal tissue, and intraductal papillomas which may occur anywhere within the breast ductal system [20]. Examples of malignant lesions are ductal carcinoma in situ, which still has not spread to surrounding tissue and invasive ductal carcinomas that have infiltrated through the wall of the ducts [21].

The majority of breast cancers present symptomatically, although screening programs have led to an increased proportion of asymptomatic detection. Evaluation of breast abnormalities should be performed by triple assessment including: clinical examination, imaging, and tissue sampling [20]. When evaluating images, different characteristics are visually inspected in order to differentiate between malignancy and benignity [20].

2.2 Breast MRI

The hydrogen nucleus, or proton, is used in MR imaging because of its high concentration in biological tissues such as fat and water. Proton T1 and T2, the longitudinal and transversal relaxation times differ between various tissue types and can also be altered by disease [22]. Studies in the early 1980s sought to differentiate benign and malignant breast lesions based on the intrinsic contrast given from T1 and T2 relaxation times. However these studies were unable to demonstrate reliable differentiation [22]. As a consequence MR imaging of the breast was at first not widely accepted. Modern MRI of the breast stems from several developments in the mid-1980s including the development of fast gradient-echo imaging sequences with small flip angles, the introduction of gadolinium based contrast agents and the introduction of dedicated surface coils [22].

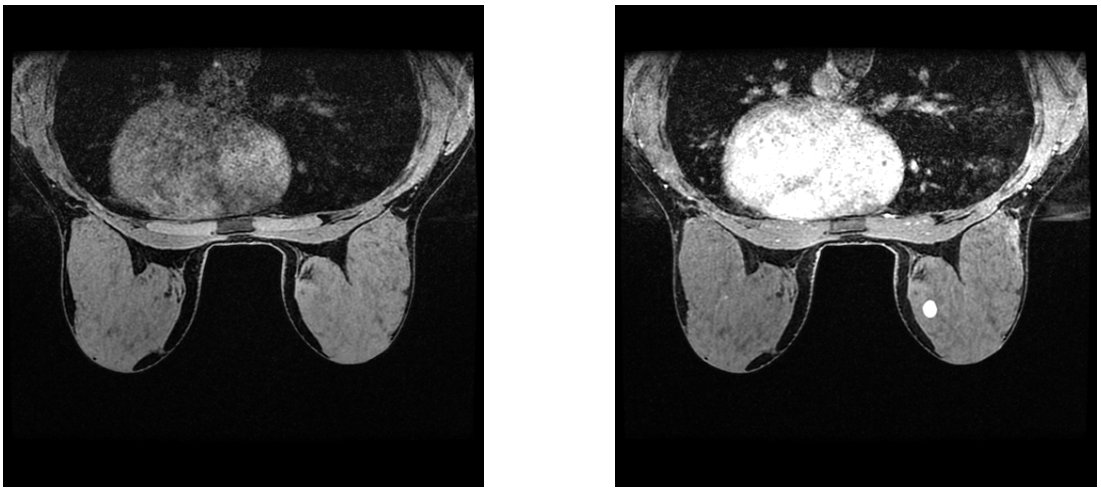
2.3 Dynamic contrast-enhanced magnetic imaging of the breast

In early studies in the field of breast DCE-MRI, images were acquired at least five minutes apart and with a slice thickness of 5 mm. It had previously been shown that breast carcinomas showed significant enhancement in the early phases after contrast agent administration [10]. Subsequent studies analysed also the intermediate and late phases of

the enhancement pattern and showed that important information for lesion discrimination is contained here [11]. As faster pulse sequences were developed, the possibility to image the whole breast at shorter time intervals became feasible [10]. These techniques allowed characterisation of lesion enhancement over shorter periods of time. By use of this data, a signal intensity curve (or time intensity curve) can be generated and its properties analysed in order to extract valuable information for distinguishing between malignant and benign tissue. As experimental observations have shown that the contrast agent concentration is proportional to the relative signal increase, the signal increase relative to the image volume before injection of contrast agent is calculated [23]

$$\mathbf{C}(t) = \frac{\mathbf{I}_t - \mathbf{I}_0}{\mathbf{I}_0}, \tag{2.1}$$

where $\mathbf{I}_t$ represents the post-contrast volumes and $\mathbf{I}_0$ the pre-contrast volume. To acquire what is called a subtraction image, the division with pre-contrast volume is omitted.



(a) Pre-contrast image.

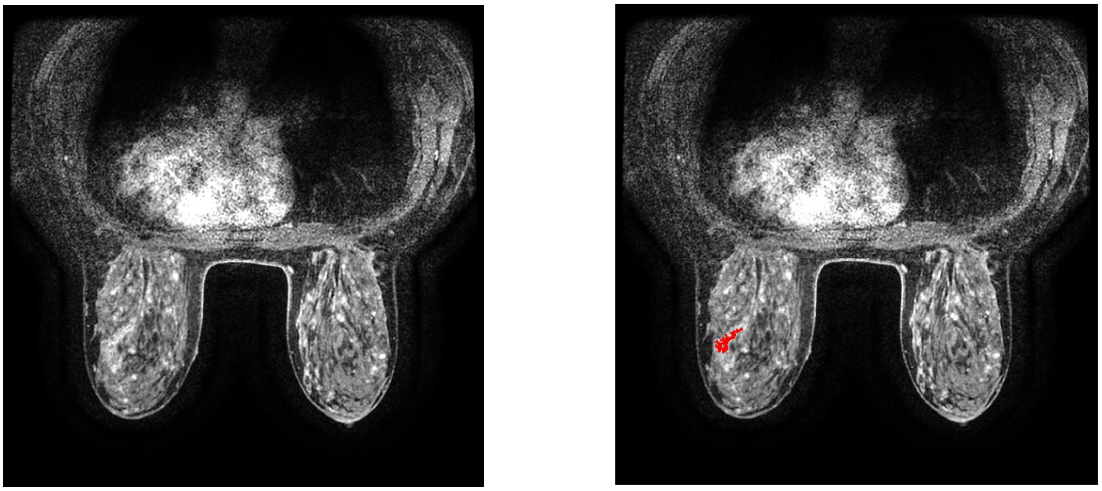
(b) Post-contrast image.

Figure 2.1: Enhancement of a benign lesion in the right breast after injection of a contrast agent. Note also the considerable enhancement within the chest cavity.

2.3.1 Pathophysiological basis of contrast enhancement and lesion characteristics

Lesion morphology, enhancement intensity and kinetics differ between benign and malignant lesions. For instance round or oval shapes of masses, see Figure 2.1, are highly indicative of benignity [10] whilst indistinct or ill-defined margins raises concern about a lesion infiltrating surrounding tissue [12] (see Figure 2.2). Qualitative analyses have concluded that the signal intensity in malignant tissue generally peaks early compared with normal tissue, in part because of increased vascularisation and tumour vessel leakiness [24]. Nevertheless, an overlap exists between the enhancement curves for many benign lesions and malignant lesions, making the analysis more difficult [10].

The importance of tumour angiogenesis (growth of new blood vessels) for tumour growth is well known. To experimentally describe micro-vessel structure and function, MRI is used [25]. The results from Gibbs et al. [26] showed that there are significant differences in texture between benign and malignant tissue. From the field of computer image analysis, various textural algorithms have been proposed for quantifying these properties.



(a) Post-contrast image.

(b) Post-contrast image with ROI superimposed.

Figure 2.2: A malignant lesion with irregular shape and non-enhancing regions.

2.4 Enhancement curve analysis

Non-model-based features such as *signal-enhancement ratio* or *initial percentage enhancement* can be extracted directly from raw data. Model-based features require parametric models to which the raw data is fitted. Several parametric models for describing contrast enhancement in breast MRI data have been proposed in literature. These models can be divided into subgroups of pharmacokinetic models and empiric parametric models.

2.4.1 Pharmacokinetic models

Models based on pharmacokinetics aim to describe the time-varying distribution of contrast agent in different tissue compartments in the body[27]. It is therefore necessary to relate the observed signal intensity increase to the contrast agent concentration. Previous studies show experimentally that this relation is linear in blood [28] and soft tissues [29]. The standard two compartmental model used in DCE-MRI has the form

$$C(t) = K^{trans} \int_0^t C_p(t') exp\left(\frac{-K^{trans}(t-t')}{v_e}\right) dt', \tag{2.2}$$

where C is the contrast agent concentration in the tissue as a whole, C_p is the concentration of contrast agent in the blood plasma, v_p is the fraction of the tissue volume occupied by blood plasma, v_e is the fraction occupied by the extracellular space extravascular space, and K^{trans} is the volume transfer constant relating v_p to v_e [27]. The basis for this model is illustrated in Figure 2.3.

Pharmacokinetic models require an accurate pre-measurement of the arterial input function along with an accurate quantisation of the signal intensity. Given that typical dynamic breast MRI data consists of less than ten time points and are often influenced by noise, fitting such a complex model to this data may be ambitious [19].

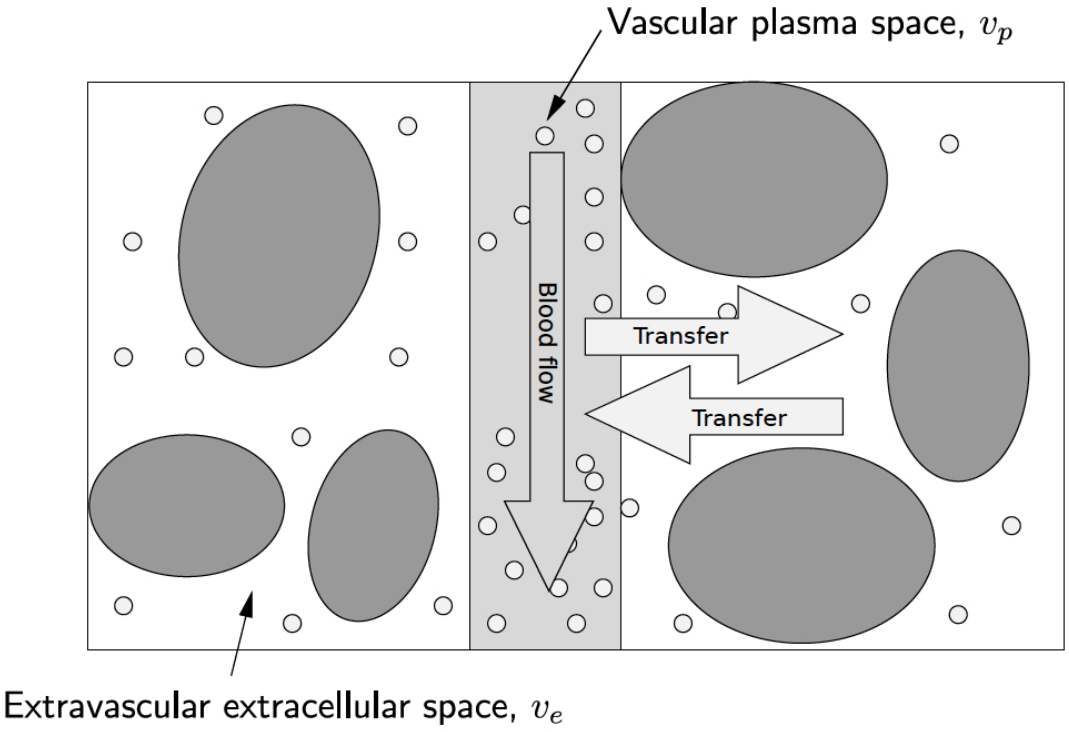


Figure 2.3: Illustration of two compartmental model. [27]

2.4.2 Empirical parametric models

Describing the contrast enhancement pattern based on empiric parametric models does not require any pre-measurements of physical properties. These models simply describe the shape of the enhancement curve.

As shown in Mehnert et al. [27] simple empiric parametric models such as the Linear Slope model and the Ricker model can be fitted to the data points using linear least squares. This means that fitting these models to the data can be done very fast and thereto without the need of specifying starting values for parameters, performing pre-measurements of physical properties of tissue or dealing with convergence issues. The results of Mehnert et al. [27] also show that the Linear Slope shows a higher goodness of fit to real clinical data than the more sophisticated pharmacokinetically inspired model by Hayton. The Ricker model showed a remarkably good fit to the data given that it has two rather than three parameters.

2.5 Statistical pattern recognition

Statistical pattern recognition is a broad term that comprises several concepts ranging from data collection and discrimination to interpretation of results. The field originated from multiple disciplines, including: statistics, engineering, computer science and psychology. The term pattern is refers to the entity or object of interest [30], e.g. an MR image, a human face or a speech signal. The recognition problem is often posed as a categorisation task, where a pattern is assigned a predefined class (supervised classification) or learned based on similarities of patterns (unsupervised classification) [31]. Measurements, more commonly termed features, are used to characterise each pattern, including e.g. quantitative measurements of object descriptors and derived numerical parameters [30].

2.5.1 The curse of dimensionality

Typically when adding features to a statistically based classifier the additional features should improve the classification performance due to the increased amount of information they hold. However a consequence of the increased dimensionality is that the observations in each class become less representative of that class. In other words, if the number of samples stay constant as new features are added the feature space will grow more and more sparse. In order to form a representative sample in this higher dimensional space more observations are needed. As more features are added, the classification performance on the training set will increase but the performance on unseen data will decrease [30]. This phenomena is referred to as the "curse of dimensionality". A rule of thumb for avoiding the curse of dimensionality is to use at least ten times as many training instances per class as the number of features.

2.5.2 Feature selection strategies

Feature selection strategies address the following problem: "given a set of d features, select the subset of size m that leads to the smallest classification error" [31]. A set that contains d features will have $2^d - 1$ possible subset candidates, without considering the case of choosing zero features [30]. In general the exponential increase in the number of candidate subsets as d increases makes an exhaustive search unfeasible, except in cases where only a small total number of features is considered. Many methods for feature selection have been presented in the literature. For an overview, see [31]. The methods used in this thesis are discussed below.

A feature selection algorithm typically consists of four steps. A choice of initial feature set, a criterion for evaluating the discriminatory power of a feature subset, a strategy for adding or removing features from the subset or from the feature set and finally a stopping criterion to decide when to stop the algorithm [31]. The criterion used to evaluate the performance of the feature set can either be formulated in a statistical sense, by adding the feature that is most statistically significant. Or it can be done in a heuristic sense by adding the feature that together with preceding features gives the best classification result on a training data set. Naturally each feature selection method therefore has its own properties, which depending on the nature of the features and the particular pattern recognition problem may result in a different final set of features and ultimately different classification performances.

The sequential forward selection strategy is one of three basic traditional approaches for feature selection, the other two being sequential backward selection and stepwise selection. The strategy of sequential forward selection is to start out without any features, then adding the best single feature according to a certain criterion function. Further features are then added, one at a time. The feature that together with the previously chosen feature or subset of features that performs best according to the criterion function, is the one that is added to the feature subset [31]. Sequential backward selection is performed in a similar fashion, but starts out with all features and successively deletes one feature at a time. Stepwise selection methods starts with a set of features, either the entire set or a subset. In each iteration of the algorithm a feature is either added or removed from the feature subset. More sophisticated methods such as the "Plus l - take away r" strategy have been found to outperform the more simple straight sequential searches on large feature selection problems [31]. The forward search, however, is computationally attractive and does not require a selection of l and r [31].

2.5.3 Cross-Validation

Rotation estimation or k-fold cross-validation, is an error estimation technique used to asses how well the results from a statistical analysis generalise to unseen data [31]. A dataset D, containing vectors of observations, is randomly split into k number of subsets (folds) $D_1, ..., D_k$ of approximately equal size. For each fold $n \in \{1, 2, ..., k\}$ the classifier is trained on $D \setminus D_n$ and tested on D_n. The estimate of error or accuracy rate can be evaluated in terms of the number of incorrect/correct classifications for each subset D_n, divided by the total number of instances in that same set [32], or the AUC (see 2.5.5). Stratified cross-validation is when the folds contain roughly the same proportion of labels as the full training set D.

Other error estimation techniques include the leave-one-out method where the classifier is trained n times on $(n-1)$ samples and evaluated on the remaining sample, where the training set and test sample is rotated over the full set. This can be viewed as an extreme case of k-fold cross-validation. Another technique is the holdout method, where, for instance, half the data is used for training and the remaining data as test set [31]. The k-fold cross validation approach has lower bias than the holdout method and is computationally less expensive than the leave-one-out method [31].

2.5.4 Model selection

Model selection generalises the concept of feature selection to include the selection of the type of classifier as well. Ideally these steps should not be considered independently of each other [30]. The idea behind these concepts is to make full use of the available training data [31], without overfitting to it and thus reducing performance on unseen data.

2.5.5 Evaluating classifier performance

Several ways to quantify the performance of a classifier exist. The receiver operating characteristic (ROC) curve is a plot of the probability of detecting a false positive (1-specificity) against the probability of detecting a true positive (sensitivity) over the range of all possible classifier threshold values. The ROC curve always passes through the two points (0,0) and (1,1) and the straight line through these points is the ROC of a classifier that does no better than random (see Figure 2.4).

The Area Under the Curve (AUC) refers to the area under the ROC curve. The AUC represents a summary measure of sensitivity and specificity over all possible classification thresholds [33]. With the shape of the ROC curves in mind, an AUC value above 0.5 means a performance better than that of a random classifier. A higher AUC value indicates better performance, where a perfect classifier has an AUC value of 1. A method used for computing the standard error of the AUC is to compute the standard error of the Wilcoxon statistic as follows [34]:

$$\hat{SE} = \sqrt{\frac{\theta(1-\theta) + (n_p - 1)(Q_1 - \theta^2) + (n_n - 1)(Q_2 - \theta^2)}{n_p n_n}}, \tag{2.3}$$

where $\theta = \hat{AUC}$ (the estimate of the AUC), $Q_1 = (2 - \theta)$, $Q_2 = (2\theta^2)/(1 + \theta)$, n_p is the number of positive examples, and n_n is the number of negative examples.

2.6 Classification approaches

Given a classification problem, several different approaches can be taken where classifiers using different design principles are constructed. Three different approaches are identified

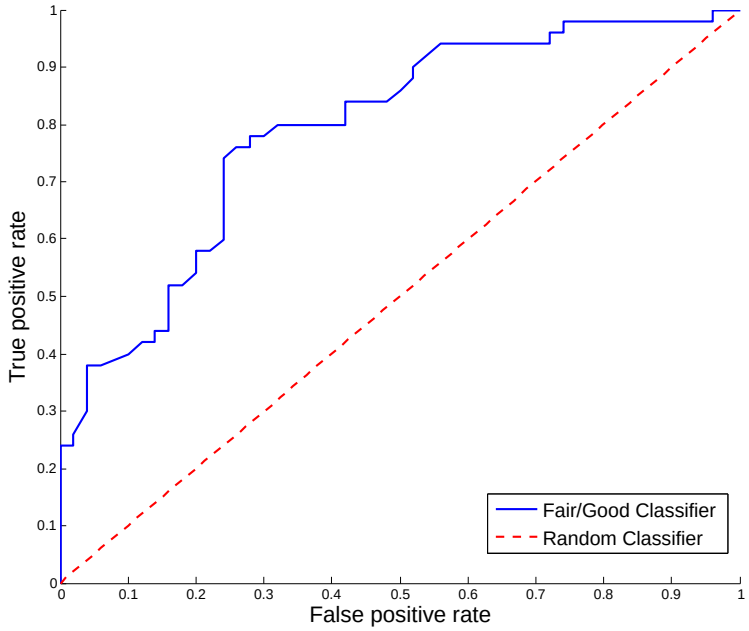


Figure 2.4: A comparison between a classifier with fair performance and a random one.

by Jain et al. [31]; concepts based on measures of similarity, probabilistic approaches and classifiers that construct decision boundaries. There is no context- or usage-independent reasons to choose one classification method over another [35]. Here, the three different classification methods used in this thesis are presented; k-nearest neighbour classification, logistic regression and support vector machines.

2.6.1 k-Nearest Neighbours classifier

The k-nearest neighbour (k-NN) rule is an extension of the nearest neighbour rule for a set of n pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where the $\mathbf{x}_i$'s take values in a metric space $\mathcal{X}$ (e.g. Euclidean) and y_i's take values from a set of labels $n \in \{1, 2, \ldots, M\}$. Given a new instance $(\mathbf{x}, y)$, where $\mathbf{x}$ is observable, it is desired to predict y from the set of known pairs by assigning, according to a metric (e.g Euclidean, Mahalanobis or Chebychev distances), the label of the closest neighbour in that space [36].

Extending this rule to the k-nearest neighbours of $\mathbf{x}$, this observation is assigned to label y_i if a majority of its nearest neighbours are members of this class [37]. Figure 2.5 illustrates how the observed instance will be assigned a different label, depending on how many neighbours are taken into consideration.

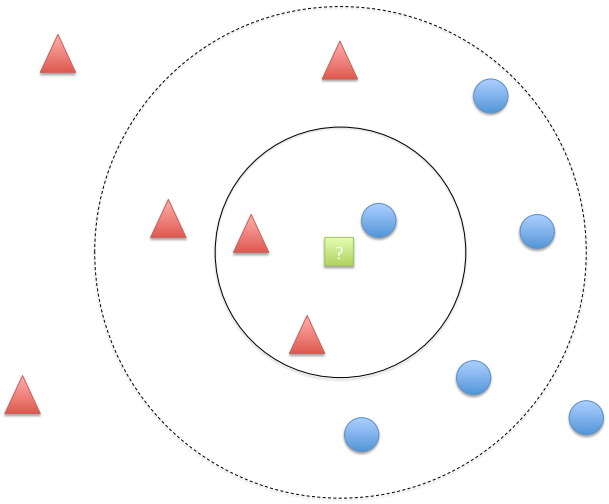


Figure 2.5: Distribution of features in a 2-dimensional k-NN classification scenario. The label of the unknown instance will be different, depending on how many neighbours that are considered. The square in the centre will be classified as a triangle if $k = 3$, but as a circle if $k = 9$ (assuming that Euclidian distance is used as the metric).

2.6.2 Linear and Logistic Regression classifiers

In linear regression, a linear model is used in order to identify the relationship between the predictors and the outcome. Given a set of known data $\mathbf{x}$, the model can be expressed as [38]

$$z = \alpha_0 + \mathbf{x} \cdot \boldsymbol{\alpha} + \epsilon, \tag{2.4}$$

giving a model for estimating variables in a continuous range. Here, $\boldsymbol{\alpha}$ is a vector containing what are called the regression coefficients, which influence how each predictor affects the outcome. These coefficients are estimates using the least squares method [30]. The error term, ϵ, captures influence on z other than that from the predictors.

The linear concept can be extended to that of logistic regression. Elaborating on the details a bit further, this is a parameter-free discriminative classification method that, given a set of instance-label pairs $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y \in \{0, 1\}$, builds a model approximating the posterior distribution $P(y|\mathbf{x})$. Logistic regression can be expressed on a functional form $P(y|\mathbf{x}) = f(z)$, with z given from (2.4).

As in the case with linear regression, the vector $\boldsymbol{\alpha}$ holds the regression coefficients of $\mathbf{x}$, although usually determined by maximum-likelihood estimation based on the dataset [33, 30] and α_0 is where all independent variables are zero. Each of these coefficients can be considered as weights on how much each of the variables, or features, are allowed to influence the outcome. The model calculates the probability of a certain class membership by use of the logistic function

$$f(z) = \frac{1}{1 + e^{-z}} = P(1|z), \tag{2.5}$$

and $P(0|z) = 1 - P(1|z)$. The model complexity is low, consequently overfitting is less likely to occur than for more flexible models [33]. Relating back to 2.5.2, these methods can be used in a stepwise manner for feature selection. In regression based feature selection methods the feature that produces the greatest statistically significant (for a certain confidence interval) change in a criterion function relative to a model not containing the

feature is selected [39]. Or, expressed in other terms; investigating the influence on a criterion, by adding or removing the regression coefficients, a heuristically deduced model is found. Normally the feature subset tends to grow larger and larger, because more information is included for each feature added. However, it is possible that a smaller subset of features can produce a sufficiently accurate classifier. By penalising the fit by a measure of complexity of the model, for example the number of features in the subset, the smallest adequate model may be chosen [40].

When predicting probabilities or outcomes in the range zero to one [39], logistic regression models are commonly used as they map values from minus to plus infinity to this interval. The linear and logistic models are used in a stepwise fashion with the same motivation; to find the estimators that agree most closely with the observed data, $\mathbf{x}$.

2.6.3 Support Vector Machines

In the case of a two-class (binary) classification problem, given a set of instance-label pairs $(\boldsymbol{x}_i, y_i)$, $i = 1, ..., n$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\boldsymbol{y} \in \left\{1, -1\right\}^n$ (or possibly $\boldsymbol{y} \in \left\{1, 0\right\}^n$), the support vector machine (SVM) is a concept for assigning a new unknown instance $\boldsymbol{x}$ to one of these classes [41]. Algorithms operating in feature spaces generally use the idea: the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^d$ is via a nonlinear mapping,

$$\Phi : \mathbb{R}^d \to \mathcal{F}$$

$$\boldsymbol{x} \mapsto \Phi(\boldsymbol{x})$$

mapped to a higher or potentially infinite [41] dimensional feature space $\mathcal{F}$. Given a learning problem, the same algorithm is considered in $\mathcal{F}$ instead of in $\mathbb{R}^d$, which means that one works with the sample

$$(\Phi(\boldsymbol{x}_1), y_1), \ldots, (\Phi(\boldsymbol{x}_n), y_n) \in \mathcal{F} \times \boldsymbol{y}.$$

Behind this is the motive to find, given this mapped representation, a simple classification or regression in $\mathcal{F}$ [42]. For the SVM, the aim is to find a linear separating hyperplane in this higher dimensional space, which maximizes the margin of classification. In the non-separable case, i.e. when points might end up on the wrong side of the hyperplane, this requires a solution to the optimisation problem [42, 43, 41]:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}^T \Phi(\boldsymbol{x_i}) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, n, \end{aligned} \tag{2.6}$$

where $C > 0$ is a penalty parameter of the error term [41] and ξ_i slack variables introduced to relax the hard-margin constraints by allowing for some classification error [42].

The mapping Φ is never explicitly calculated because of the possibly high dimension of this space, which would require complex computations, but rather using a kernel function, $K(x_i, x_j) \equiv \Phi(x_i)^T \Phi(x_j)$ [44], with which a scalar product can implicitly be calculated in $\mathcal{F}$ with feature values as input [42, 43, 45]. This idea is often termed the "Kernel Trick" [43]. The inner product in (2.6) is then replaced by a kernel $K(x_i, x_j)$ which allows for the construction of classifiers that are linear in the feature space, although they are non-linear in the original space [43]. Solving for Lagrange multipliers α_i, a maximal margin separating

hyperplane in $\mathcal{F}$ defined by the kernel, can be obtained. The classifier has then a decision function given by [42, 44]:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b\right) \tag{2.7}$$

New kernels are being proposed by researchers, but four basic kernels commonly used are [41]:

- linear: $K(x_i, x_j) = x_i^T x_j$.
- polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^l, \gamma > 0$.
- radial basis function (RBF): $K(x_i, x_j) = exp(-\gamma \left\|x_i - x_j\right\|^2), \gamma > 0$.
- sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$.

Here, γ,r and l are kernel parameters.

In essence what the support vector machine does, is to maximise the margin between separable classes (see Figure 2.6). As classes seldom are completely separable, the regularisation term C is introduced. This parameter controls how many samples are allowed to be misclassified and consequently the smoothness of the decision boundary. The rational behind this is to avoid a classifier with bad generalisation ability, that overfits to training data [45]. The SVM is a deterministic optimisation problem i.e. the same data, regularisation term and kernel parameters will always produce the same result [45].

The performance of the SVM depends strongly on the choice of kernel and kernel parameters. A commonly used strategy for this model selection step is to start out with the Radial Basis function (RBF) kernel and then perform a grid search in order to find the best values for the parameters C and γ [41]. One motivation for using this kernel is that it nonlinearly maps samples into a higher dimensional space, which means that it can handle cases when the relation between the class labels and attributes is nonlinear [46].

2.7 Texture analysis by co-occurrence

Several methods for analysing texture exist including Fractal methods, Markov random field models as well as grey-level co-occurrence matrices (GLCM) [26]. Here, the traditional method of in-plane grey-level co-occurrences is described followed by an extension of this concept to 3-D.

2.7.1 Traditional grey-level co-occurrence

The GLCM approach was introduced by Haralick et al. [47] for grey-scale images. It is based on estimating the joint probability density function $F_{\alpha_1,\alpha_2}(\alpha_1, \alpha_2; d, \theta)$. Each $\hat{F}_{\alpha_1,\alpha_2}(\alpha_1, \alpha_2; d, \theta)$ is the estimate of the probability of going from grey-level α_1 to grey-level α_2 or from α_2 to α_1, given that the inter sample spacing is d and the direction is given by the angle θ [48]. This is illustrated by an example in Figure 2.7 where $d = 1$ and $\theta = 0$. The co-occurrence matrix is then obtained from the relative frequencies in the count matrix by dividing each element by the total number of counts.

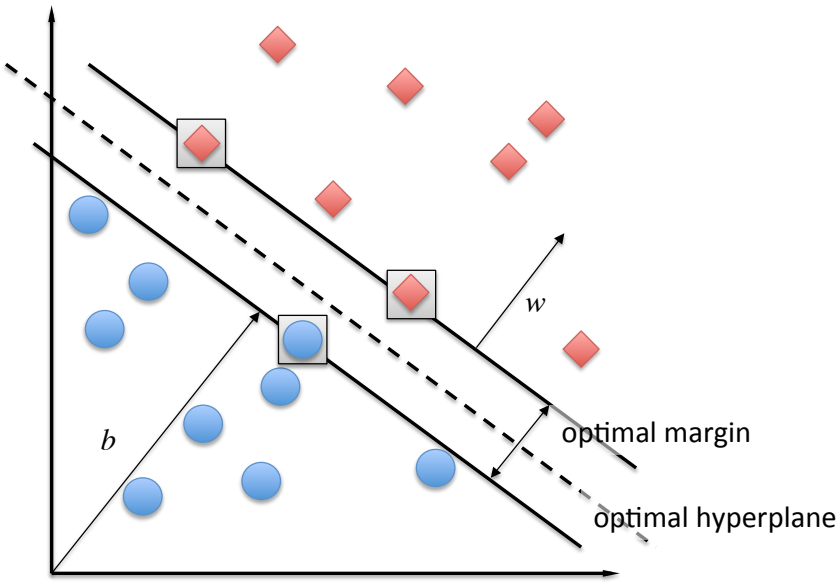


Figure 2.6: An example of a separating hyperplane in two dimensions. The support vectors, marked with grey squares, define the margin of largest separation. The orientation of the hyperplane is given by (w, b).

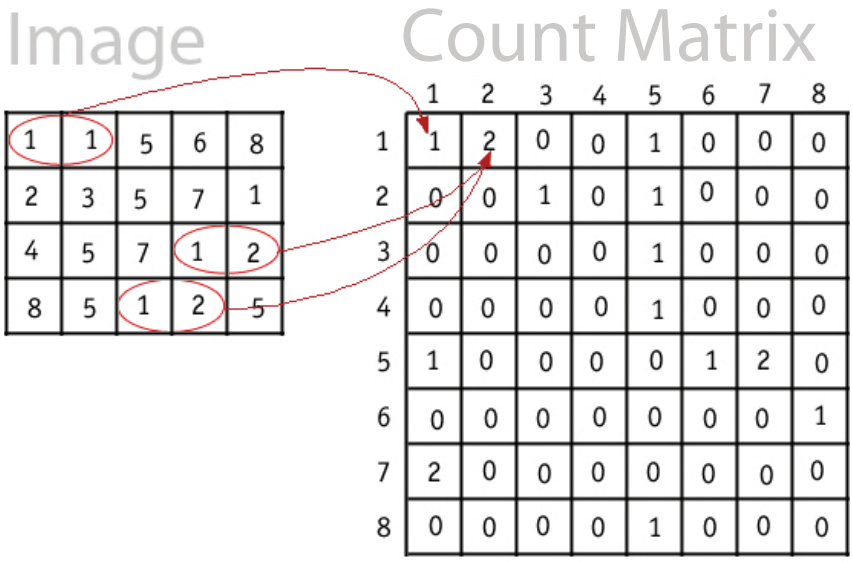


Figure 2.7: Illustration showing the principle of grey level co-occurrence. Co-occurrences with $d = 1$ and $\theta = 0$ are put into a count matrix.

Numerous features can be extracted from the resulting matrix, the most commonly used being the Haralick features. One of the Haralick features is entropy, defined as

$$Entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \hat{F}_{i,j}(i,j;d,\theta)\log\left(\hat{F}_{i,j}(i,j;d,\theta)\right). \tag{2.8}$$

For more details on the Haralick features see [47].

2.7.2 Multispectral co-occurrence

Kale et al. [48] extended the idea of GLCM to multispectral co-occurrence in 3-D. In this approach it is the co-occurrence of three different parameter values that is considered. This can be thought of as e.g. RGB values where each voxel has three values (R,G,B). Co-occurrence is considered over the three parameter planes in a window about a given voxel.

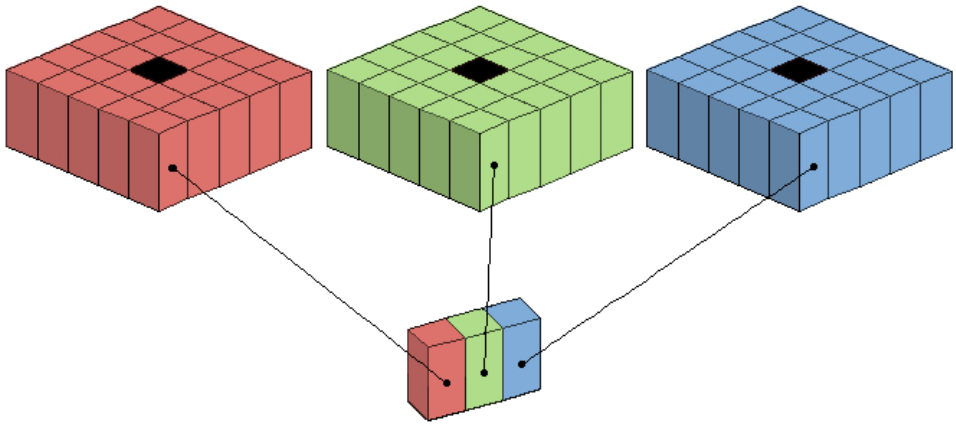


Figure 2.8: Illustration of one sample of co-occurring parameters within a volume.

By considering co-occurrences between three parameter planes, this extends the concept of the co-occurrence matrix to three dimensions. Each dimension of the 3-D co-occurrence array corresponds to the grey-level intensities of each parameter value. In a similar fashion to that of traditional grey-level co-occurrence, an entry in the co-occurrence array, $\hat{F}_{i,j,k}(i,j,k)$ is given from dividing the corresponding element in the 3-D count array by the total number of counts. $\hat{F}_{i,j,k}(i,j,k)$ is then the relative frequency of the RGB-triplet (i,j,k) or in other words an estimate of the probability of encountering this triplet within the RGB image. Features are extracted from the co-occurrence array in a similar fashion as in traditional co-occurrence. For example the feature Entropy is formulated

$$Entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\sum_{k=1}^{N_g} \hat{F}_{i,j,k}(i,j,k)\log(\hat{F}_{i,j,k}(i,j,k)). \tag{2.9}$$

For the complete set of features and their mathematical formulations, the reader is referred to Kale et al. [48].

3 Literature review

This section briefly reviews the existing literature on the topic of automatic segmentation of lesions in breast MRI. The study was conducted using resources made available through Chalmers Library and The Gothenburg University Library. These resources were mainly articles from scientific databases, such as IEEE Xplore, Medline and Inspec, but also books and dissertations. Only a handful of studies had previously been conducted in the field of automatic segmentation of breast lesions, although there existed several solutions for semi-automatic approaches. These solutions, semi-automatic as well as automatic, were evaluated in terms of limitations, how well they compared against other methods and how they could be improved.

Numerous approaches on semi-automatic segmentation exist. A selection are presented in Table 3.1. These approaches were either utilise a manually selected seed-point or a region of interest as the starting point for the segmentation procedure. Such methods, therefore, are not easily adapted to perform automatic segmentation.

Table 3.1: An overview of selected articles using semi-automatic approaches for lesion segmentation.

Title	Year	Segmentation method	Results
A Fuzzy C-Means (FCM)-based approach for computerized segmentation of breast lesions in Dynamic Contrast-Enhanced MR Images [5]	2006	Fuzzy C-Means segmentation	Correct segmentation of 98.7% of the malignant lesions and 93.2% of the benign (overlap threshold of 0.4).
Computer-aided diagnosis and visualization based on clustering and independent component analysis for breast MRI [49]	2008	Independent Component Analysis	Best AUC: 0.8388.
Malignant lesion segmentation in contrast-enhanced breast MR images based on the marker-controlled watershed [7]	2009	Marker controlled watershed transform	Overlap ratio (Jaccard index) of 64.3±10.4% with two ground truth segmentations.
Robust segmentation of mass-lesions in Contrast-Enhanced dynamic breast MR Images [50]	2010	Automatic intensity threshold estimation and connected component analysis	64% overlap with manual segmentation.

Two main categories of automatic methods were identified: those using enhancement characteristics only and those using co-occurrence analysis in combination with voxel-wise classification. These are exemplified by the methods listed in Table 3.2. The traditional way to execute this co-occurrence analysis is by extracting in plane co-occurrences and extracting statistical features, as defined by Haralick et al. [47] (see Section 2.7), from the whole lesion. Although shown to produce good results, delineation of the lesion is needed before undertaking this study, and it is also evident that this strategy would be problematic for textural analysis of smaller secondary lesions [26]. The strategies of voxel-wise segmentation overcome these issues and are suitable for automation.

Table 3.2: An overview of selected articles using automatic approaches for lesion segmentation.

Title	Year	Segmentation method	Results
Malignant-lesion segmentation using 4D co-occurrence texture analysis applied to Dynamic Contrast-Enhanced Magnetic Resonance breast image data [51]	2007	Voxel-wise classification using a co-occurrence-based texture analysis method	True positive fraction of 0.96 for a given false positive fraction of 0.0015.
Providing context for tumor recognition using the wrapper framework [16]	2007	Minimisation of distance to a priori classes using k-NN in an iterative procedure	92% correct classification of tumour types.
Texture analysis of lesion perfusion volumes in dynamic contrast-enhanced breast MRI [52]	2008	Three-time-point curve classification and textural analysis by co-occurrence on the resulting labels	Best feature accuracy: 74.3%.
Multispectral co-occurrence with three random variables in Dynamic Contrast Enhanced Magnetic Resonance Imaging of breast cancer [48]	2008	Voxel-wise classification using 3D co-occurrence analysis on the parameters of a pharmacokinetic model	True positive fraction of 0.8079 for a given false positive fraction of 0.0018.
A fully automatic lesion detection method for DCE-MRI fat suppressed breast images [13]	2009	Lesion detection using contrast uptake and covariance analysis	93% sensitivity.

In a study by Woods et al. [51], an automatised approach was proposed using the concept of a local 4D window (including the time dimension) co-occurrence texture analysis on the raw MR data. An artificial neural network (ANN) classifier was used to classify each voxel as malignant or non-malignant. Performance comparable to that of an analysis conducted by a human, using a pharmacokinetic two-compartment model as aid, was presented. As the classifier was applied to raw data, the classifier would have to be retrained for each new pulse sequence and hardware/software combination.

A similar concept using co-occurrence of three random variables was explored by Kale et al. [48]. The distributed statistical co-occurrence of three parameters from a pharmacokinetic two-compartment model, at matched spatial positions, was used to provide information on the tissue vascularisation properties. By scanning a 3-D local spatial window of observation for each voxel, a co-occurrence array was acquired for each position. 3D co-occurrence features were calculated in a similar fashion as the traditional ones. An ANN was used to classify voxels as either malignant or benign. This study used the parameters from the pharmacokinetic model as basis for the co-occurrence matrices and its performance was tested only on invasive ductal carcinomas. By use of the model parameters as input the problem of restricting these values to a limited range occurs. This must be done carefully in order to avoid letting outliers in the data govern the resolution of the co-occurrence analysis. Further, as the parameters were the sole basis for the analysis,

much temporal information was left unexplored.

From the literature study it was concluded that textural analysis by co-occurrence in a voxel-wise fashion is a promising approach to automatic segmentation of lesions. Further it was concluded that the use of a voxel score or probability between zero and one, based on the information from several temporal features, would be suitable for use in co-occurrence analysis. This would overcome the problem of truncating the parameters or features and provide flexibility, because the same co-occurrence procedure can be performed independently of which features are used to compute this probability. The proposed segmentation algorithm is presented in Section 4.

4 Proposed segmentation method

This section presents the new algorithm, developed by the authors, for automatically segmenting lesions in DCE-MRI data. The algorithm is hereinafter referred to as the TSC (temporal score co-occurrence) method. A flowchart showing the individual steps that make up the TSC algorithm is shown in Figure 4.1. The algorithm takes as input a set of DCE-MRI volumes $\{V_0, \ldots, V_n\}$ and the corresponding set of acquisition times $\{t_0, \ldots, t_n\}$. The steps are as follows:

1. Preprocessing steps include mean filtering the raw data and converting the data to relative enhancement.

2. A parametric model (e.g. Tofts model) is fitted to the data to yield a set of p parameter volumes $\{P_1, \ldots, P_m\}$ where m is typically 2 or 3.

3. For each voxel, features that characterise the voxel's temporal characteristics are extracted. These features are extracted both from raw MR data and from the intensity curve given from the parametric model. The result is a set of temporal feature volumes $\{Ft_{1,\ldots,}Ft_l\}$.

4. A (pre-trained) classifier is applied voxel-wise to each set of temporally based features to yield a volume of posterior probabilities of suspiciousness (i.e. each probability reflects the likelihood that the fitted model is that of a suspiciously enhancing tissue).

5. For each voxel in this probability volume, the co-occurrence of probabilities in a neighbourhood around this voxel is computed and then used to compute l co-occurrence features. The end result is a set of co-occurrence feature volumes $\{Fc_{1,\ldots,}Fc_m\}$.

6. The final set of features is comprised of the temporal and co-occurrence features together, yielding a set of l+m features.

7. A second (pre-trained) classifier is applied voxel-wise to each $(l+m)-$tuple of features to yield a single binary volume S containing 3D connected components that locate the suspicious tissues.

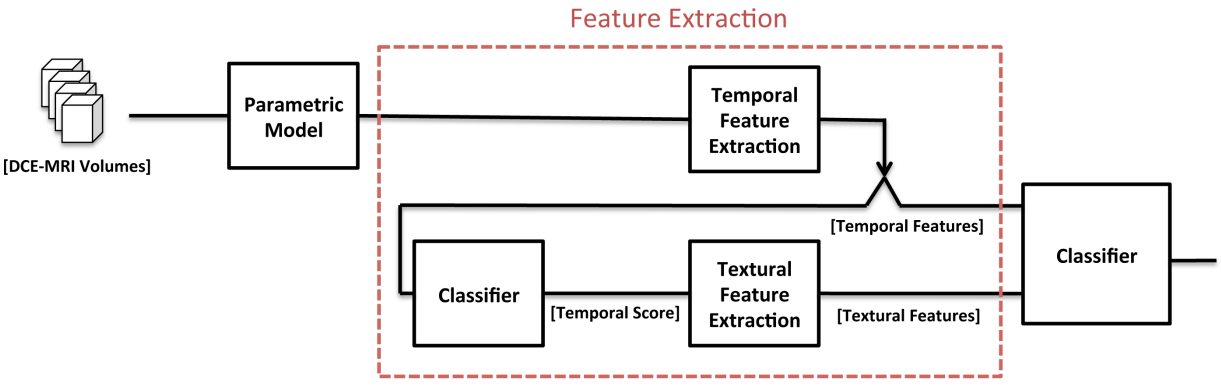


Figure 4.1: Flow chart of proposed segmentation method

5 Empirical evaluation using real clinical data

This section describes an empirical evaluation of the efficacy of the proposed new features (and segmentation method) for segmenting (voxel-wise classification) suspicious tissue in real clinical DCE-MRI data. Here "suspicious tissue" refers to voxels that an interpreting radiologist labelled as suspicious for malignancy and not tissue subsequently confirmed by cyto- or histopathology to be malignant. The evaluation was done in two stages. In the first stage the aim was to identify the most discriminatory subset of features from among the proposed new features and features extracted from only a single voxel time series (enhancement curve). In the second stage the aim was to compare the classification performance of a classifier based on this subset of features relative to one based on only the competing features of Kale [48].

5.1 MR image data

The DCE-MRI image data used for the evaluation originates from clinical breast examinations of 32 subjects. The data contains 45 lesions marked as suspicious by the reporting radiologist. Cytology or histology findings were available for all but one lesion confirming that 28 were malignant and 16 were benign.The DCE-MRI data were acquired on a 1.5T Signa HDxt (GE medical systems), at Queensland X-ray, Queensland, Australia. Pulse-echo time was 3.412 ms, repetition time 6.516 ms and flip angle 10 degrees. The data comprises axial slice images of size 512×512 pixels covering both breasts acquired using a fast spoiled gradient-recalled sequence. The in-slice resolution is typically 0.625 mm and slice thickness varies from 1-1.4 mm. Images in the data sets were fat-suppressed. Each DCE-MRI data set comprises a single pre-contrast volume and 4-5 post-contrast volumes. A subset of the DCE-MRI data sets (15) provided for the study had been preprocessed to correct for patient motion. Each lesion was located and segmented manually by an expert reader. This was performed using the region-growing tool in OsiriX. More specifically the tool was used to define a volume of interest (VOI) in the first difference volume (subtraction of the pre-contrast volume from the first post-contrast).

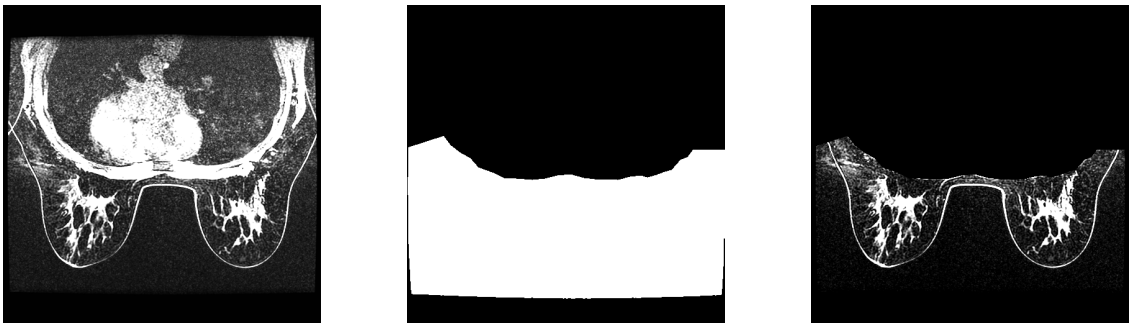
5.2 Pre-processing steps: filtering, normalisation, segmentation

In order to mitigate the influence from noise and mild motion artefacts in the raw MRI data sets, a local 3×3 mean filter was applied to each slice of each image volume. The chest-wall and breast-air boundaries were segmented as well, in order to reduce the data volume. The relative enhancement maps were computed from the raw MR data, using Equation 2.1.

5.2.1 Manual segmentation of the chest wall using OsiriX

The chest-wall boundary was segmented using OsiriX[5]. For a given DCE-MRI data set a 3D maximum intensity projection was computed for the precontrast volume and the scissor tool applied interactively to cut away the chest cavity. The remaining non-zero voxels were used to define a binary mask which was then applied to the remaining volumes in the data set. This procedure is illustrated in Figure 5.1.

[5]OsiriX is an image processing software dedicated to DICOM images, URL: http://www.osirix-viewer.com/

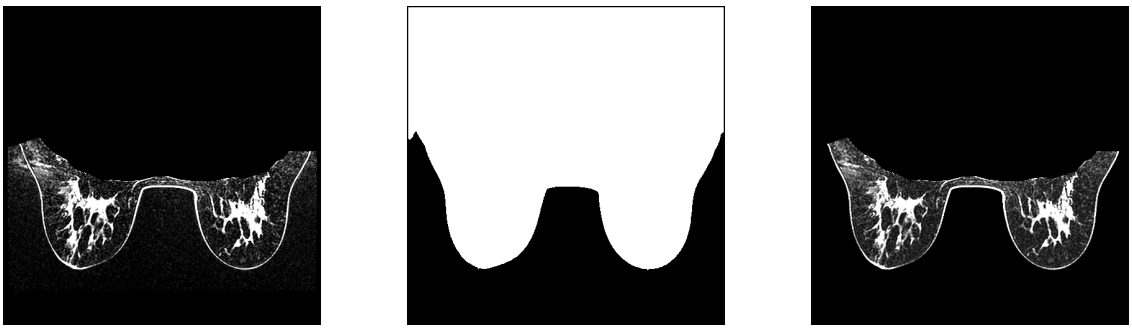


(a) Raw MR image (b) Manual mask (c) Masked result

Figure 5.1: The manual masking procedure, removing enhancing tissue inside of the chest wall.

5.2.2 Automatic segmentation of the breast-air boundary

The breast-air boundary was automatically segmented in each volume using the alogrithm proposed by [53]. The implementation of this algorithm in MATLAB was courtesy of Darryl McClymont, School of ITEE, The University of Queensland, Brisbane, Australia. Figure 5.2 shows a typical result.



(a) Manual segmentation result (b) Hayton mask (c) Final breast segmentation result

Figure 5.2: The automatic masking procedure, removing noise outside of the breasts.

5.3 Partitioning into training and validation sets

The data were divided into two disjoint sets: one for training the classifiers (training set) and one for validating the performance of each on unseen data (validation set). The training set was constructed using stratified random sampling, so that each set had approximately the same ratio of malignant to benign voxels. Approximately two-thirds of the data were used to define the training set as shown in Table 5.1.

5.3.1 Selection of suspicious and non-suspicious voxels for feature selection and classifier training

The smallest VOI in the training set comprises 54 voxels. Thus to ensure that each VOI had equal contribution to the class of suspicious voxels it was decided to sample only 54

Table 5.1: Cyto-/histopahtology status of the lesions in the training and validation sets

	Training set (22 subjects)	**Validation set (10 subjects)**
Malignant	19	9
Benign	12	4
Unknown	1	0

voxels from each. First all voxels not enhancing 50% or more in at least one post-contrast volume were excluded from the VOIs (this is the rate of enhancement that a radiologist typically considers when looking for suspicious tissue). Then for each VOI, 54 of the remaining voxels were selected for inclusion in the class of suspicious voxels. In the case of the smallest VOI only 52 of the 54 voxels passed the enhancement criterion. Consequently the class of suspicious voxels contained $54 \times 31 + 52 = 1726$ voxels. Voxels comprising the non-suspicious class were randomly sampled from regions within the breasts of slices with no VOI. The partitioning of the training data is shown in Table 5.2.

Table 5.2: Number of training samples in the two classes.

Class	**Number of training examples**
Suspicious (class label 1)	1726
Non-Suspicious (class label 0)	1738

5.3.2 Selection of suspicious and non-suspicious voxels for classifier validation

All voxels demonstrating negative relative enhancement or negative raw MR values at any time point were excluded from the VOIs in the validation set. All of the remaining VOI voxels were then used to define the class of suspicious voxels. All of the remaining voxels in the breast tissue were used to define the class of non-suspicious voxels. These two classes are hereinafter referred to as the full validation data set. A reduced version of this data set, hereinafter referred to as the reduced validation data set, was defined by removing the data for two of the subjects (two malignant lesions). The need for this reduced data set is explained in Section 5.6. The partitioning of the validation data is shown in Table 5.3.

Table 5.3: Number of validation samples in the two classes for the full and reduced validation data sets.

Class	**Number of validation examples**	
	Full	**Reduced**
Suspicious (class label 1)	22130	9490
Non-Suspicious (class label 0)	33941654	28145930

5.4 Fitting of parametric models of enhancement

Three different parametric models of enhancement were fitted to the normalised data. These models were the Hayton model, the Linear Slope model and the Ricker model shown in Table 5.4. The different shapes of the model curves are shown in Figure 5.3 where these models have all been fitted to the same voxel. For more detailed descriptions of these models the reader is referred to Appendix A or [27].

Table 5.4: The utilised parametric models listed.

Model	Mathematical formulation
Hayton model	$C(t) = \frac{A}{a-b}(e^{-bt} - e^{-at})$
Linear Slope model	$C(t) = \begin{cases} \beta_1 t & \text{if } t \leq \alpha \\ \beta_1 \alpha + \beta_2(t-\alpha) & \text{if } t > \alpha \end{cases}$
Ricker model	$C(t) = ate^{-bt}$

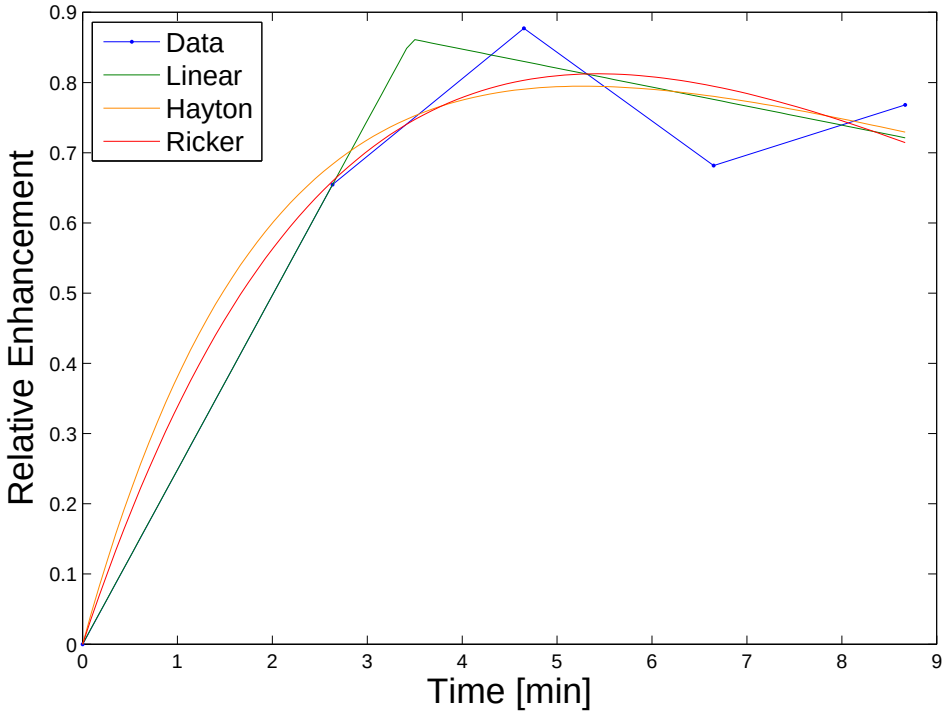


Figure 5.3: All considered models fitted to the same data points.

5.4.1 Fitting the parametric models

The Hayton model was fitted voxel-wise to each DCE-MRI data set using the function `lsqcurvefit` in MATLAB and the non-linear least squares (NLS) trust region reflective algorithm. Starting values for the model were those recommended by Hayton [53].

The Linear Slope model was fitted using a finite number of linear least squares (LS) fits as described in [54].

The Ricker model was implemented using LS.

5.5 Feature extraction: Temporal-score co-occurrence method

TSC features were extracted separately for each of the three fitted parametric models. For a given model the extraction of these features involved five steps. In the first step, features from the signal-intensity time curve and raw data were extracted. In the second

step stepwise multilinear regression was used to select a subset of these features. In the third step these features were used to define a feature space. A k-NN classifier was used to generate a score, representing degree of suspiciousness, for each voxel. In the fourth step a grey-level co-occurrence matrix was computed for each voxel (or rather a neighbourhood about the voxel) in the resulting score volume. Finally, in the fifth and last step, each GLCM was used to compute several GLCM features per voxel. See Figure 5.4 for an overview of the complete implementation. As shown in this flow chart, the signal-intensity time curve and raw features were also passed on to the final classification step without processing.

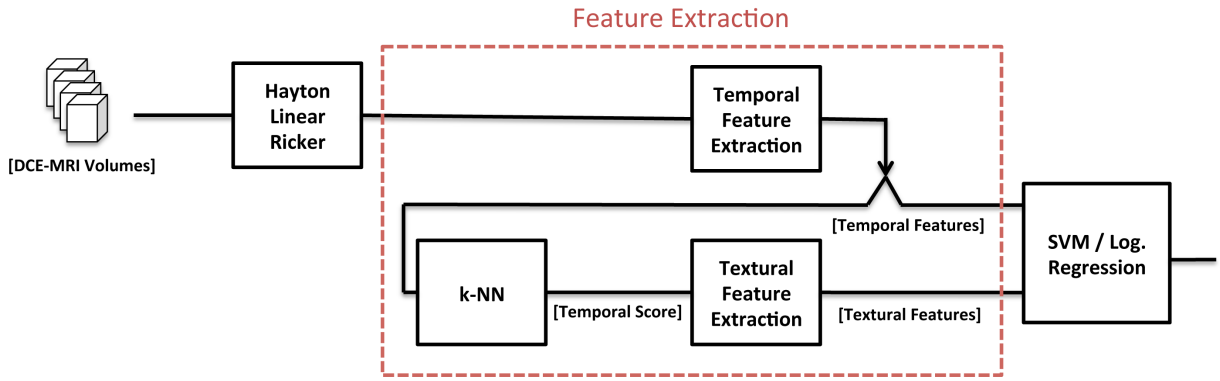


Figure 5.4: Flow chart of the implemented segmentation method.

5.5.1 Step 1: Signal-intensity time curve and raw data features

Signal-intensity time curve and raw data features, hereinafter denoted as the trivial features, were extracted for the training data set. Table C.1 and Table C.2 of Appendix C lists all trivial features.

5.5.2 Step 2: Feature selection by stepwise linear regression

Trivial features with low probability of influencing the predictive power of the k-NN models positively, were excluded using a stepwise multilinear regression method. This was done using MATLAB's stepwise regression function, `stepwisefit`, starting with an empty model. An F-statistic was computed to test the models with and without a potential term.

5.5.3 Step 3: Temporal score extraction

For each fitted enhancement model a k-NN classifier was used to generate a probability of suspiciousness (i.e. of membership of the suspicious class) for each voxel. This probability was defined to be the proportion of the k nearest neighbours that have the label suspicious. The features remaining after the linear regression step were used as input to the k-NN classifier. Features were not scaled before use in this procedure.

An example of what this score map produced by the k-NN classifier may look like for one slice is given in Figure 5.5, where higher intensities represent a higher score.

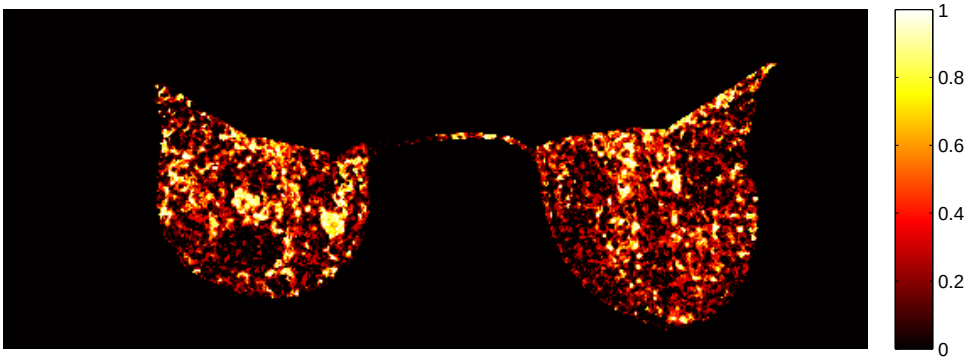


Figure 5.5: An example of resulting scores from applying the k-NN step.

The resolution of the score feature was limited by the number of neighbours, k. This number, k, and the distance metric used were selected via an exhaustive leave-one-out strategy. For evaluating k and distance metrics, correct classification rate (CCR) was used as performance measure at a cut-point of 0.5. The range of k was [1, 100] and the distance metrics tested were the default ones present in MATLAB's `knnsearch` implementation, e.g. Euclidean, Mahalanobis and Chebyschev. The k-value resulting from this procedure was set as the maximum resolution for the construction of the GLCM from values, by quantising score data into this number of bins.

5.5.4 Step 4: Spatial co-occurrence feature extraction

The textural characteristics surrounding each voxel were quantified by evaluating the co-occurrence of scores within a window containing the current voxel. In order for the window to be as close as possible to the shape of a cube, the in-plane dimensions of the window was 5×5 voxels and the through-plane dimension was 2. This was due to the anisotropic voxel size, in other words, the slice thickness being greater than the in-plane dimensions. See Figure 5.6 for an illustration of this window.

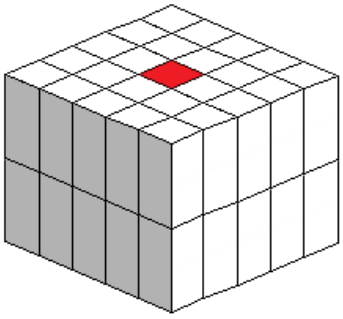


Figure 5.6: Illustration of window, were the red element is representing the centre voxel.

For all voxels that fulfilled a criterion, textural features were extracted. This criterion stated that the relative enhancement was to be positive at all post-contrast injection time points and that the pre-contrast intensity raw-value was to be larger than zero. In order to extract features from the window each voxel inside the window needed to be assigned a score from the k-NN step. For voxels within the window that did not fulfil the criterion the

score was set to zero. Textural feature extraction was not made in cases were the boundary of the acquisition volume was within the window.

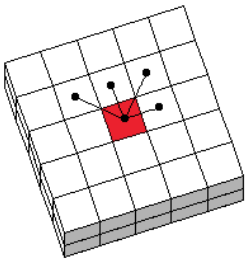


Figure 5.7: Illustration of in-plane co-occurrence directions 0, 45, 90, 135 degrees, for one voxel.

For each slice within the current window, texture analysis based on co-occurrence was performed as described in 2.7. Co-occurrence analysis was performed on the two slices separately, considering in-plane directions 0, 45, 90 and 135 degrees and the distance 1. Figure 5.7 shows an illustration of these co-occurrence directions. The resulting count matrices from the two slices of the window were pooled into one matrix containing all in-plane counts within the window. The total number of counts in this matrix was 288, since counts were given for moving in both directions between voxels.

5.5.5 Step 5: GLCM features

Features were extracted from the resulting co-occurrence matrix using standard Haralick features. A list of these features can be found in Table 5.5. For a further description of these features and a mathematical formulation see Haralick et al. [47].

Table 5.5: Textural features extracted from the 3D co-occurrence matrix; TSC method.

Textural Features, TSC
X1. Energy
X2. Entropy
X3. Contrast
X4. Variance
X5. Sum Mean
X6. Inertia
X7. Cluster Shade
X8. Cluster Tendency
X9. Cluster Shade
X10. Homogeneity
X11. MaxProbability
X12. Inverse Variance

5.6 Feature extraction: Multispectral co-occurrence method

The feature extraction for the MSC method was performed in two steps. The first step limited the span of model parameter values by setting outliers to the limit value. Both

training and validation data was subject to this truncation step. The implementation of the multispectral method differed in several respects from the original one. However the second step, feature extraction, remained unchanged but utilised parameter values from the Hayton and Linear Slope enhancement models used as input.

5.6.1 Step 1: Truncation of parameters

In order to make sure that outliers among the parameter values would not dictate the span of values, the parameters were truncated. This truncation was conducted for each parameter by observing the parameter values from five slices of two patients and the corresponding parameter values from within the pre-segmented VOIs of these patients in a box-plot. These two patients were selected from the full validation data set. In order to avoid bias in the classification result by including these patients in the final evaluation, these subjects were removed from this data set. See Table 5.3 for the partitioning of this reduced data set. Figure 5.8 shows an example of this for the β_2 parameter of the Linear Slope model. The whiskers of these box-plots were set so that for normally distributed sets 99.3 % or $\pm 2.7\sigma$ of the parameter values were within the whiskers. The number of voxels in the set containing parameters from five slices of the two patients was significantly higher than the set containing parameters from the VOI. Therefore, in order not to remove valuable parameters, the whiskers that spanned the largest values of the breast and VOI parameters was chosen as the limit of the parameter. Values outside of this limit were considered outliers. The truncation was performed for the training data set and the reduced validation set.

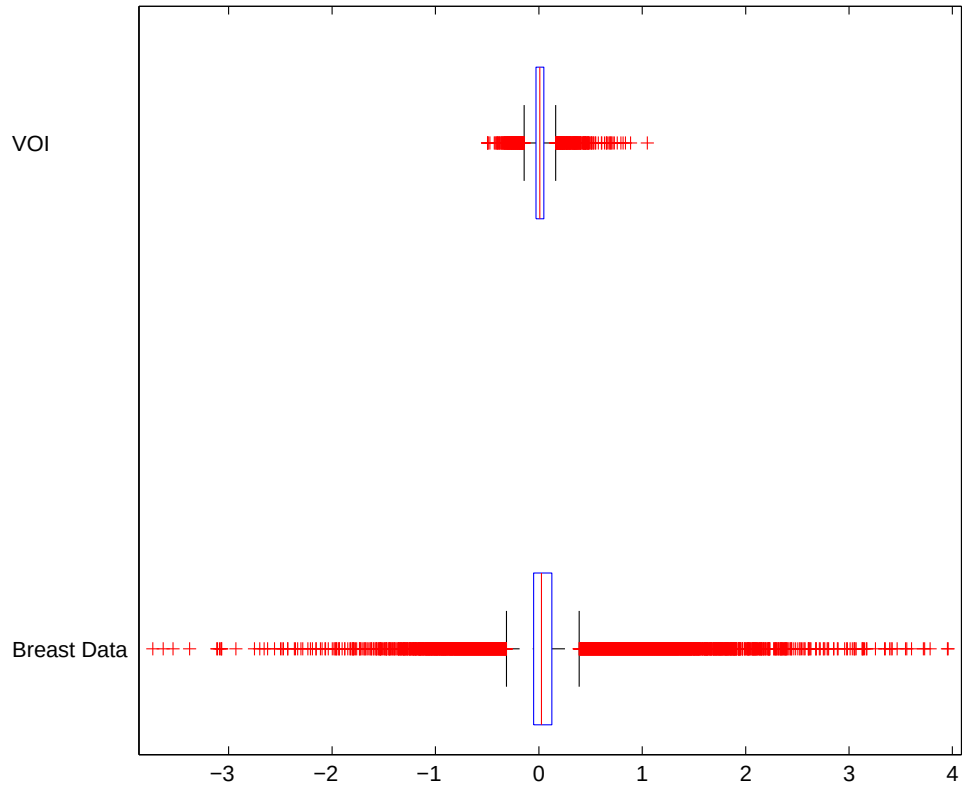


Figure 5.8: A boxplot of the β_2 parameter of the Linear Slope model, parameters from five slices of the breast and parameters from only within the VOI.

5.6.2 Step 2: Implementation of the MSC method

The implementation of multispectral co-occurrence analysis of this study was influenced by that of Kale et al. [48]. The model selection steps used together with the MSC feature extraction method was the same as that used with the TSC feature extraction method, see Figure 5.9 and Figure 5.4. For the theoretical methodology of multispectral co-occurrence analysis, see 2.7.2.

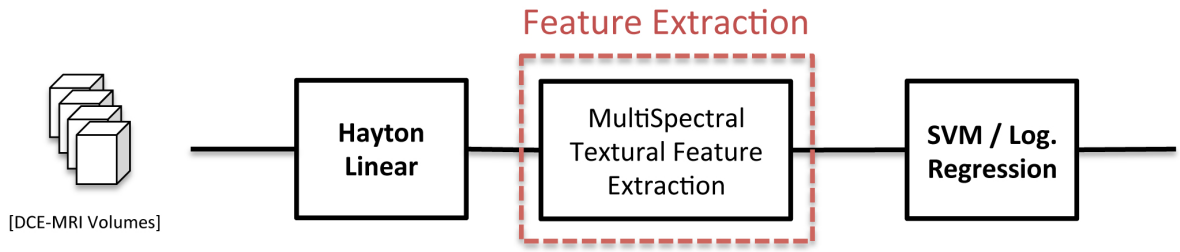


Figure 5.9: Flow chart of the implemented segmentation method.

The MSC method utilised the three parameter planes from the Hayton and the Linear Slope model. The size of the window on which the multispectral co-occurrence was applied was $5 \times 5 \times 2$ voxels. The size of the co-occurrence array was set to $32 \times 32 \times 32$ in accordance with the method of Kale et al. [48]. This size dictated the number of bins that the parameter values were quantised into. The co-occurrence array thus consisted of a total of $32^3 = 32768$ cells. An illustration of a parameter triplet giving rise to a count in the count array is given in Figure 5.10.

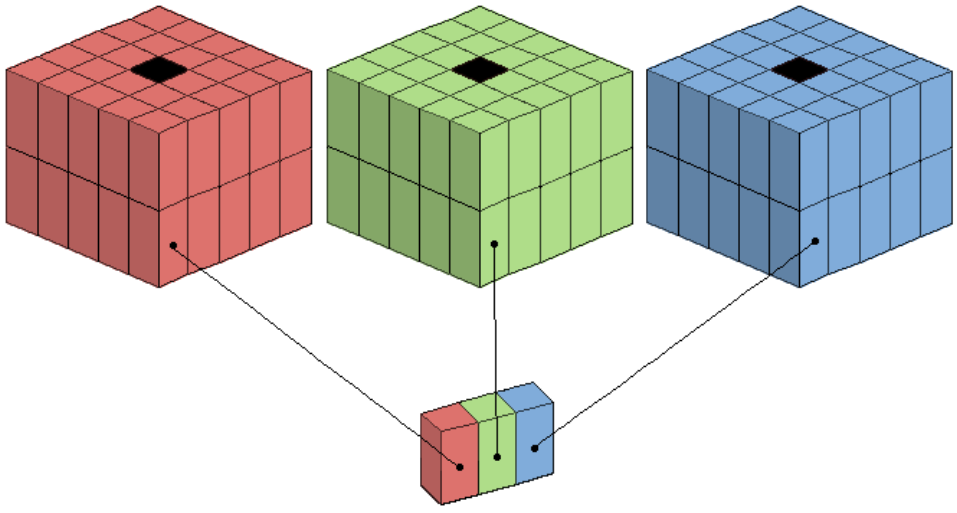


Figure 5.10: Illustration of one sample of co-occurring parameters within a window of size $5 \times 5 \times 2$.

The features extracted from the co-occurrence array are stated in Table 5.6, for the complete mathematical formulation of these features, see Kale et al. [48].

Table 5.6: Textural features extracted from the 3D co-occurrence matrix; MSC method.

Textural Features, MSC
XK1. Angular Second Moment
XK2. Central Moment
XK3. Sum Average
XK4. Sum Central Moment
XK5. Sum Entropy
XK6. Entropy
XK 7,8,9. Contrast
XK 10,11,12. Inverse Difference Moment
XK 13,14,15 Correlation
XK 16,17,18 Difference Variance
XK 19,20,21 Difference Entropy

5.7 Feature selection

Two feature selection strategies were used, sequential forward selection and stepwise logistic regression in order to find the most discriminatory feature subset. In the feature selection process of the TSC method, the trivial features were reintroduced.

5.7.1 Feature selection using Sequential Forward Selection

The sequential forward selection (SFS) method chose the feature that increased the AUC measure most, and added it to the previously selected ones. The AUC was evaluated using a logistic regression classifier with 10-fold cross-validation. This procedure was repeated until all features had been selected. At this point all best performing feature sets, over the full range of features, had a corresponding AUC value. For the TSC method six features were selected for all different time-curve models. The AUC values had for this number reached a point where adding more features had only marginal influence on the result. For the MSC method ten features were selected for both the time-curve models, with the same motivation.

5.7.2 Feature selection using Stepwise Logistic Regression

In order to avoid choosing too large feature sets, the Bayesian information criterion which applies a logarithmic penalty as features are added, was used. This procedure was coded in the statistically oriented programming language R, code is given in Appendix B. The model was trained using all features, and features were added and removed iteratively.

5.8 Model selection

The model referred to in this section is the final set of features given by the feature selection step and the final classifier. Two different types of classifiers were used[6], one from the family of SVM classifiers and one logistic regression classifier. As the logistic regression classifier does not require any tuning of parameters, this section primarily concerns the SVM.

[6]The classifiers were implemented in MATLAB using PRTools (URL: http://www.prtools.org/) and LIBSVM (URL: http://www.csie.ntu.edu.tw/~ cjlin/libsvm/).

5.8.1 Feature scaling

Large feature scale differences can have a great effect on the performance of the SVM. All features were thus scaled by subtracting the mean of that particular feature and dividing by its standard deviation. Even though this is not required for logistic regression, the same step was performed in order to be consistent.

5.8.2 Kernel and parameter choice for the SVM

As mentioned in 2.6.3, the strategy of choosing kernel and kernel parameters was made according to Hsu et al. [41]. The RBF kernel had two tuning parameters, (C, γ).

In order to reduce the risk of overfitting kernel parameters to the training data, 5-fold cross-validation was used along with the grid search. Initially parameter pairs were selected as all combinations of the exponentially growing sequences $C = 2^{-5}, 2^{-3}, ..., 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$. For each tested parameter pair, the performance of the classifier was evaluated by averaging together the AUC over each of the five folds. When the best combinations of C and γ were found in the initial grid, the resolution of the grid was increased and the performance for each parameter pair was evaluated. This procedure continued until the values had stabilised within the current grid.

5.9 Evaluating classifier performance

The measure used for evaluating the segmentation performance of the competing methods was the AUC. Evaluation of the final models was performed by use of the validation data set, for which each voxel was classified.

5.10 Results

In this part the results from the feature and model selection strategies, as well as the classification results from the final model, are presented. The performance measure under which the models were evaluated, was the area under the receiver operating characteristic curve, in short AUC.

5.10.1 Feature extraction; Temporal score co-occurrence method

By the result from the stepwise linear regression step, the trivial features T14 and T15 (see Appendix C) for the Hayton model were not included in the k-NN model. Features TL2 and TL3 for the Linear Slope model and T2, T4, T7, T9, T10, T11, T13 for the Ricker model were not included in their respective k-NN model. See Table 5.7 for a summary of these results.

Table 5.7: Resulting features from the stepwise linear regression step:

Model	Selected Features
Hayton	T1, T2,T3,T4,T5,T6,T7,T8,T9,T10,T11,T12,T13
Linear Slope	TL1,TL4,TL5,TL6,TL7,TL8,TL9
Ricker	T1,T3,T5,T6,T8,T12,T14

The results from the exhaustive search for a distance metric and optimal number of neighbours are presented in Table 5.8. In comparison to the MSC method which had a resolution of 32 grey-levels, these k-values are more than a halving in resolution.

Table 5.8: Resulting distance metrics and neighbours from the leave-one-out cross-validation for each model:

Model	**Distance Metric**	**Neighbours (k)**
Hayton	Mahalanobis	13
Linear Slope	Mahalanobis	13
Ricker	Mahalanobis	7

5.10.2 Feature selection using Sequential Forward Selection

From the sequential forward selection method, six and ten features were selected for the TSC and MSC models respectively. No trivial features were selected for either of the TSC models. The textural features that were selected, were frequently appearing for all the three different enhancement models. Half of the features for the MSC model were common to both the Hayton and Linear Slope models. See Table 5.9 for complete results from the SFS strategy.

Table 5.9: Features selected by SFS, in order of selection:

Model	**Selected Features**	$\overline{AUC} \pm \hat{SE}$
Hayton (TSC)	X7, X5, X12, X4, X9, X6	0.9395 ± 0.0037
Linear Slope (TSC)	X7, X5, X12, X11, X9, X6	0.9423 ± 0.0032
Ricker (TSC)	X7, X5, X12, X4, X9, X6	0.9312 ± 0.0016
Hayton (MSC)	XK11, XK3, XK16, XK21, XK10, XK19, XK20, XK8, XK9, XK17	0.8972 ± 0.0016
Linear Slope (MSC)	XK9, XK16, XK3, XK5, XK10, XK18, XK7, XK12, XK20, XK14	0.8953 ± 0.0012

5.10.3 Feature selection using Stepwise Logistic Regression

By the stepwise logistic regression feature selection method, similar textural features were extracted for the TSC method. Trivial features were also included, in contrast to the SFS strategy. No strong tendency to choose the same features as the SFS method was observed. For the MSC method, half of the features were consistent between the Hayton and Linear Slope model, as for the SFS strategy. A large overlap of selected features can be observed between the feature selection methods for the MSC method. See Table 5.10 for complete results from the SLR feature selection strategy.

5.10.4 Model selection

Three grid searches per model were conducted in order to get adequate resolution. The γ-parameter tended to move towards the upper bounds of the grid search, while no clear behaviour was observed for C. The values of C and γ ranged from 0.125 to 2218.7 and 3.04 to 8 respectively, for the TSC method. For the MSC method, C and γ ranged from 912 to 14564 and 5.45 to 7.53. Table 5.11 summarises the results.

A typical grid surface is illustrated in Figure 5.11. A steep change between two plateaus was a characteristic behaviour for the AUC surface of the initial search area. AUC values stabilised quickly as the resolution of the search grid was increased.

Table 5.10: Features selected by SLR, in no particular order:

Model	**Selected Features**
Hayton (TSC)	X1, X2, X4, X6, X7, T4, T10, T11
Linear Slope (TSC)	X1, X4, X6, X7, TL1, TL4, TL8, TL9
Ricker (TSC)	X1, X4, X6, X7, T1, T3, T4, T9, T10, T14
Hayton (MSC)	XK1, XK3, XK4, XK5, XK7, XK8, XK9, XK12, XK16, XK17, XK18, XK21
Linear Slope (MSC)	XK1, XK3, XK5, XK7, XK8, XK9, XK10, XK12, XK13, XK14, XK19

Table 5.11: Resulting RBF kernel parameters:

Model	C_{SFS}	γ_{SFS}	$\overline{AUC}_{SFS} \pm \hat{SE}$	C_{SLR}	γ_{SLR}	$\overline{AUC}_{SLR} \pm \hat{SE}$
Hayton (TSC)	0.125	7.55	0.9300 ± 0.0045	512	3.04	0.9447 ± 0.0040
Linear Slope (TSC)	2218.7	6.6	0.9301 ± 0.0045	554.7	5.3	0.9538 ± 0.0037
Ricker (TSC)	0.292	8	0.9256 ± 0.0047	424.89	7.69	0.9287 ± 0.0046
Hayton (MSC)	14564	7.53	0.9237 ± 0.0047	912	5.45	0.9387 ± 0.0042
Linear Slope (MSC)	3297.55	6.22	0.9316 ± 0.0045	3236.55	6	0.9297 ± 0.0046

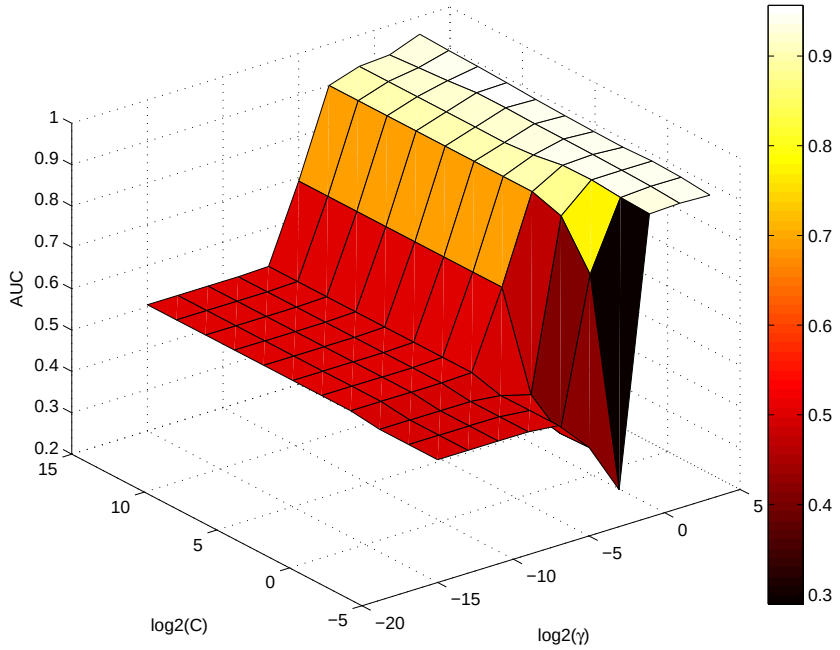


Figure 5.11: Example of grid search surface from the initial coarse grid using the Linear Slope model features from forward selection, to train the SVM.

5.10.5 Performance of the TSC-classifiers on the full validation set

The results from computing the AUC measure from the ROC curves for all different models, on the ten patient validation set, is presented in Table 5.12.

Table 5.12: Resulting AUC values from validation set containing ten patients, for the TSC models:

AUC				
Model:	**SFS**		**SLR**	
	SVM	LR	SVM	LR
Hayton	0.8472± 0.0016	**0.8788 ± 0.0015**	0.5394 ± 0.0020	**0.8683 ± 0.0016**
Linear Slope	0.7280± 0.0019	0.8271 ± 0.0017	**0.6844 ± 0.0015**	0.8260 ± 0.0017
Ricker	**0.8638 ± 0.0016**	0.8748 ± 0.0015	0.6254 ± 0.0020	0.6172 ± 0.0026

5.10.6 Performance of the TSC-classifiers and MSC-classifiers on the reduced validation set

The resulting AUC values for all models are shown in Table 5.13. The best classifiers in terms of the AUC value were a logistic regression classifier used on SLR features from the Linear Slope model (AUC=0.8989±0.0021) for the TSC method, and a logistic regression classifier used on SFS features from the Linear Slope model (AUC=0.9330±0.0018) for the MSC method. How the performance for each of these models at different operating points varied, given in terms of sensitivity, specificity and CCR, is shown in Table 5.14 and Table 5.15. The corresponding ROC-curves are shown in Figure 5.12.

Table 5.13: Resulting AUC values from validation set containing eight patients:

AUC				
Model:	**SFS**		**SLR**	
	SVM	LR	SVM	LR
Hayton (TSC)	0.8621±0.0024	0.8964±0.0022	0.5995±0.0027	0.8957±0.0022
Linear Slope (TSC)	0.8345±0.0026	0.8801±0.0023	0.7441±0.0020	0.8989±0.0021
Ricker (TSC)	**0.8747±0.0023**	0.8877±0.0022	0.6185±0.0026	0.7000±0.0030
Hayton (MSC)	0.8383±0.0026	0.9279±0.0019	0.7298±0.0030	0.9243±0.0019
Linear Slope (MSC)	0.8178±0.0027	**0.9330±0.0018**	**0.8671 ±0.0024**	**0.9266±0.0019**

Table 5.13 shows that for eight test patients and for the TSC method, the Ricker model was the best model for SFS and SVM, the Hayton model for the SFS and LR, whilst the Linear Slope was the best model for SLR and LR. The logistic classifiers for all MSC models performed similarly on all feature sets and outperformed the SVM. The best model was logistic regression in combination with SFS, for the MSC method.

5.10.7 Visual assessment of segmentation result

Segmentation results are presented for two different patients, at cut-point levels 0.7, 0.9 and 0.99. Under the 0.7 level, considerable over-segmentation was observed. For the first example, segmentation results for the TSC and MSC methods are presented in Figure 5.13 and 5.14. For the second example, segmentation results for the TSC and MSC methods are presented in Figure 5.15 and 5.16. True Positives (TP) are colour coded as green, false positives (FP) are coded as yellow, false negatives (FN) are coded as red and true negatives (TN) are coded in grey-scale.

Table 5.14: Classification summary for the best performing logistic classifier using features from the TSC method:

Cut-point	**Sensitivity**	**Specificity**	**CCR (%)**
0.0	1.00	0.00	0.03
0.1	0.98	0.18	18.29
0.2	0.97	0.30	29.88
0.3	0.96	0.39	39.05
0.4	0.94	0.46	46.22
0.5	0.93	0.52	52.32
0.6	0.92	0.58	57.97
0.7	0.90	0.64	63.63
0.8	0.88	0.70	69.80
0.9	0.84	0.78	77.59
1.0	0.00	1.00	99.97

Table 5.15: Classification summary for the best performing logistic classifier using features from the MSC method:

Cut-point	**Sensitivity**	**Specificity**	**CCR (%)**
0.0	1.00	0.00	0.03
0.1	1.00	0.12	12.10
0.2	0.99	0.25	24.74
0.3	0.99	0.36	36.14
0.4	0.98	0.46	46.10
0.5	0.96	0.54	54.16
0.6	0.95	0.61	61.36
0.7	0.93	0.68	68.31
0.8	0.90	0.76	75.61
0.9	0.86	0.84	84.19
1.0	0.00	1.00	99.97

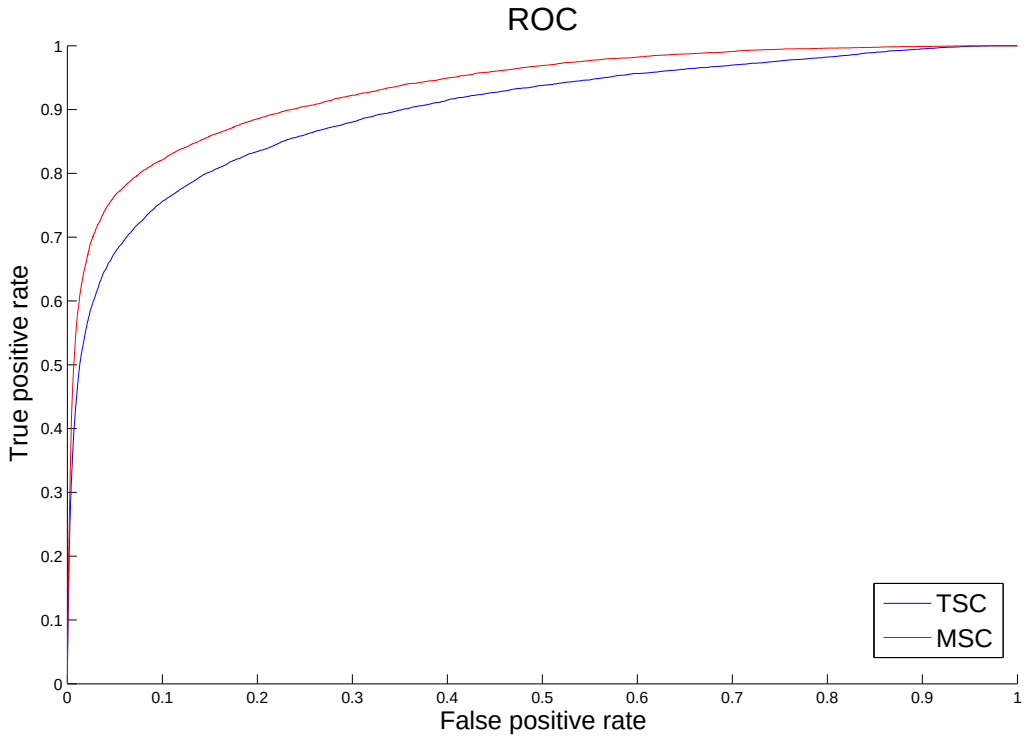


Figure 5.12: ROC-curves for the best performing logistic classifiers using TSC and MSC methods for feature extraction.

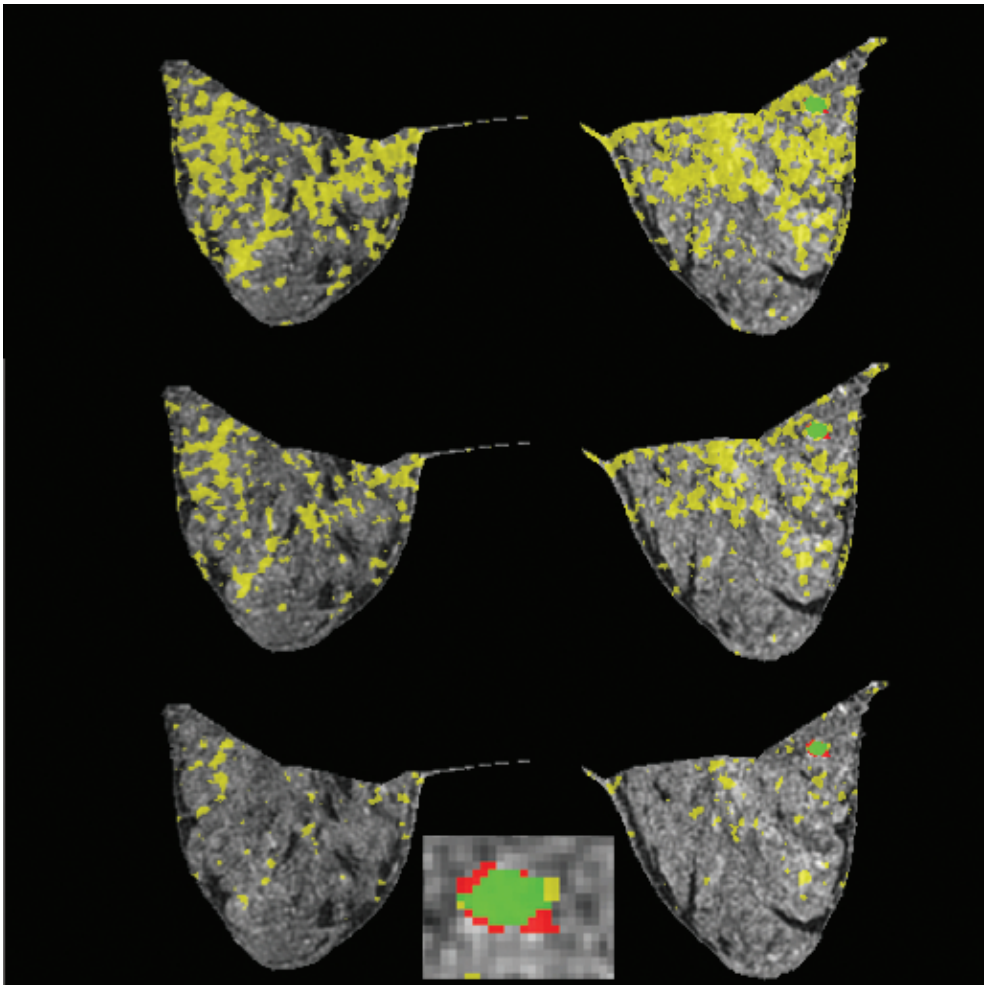


Figure 5.13: Example 1 of segmentation results from the TSC method at cut-point levels 0.7, 0.9 and 0.99, from top to bottom. The lesion segmentation is magnified at the bottom, at a cut-point level of 0.99 (TP = green, FP = yellow, FN = red, TN = greyscale).

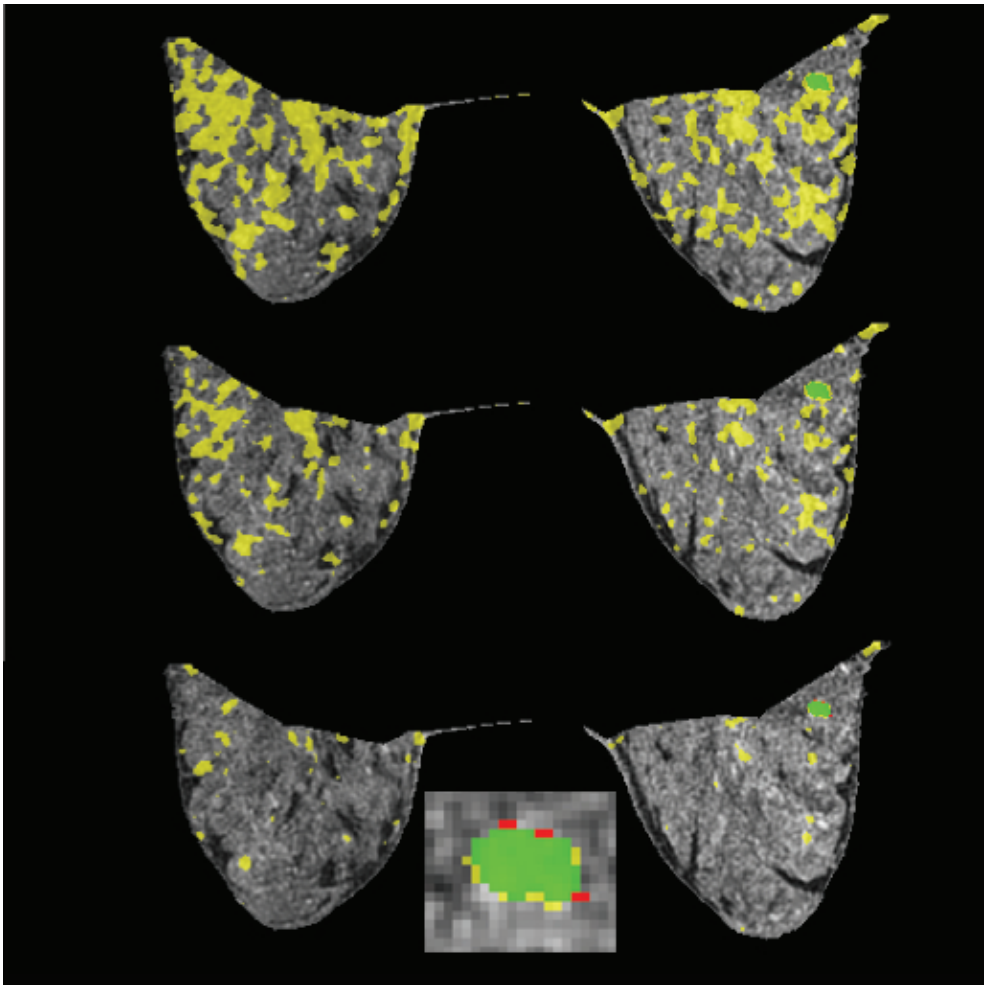


Figure 5.14: Example 1 of segmentation results from the MSC method at cut-point levels 0.7, 0.9 and 0.99, from top to bottom. The lesion segmentation is magnified at the bottom, at a cut-point level of 0.99 (TP = green, FP = yellow, FN = red, TN = greyscale).

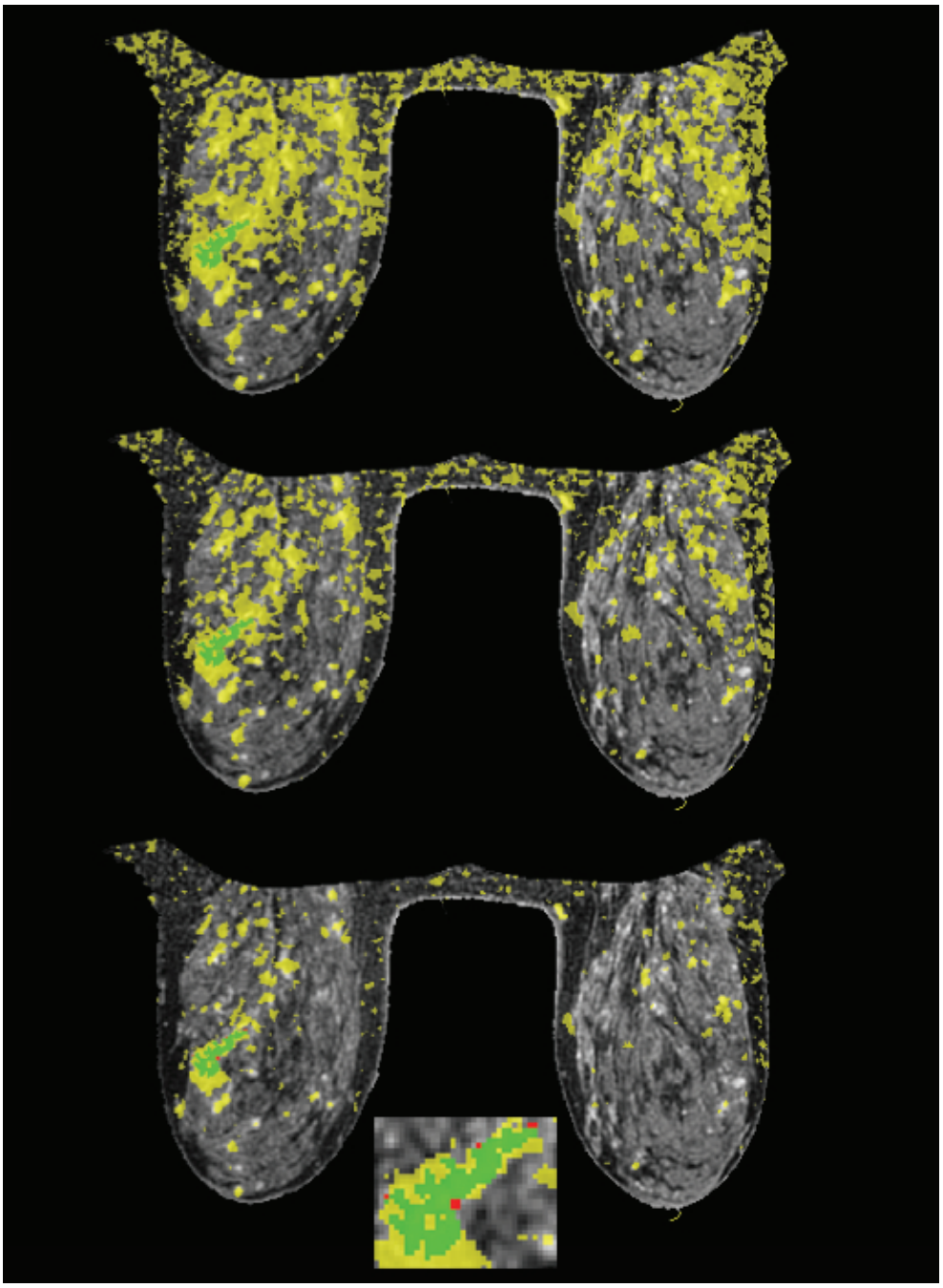


Figure 5.15: Example 2 of segmentation results from the TSC method at cut-point levels 0.7, 0.9 and 0.99, from top to bottom. The lesion segmentation is magnified at the bottom, at a cut-point level of 0.99 (TP = green, FP = yellow, FN = red, TN = greyscale).

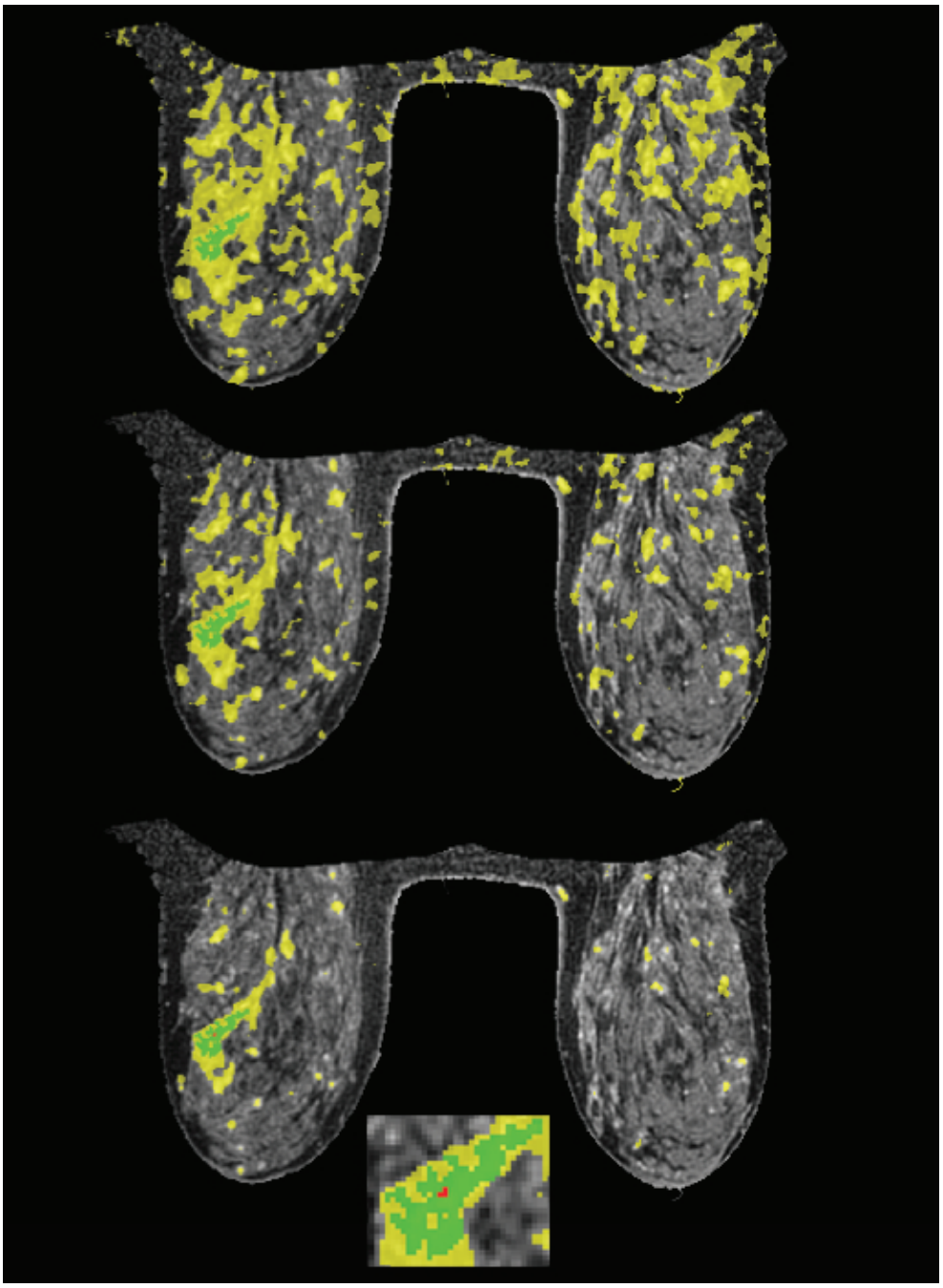


Figure 5.16: Example 2 of segmentation results from the MSC method at cut-point levels 0.7, 0.9 and 0.99, from top to bottom. The lesion segmentation is magnified at the bottom, at a cut-point level of 0.99 (TP = green, FP = yellow, FN = red, TN = greyscale).

5.11 Discussion

In this section the the results and findings are discussed, as well as the choices made in the empirical evaluation.

5.11.1 Results

Selection of features with stepwise linear regression for the k-NN classifier resulted only in a small reduction of features for the Hayton and Linear Slope model, as seen in Table 5.7. For the Ricker model, a subset including half of the features was selected. As the trivial features were defined in the exact same way for the Ricker and Hayton model, and the enhancement curves show similar behaviour, the cause of this difference is unclear.

The results from Table 5.8 show that the Mahalanobis distance was the best metric and that a low number of nearest neighbours, k, was favoured over a high. One reason for why the Mahalanobis distance was favoured over other distance metrics may be that it takes correlation between features into consideration.

From the results in Table 5.9, it is clear that the SFS method favoured the proposed spatio-temporal features, for the TSC models. The TSC models showed an overall higher AUC score than the MSC models. Table 5.9 and 5.10 show that larger subsets were selected by the SLR feature selection step than for SFS. However for the SFS method the number of features was chosen manually by finding at what number of features the performance did not improve. In the SLR method the number of features selected was governed by the BIC.

The grid searches for RBF kernel parameters show similar AUC results for both the TSC and MSC models, as observed in Table 5.11. A small tendency for the models including SLR features to perform better is observed. Figure 5.11 show an AUC-surface for the grid searches where the strongest influence on the AUC is governed by the γ-parameter. This behaviour is connected to the observed values in Table 5.11, where the γ-parameters tend to take values close to the upper search bound.

The results from using the TSC-classifiers, shown in Table 5.12, on the full validation set show that the Hayton models based on LR performed best overall, but the Linear Slope models was the most consistent across all models. An other observation is that the SVM classifiers with SLR features produced overall poorer results. This is in contrast to the model selection step, where these models performed best. This is perhaps indicative of these models being too tightly fitted to the training data.

In the comparison between the performance of the TSC- and the MSC-models on the reduced validation set, presented in 5.13, the Linear Slope model using MSC-features performs best overall and most consistently across all models. The Hayton model using MSC-features does however produce similar results. Using the SFS feature sets and the SVM as classifier, the Ricker model using TSC-features resulted in highest AUC, despite the lower flexibility of this enhancement model (because it has two rather than three parameters).

Observing the classification summaries in Table 5.14 and Table 5.15, both the best performing TSC- and MSC-classifier show similar values of sensitivity. The specificity is higher for lower cut-points for the TSC-model compared to the MSC-model, but lower for higher cut-points. As the reduced validation set contained significantly fewer observations of the "non-suspicious" class, the CCR closely follows the specificity. The AUC is a better measure of performance in the sense that it is independent of class prior probabilities. The ROC-curves for the best performing TSC- and MSC-classifiers in Figure 5.12 shows that for the MSC-classifier the curve is closer to the upper left corner, a behaviour which is reflected in the higher AUC.

Figures 5.13, 5.14, 5.15 and 5.16 show that the clustering of the segmented pixels seems to be more clearly grouped into well defined regions with the MSC method.

5.11.2 Pre-processing

The use of the 3×3 mean filter in order to mitigate noise is a simple and effective. Nevertheless extreme values within the filter window (due to noise) centred on a given voxel can have a great influence on the value assigned to that voxel and hence can impact local texture. Moreover because it is a low-pass filter high frequencies are attenuated which also has an impact on local texture. This may have had a negative impact on the discriminatory power of the co-occurrence features computed for both the TSC and MSC methods. A more advanced noise reduction method could have been used, e.g. the dynamic non-local means algorithm [55], but would have required significantly more computational time.

Hayton's algorithm for segmentation of the breast-air boundary in many cases yielded an over-segmentation. This is possibly due to MR intensity inhomgeneities due to the bias field. Consequently many noisy voxels will have been included in the training and test data likely leading to reduced performance. Better results may have been obtained if bias correction had been included in the preprocessing steps.

5.11.3 Partitioning into training and validation sets

Radiologists typically review tissue that exhibits 50% or higher relative enhancement. This threshold was applied to the suspicious voxels of the training set (see Section 5.3.1). A refinement to the selection of non-suspicious voxels for the training data set is to apply this same threshold for these voxels and also for all voxels of the validation set. The segmentation method would then only discriminate between suspicious and non-suspicious voxels that enhance 50% or more. This would not only reduce the amount of data to be processed but is likely to improve classification performance.

5.11.4 Feature extraction

The decision to include both malignant and benign lesions in a single "suspicious" class may have had a negative impact on classification performance for both methods. The reason for this is that a benign lesion will typically have a homogenous texture whilst a malignant lesion will have a heterogenous texture. In particular this may have affected the discriminatory power of the spatial co-occurrence features used in both methods. In the case of the TSC method this issue may have negatively influenced the performance of the k-NN classifiers. This is because some voxels in an enhancing malignant lesion may exhibit enhancement characteristic of "non-supicious" tissue and yet be assigned a class label of "suspicious". Kale et al. [48] considered only invasive ductal carcinoma tumours in their study and their reported results and segmentation examples have less false positive classification errors than for the data in this study. Thus the performances of both the TSC and MSC methods for segmenting malignant-only lesions remains an open question.

5.11.5 Fitting of parametric models of enhancement

Given that the Linear Slope model and the Ricker model were able to be fitted using LS, this could be done considerably quicker than for the Hayton model, which was fitted using NLS. On an Intel Core i5 (2.5GHz) processor, a single slice was processed in a matter of seconds for the Ricker model, minutes for the Linear Slope model and could range to well

over an hour for the Hayton model. As a typical volume consisted of 110-170 slices, this must be considered for most practical applications.

5.11.6 Temporal-score co-occurrence method

A k-NN classifier was used here because of its simplicity and because it does not make any assumptions about the underlying distribution of the data. The features used in the k-NN classifiers were not scaled which potentially means that the some features may have dominated feature space. Whether or not feature scaling would have influenced the overall performance of the TSC method thus remains an open question.

5.11.7 Multispectral co-occurrence method

The truncation of the enhancement curve parameter values was not discussed in the original paper of Kale et al. [48]. This is an important consideration in the binning of each parameter for the purposes of computing co-occurrence. Clearly the choice of range and bin-widths here can impact on the overall performance of the MSC method.

5.11.8 General comments

For the MSC method, the features extracted were the ones described in the original paper, albeit derived from different underlying parametric models of enhancement. The MSC method yielded a very sparse GLCM matrix with only 50 counts distributed over 32^3 elements. As a comparison the TSC method had a co-occurrence matrix consisting of 7^2 (Ricker) and 13^2 (Hayton, Linear Slope) elements. Each window yielded a total number of 288 counts. This implies that the sparsity of the MSC method's GLCM was significantly higher than that of the TSC. It is possible that this difference in sparsity of the GLCM may have impact on the discriminatory power of the co-occurrence features.

The choice of window size used in the MSC method represents a trade-off between the sparseness of the resulting co-occurrence matrix and the need of sufficient localisation [48]. A larger window size is beneficial in the sense that the co-occurrence matrix contains more counts. However the resulting textural features then come from a larger window and have less to do with the local texture. Kale et al. [48] used a window size of $5 \times 5 \times 2$ (i.e. 5 by 5 in-plane window over two slices). The proportions between in-plane and through-plane resolution of our data sets coincided with the proportions of that of Kale et al. such that the window size of $5 \times 5 \times 2$ voxels were the proportions closest to a cube shaped window. This window size also led to the same localisation and sparsity trade-off. The impact that this window size has on the performance of both methods remains an open question.

5.11.9 Feature selection

Only two different feature selection strategies were considered: step-wise logistic regression based on the BIC, and logistic regression with SFS. Logistic regression was chosen because it is a simple linear classifier that does not have any tuning parameters. A stop criterion or maximum number of features allowed, should have been defined in order to avoid the need of deciding what number of features to include from the method. As the Bayesian information criterion was used for the SLR method, small models were favoured, thus reducing the risk of overfitting.

5.11.10 Model selection

The grid search method for tuning the SVM kernel parameters requires user input (number of levels in the resolution pyramid) and is thus not objective. To get adequate resolution, only three grid searches had to be conducted but as this might not be the case, an automatic search method might be attractive in order to save time. As cross-validation was used in several steps, large parts of the model selection could have been done in a nested manner instead of a sequential manner.

5.11.11 Evaluating classifier performance

The ground truth data for this thesis was acquired by manual segmentation of an expert reader. Given that this data was segmented in a subjective fashion, there is a certain possibility of individual voxels being incorrectly segmented by the reader (at the very least at the border of lesions). Incorrectly labeled voxels suffer great risk of being incorrectly segmented by the proposed method since the temporal features extracted for these voxels will show non-lesion characteristics.

5.11.12 Trade-off between sensitivity and specificity

The resulting voxel-wise specificity and sensitivity is naturally dependent on the choice of cut-point for the final classifier. In general a high sensitivity is of importance for lesion segmentation applications. However, the importance of specificity may vary depending on the specific application. If the application aims at assisting a radiologist in terms of verifying suspicious lesions, high specificity is the first priority. If the application on the other hand is functioning as a pre-screener (e.g. for high risk women) a high sensitivity is necessary. Specificity increases with increasing thresholds as seen in Table 5.14 and visual examples of this can be seen in Figures 5.13, 5.14, 5.15 and 5.16. As the specificity of the voxel-wise segmentation increases, the sensitivity declines leading to a trade-off between a useful segmentation and the risk of missing lesion voxels. One way of handling the trade-off between sensitivity and specificity, may be to include a slider with which the radiologist can vary the threshold as preferred or to plot the score of the final classifier using a heat map with transparency on top of the original image.

6 Summary and Conclusions

This chapter briefly reviews the thesis, summarises the key contributions and findings, outlines the limitations of the research undertaken, and discusses opportunities for future research.

6.1 Thesis summary

Chapter 1 provided an introduction to the field of the research and a statement of the hypothesis underlying the research; i.e.

> *-that it is possible to automatically and accurately segment 3D lesions in DCE-MRI breast data by means of voxel-wise classification based on quantitative features that describe both spatial and temporal change in contrast enhancement.*

It was stated that the aim of this research was to explore the validity of this hypothesis by:

1. *Developing a spatio-temporal segmentation method for 3D breast lesions based on characterising the spatial co-occurrence of enhancement; and*
2. *Evaluating the performance of the new method using real clinical breast MRI data.*

Chapter 2 provided the necessary background and theory needed for the remainder of the thesis. This included material on breast cancer, magnetic resonance imaging (MRI), dynamic contrast-enhanced MRI, the analysis of contrast enhancement in DCE-MRI, statistical pattern recognition, and texture analysis by co-occurrence.

Chapter 3 presented a review of the literature dealing with the automatic segmentation of lesions in breast MRI. The review concluded that voxel-wise textural analysis by co-occurrence is one the best approaches for characterising the spatio-temporal enhancement of lesions and thus for automatically segmenting them. It was noted that a limitation of existing co-occurrence approaches is that it is necessary to truncate co-occurrence values in order to construct co-occurrence matrices. This decision is somewhat arbitrary. This then was the motivation for the new method proposed in Chapter 4.

Chapter 4 presented the proposed method based on first converting each voxel-wise time series to a single probability-like score (between 0 and 1) and then extracting co-occurrence features for each voxel in the resulting probability volume.

Chapter 5 presented an empirical evaluation of the performance of the proposed method relative to the competing method of Kale [48].

6.2 Key contributions and findings

This research proffers a new method (voxel-wise classifier) for automatically segmenting 3D lesions in breast DCE-MRI data suspicious for malignancy. It also provides additional evidence in support of the underlying hypothesis. In particular the empirical results demonstrate that it is possible to automatically differentiate between suspicious and non-suspicious tissue in the breast, by combining both spatial and temporal information using the notion of co-occurrence originating from classical texture analysis. The empirical results demonstrate that the proposed method achieves a level of performance comparable to the selected benchmark method (AUC of 0.8989 ± 0.0021 versus AUC of 0.9330 ± 0.0018). The advantage of the proposed method over the competing method is that it does not require subjective specification (truncation) of feature ranges for computing co-occurrence.

The empirical results also suggest that a simple empirical model of contrast-enhancement (linear-slope model) can be used in lieu of more complicated pharmacokinetically inspired models. The advantage of this model is that it can be fitted rapidly using simple least squares.

6.3 Opportunities for further research

The proposed segmentation method may potentially be improved by alternating several implementation steps. A weakness in the empirical evaluation of the proposed method is that it was used to discriminate suspicious and non-suspicious tissue. The suspicious class included both benign and malignant lesions. Given that benign lesions tend to be more homogeneous in enhancement than malignant this may have impacted on the quality of discrimination that could be obtained using the co-occurrence texture features. Future work could consider discriminating malignant from all other tissue.

The heterogeneity of malignant tissue also suggests that the generation of a single temporal score for each voxel may not be the best approach. An alternative approach would be to consider the use of regular co-occurrence features extracted for each model parameter volume separately.

Future work could also consider the combination of features from additional MR techniques such as diffusion weighted MRI and MR spectroscopy.

6.4 Limitations

In this project several limitations were encountered, these included:

- The empirical evaluation performed was limited to segmenting suspicious lesions, i.e. discriminating both benign and malignant tissue from all other tissue. It did not consider the problem of discriminating malignant tissue from all other tissue.
- In the empirical evaluation no bias field correction or motion correction was performed (although a subset of the data provided for this study had already been motion corrected).
- The " ground truth " segmentations used in the study originate from a single expert reader's manual segmentation. Ideally several independent experts should have segmented the same data to be able to account for intersubject variability.

References

[1] Monique D. Dorrius, Marijke C. Jansen van der Weide, Peter M.A. van Ooijen, Ruud M. Pijnappel, and Matthijs Oudkerk. Computer-aided detection in breast MRI: a systematic review and meta-analysis. *European Radiology*, 8:1600–1608, 2011.

[2] Robert M. Nishikawa. Current status and future directions of computer-aided diagnosis in mammography. *Computerized Medical Imaging and Graphics*, 31:224–235, 2007.

[3] S. Sinha and U. Sinha. Recent advances in breast MRI and MRS. *InterScience*, 1:3–16, 2009.

[4] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D.M. Parkin. Globocan 2008, cancer incidence and mortality worldwide: IARC cancerbase no. 10, 2010.

[5] Weijie Chen, Maryellen L. Giger, and Ulrich Bick. A fuzzy C-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Academic Radiology*, 13:63–72, 2006.

[6] S. Vinitha, E. Yin-Kwee, R. Acharya, and O. Faust. Breast imaging: A survey. *World Journal of Clinical Oncology*, 2:171–178, 2011.

[7] Yunfeng Cui, Yongqiang Tan, Binsheng Zhao, Laura Liberman, Rakesh Parbhu, Jennifer Kaplan, Maria Theodoulou, Clifford Hudis, and Lawrence H. Schwartz. Malignant lesion segmentation in contrast-enhanced breast MR images based on the marker-controlled watershed. *Med. Phys.*, 36(10):4359–4369, 2009.

[8] Thomas Bülow, Lina Arbash Meinel, Rafael Wiemker, Ursula Kose, Akiko Shimauchi, and Gillian Newstead. Segmentation of suspicious lesions in dynamic contrast-enhanced breast MR images. *Medical Imaging*, 6514:1076–1083, 2007.

[9] E. Warner, D.B. Plewes, R.S. Shumak, G.C. Catzevalos, L.S. di Prospero, M.J. Yaffe, V. Goel, E. Ramsay, P.L. Chart, D.E.C. Cole, G.A. Taylor, M. Cutrara, T.H. Samuels, J.P. Murphy, J.M. Murphy, and S.A. Narod. Comparison of breast magnetic resonance imaging, mammography, and ultrasound for surveillance of women comparison of breast magnetic resonance imaging, mammography, and ultrasound for surveillance of women at high risk for hereditary breast cancer. *Journal of Clinical Oncology*, 19(15):3524–3531, 2001.

[10] Elizabeth A. Morris and Laura Liberman. *Breast MRI: Diagnosis and intervention*. Springer, 2004.

[11] Christiane Katharina Kuhl, Peter Mielcareck, Sven Klaschik, Claudia Leutner, Eva Wardelmann, Jürgen Gieseke, and Hans H. Schild. Dynamic breast imaging: Are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211:101–110, 1999.

[12] American College of Radiology. *Breast Imaging Reporting and Data System (BI-RADS)*. American College of Radiology, Reston, VA, 1 edition, 2003.

[13] Anna Vignati, Valentina Giannini, Alberto Bert, Massimo Deluca, Lia Morra, Diego Persano, Laura Martincich, and Daniele Regge. A fully automatic lesion detection method for DCE-MRI fat-suppressed breast images. *Medical Imaging*, 7260:3146–3149, 2009.

[14] C. Boetes, R.D. Mus, and R. Holland. Breast tumors: Comparative accuracy of MR imaging relative to mammography and us for demonstrating extent. *Radiology*, 197(743):743–747, 1995.

[15] Peter L. Davis, Melinda J. Staiger, Kathleen B. Harris, Marie A. Ganott, Jolita Klementaviciene, Kenneth S. McCarty Jr., and Hector Tobon. Breast cancer measurements with magnetic resonance imaging, ultrasonography, and mammography. *Breast Cancer Research and Treatment*, 37:1–9, 1996.

[16] Hosein Rabiei, Ali Mahloojifar, and Michael E. Farmer. Providing context for tumor recognition using the wrapper framework. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1252 – 1255, 2007.

[17] Matthew Kupinski and Maryellen L. Giger. Automated seeded lesion segmentation on digital mammograms. *IEEE Transactions on Medical Imaging*, 17(4):510–517, 1998.

[18] Jr. David M. Catarious, Alan H. Baydush, and Jr. Carey E. Floyd. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system. *Med. Phys.*, 31(6):1512–1520, 2004.

[19] Yaniv Gal, Andrew Mehnert, Andrew Bradley, Kerry McMahon, and Stuart Crozier. Automatic segmentation of enhancing breast tissue in dynamic contrast-enhanced MR images. In *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on*, pages 124–129, 2007.

[20] IARC. *Pathology and genetics of tumours of the breast and female genital organs.* International Agency for Research on Cancer, 2003.

[21] American Cancer Society. Breast cancer overview. URL: http://www.cancer.org/Cancer/BreastCancer/OverviewGuide/breast-cancer-overview-what-is-breast-cancer (Acquired: 2011-10-07), 6 2011.

[22] Uwe Fischer and Ulrich Brinch. *Practical MR mammography.* Georg Thieme Verlag, 2004.

[23] Alan Jackson, David Buckley, and Geoffrey J.M. Parker. *Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Oncology.* Springer, 2005.

[24] Wei Huang, Paul R. Fisher, Khaldoon Dulaimh, Luminita A. Tudorica Brian O'Hea, and Terry M. Button. Detection of breast malignancy: Diagnostic MR protocol for improved specificity. *Radiology*, 232(585-591):585–591, 2004.

[25] Anwar R. Padhani. Dynamic contrast-enhanced MRI in clinical oncology: Current status and future directions. *Journal of Magnetic Resonance Imaging*, 16:407–422, 2002.

[26] Peter Gibbs and Lindsayy W. Turnbull. Textural analysis of contrast-enhanced MR images of the breast. *Magnetic Resonance in Medicine*, 50:92–98, 2003.

[27] A. Mehnert, M. Wildermoth, and S. Crozier. Two non-linear parametric models of contrast enhancement for DCE-MRI of the breast amenable to fitting using linear least squares. In *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, pages 611–616, 2010.

[28] Seymour Koenig, Marga Spiller, Rodney Brown, and Gerald Wolf. Relaxation of water protons in the intra- and extracellular regions of blood containing Gd(DTPA). *Magnetic Resonance in Medicine*, 3:791–795, 1986.

[29] G. Strich, P. Hagan, and K. Gerber. Tissue distribution and magnetic resonance spin-lattice relaxation effects of gadolinium-DTPA. *Radiology*, 154:723–726, March 1985.

[30] A. Mehnert. *Image Analysis for the Study of Chromatin Distribution in Cell Nuclei*. PhD thesis, The University of Queensland, 2003.

[31] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:4 – 37, 2000.

[32] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.

[33] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35:352–359, 2002.

[34] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[35] R.O. Duda, P.E. Hart, and D.G. Storck. *Pattern Classification*. John Wiley and Sons, New York, 2 edition, 2001.

[36] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[37] Thierry Denoeux. A k-nearest neighbor classification rule based on dempster-shafer theory. *Studies in Fuzziness and Soft Computing*, 25:804–813, 1995.

[38] R.R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[39] D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Propability and Statistics. Wiley and Sons, second edition, 2000.

[40] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, seventh printing edition, 1996.

[41] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University., 2003.

[42] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[43] Dmitriy Fradkin and Ilya Muchnik. Support vector machines for classification. In James Abello and Graham Cormode, editors, *Discrete Methods in Epidemiology*, volume 70, pages 13–20. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 2006.

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

[45] Johan Löfhede. *The EEG of the Neonatal Brain - Classification of Background Activity*. PhD thesis, Chalmers University of Technology, 2009.

[46] Yi-Wei Chen. Combining SVMs with various feature selection strategies. *Studies in Fuzziness and Soft Computing*, 207:312–324, 2006.

[47] R. M. Haralick, R. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. Syst., Man, Cybernitics*, SMC-3:610–621, 1973.

[48] Mehmet C. Kale, Bradley D. Clymer, Regina M. Koch, Johannes T. Heverhagen, Steffen Sammet, Robert Stevens, and Michael V. Knopp. Multispectral co-occurrence with three random variables in dynamic contrast enhanced magnetic resonance imaging of breast cancer. *IEEE Transactions on Medical Imaging*, 27(10):1425 – 1431, 2008.

[49] A. Meyer-Baese, O. Lange, T. Schlossbauer, and A. Wismuller. Computer-aided diagnosis and visualization based on clustering and independent component analysis for breast MRI. In *15th IEEE International Conference on Image Processing*, pages 3000 – 3003, 2008.

[50] Lina Arbash Meinel, Thomas Bülow, Dezheng Huo, Akiko Shimauchi, Ursula Kose, Johannes Buurman, and Gillian Newstead. Robust segmentation of mass-lesions in contrast-enhanced dynamic breast MR images. *Journal of magnetic resonance imaging*, 32:110–119, 2010.

[51] Brent J. Woods, Bradley D. Clymer, Tahsin Kurc, Johannes T. Heverhagen, Robert Stevens, Adem Orsdemir, Orhan Bulan, and Michael V. Knopp. Malignant-lesion segmentation using 4d co-occurrence texture analysis applied to dynamic contrast-enhanced magnetic resonance breast image data. *Journal of magnetic resonance imaging*, 25:495–501, 2007.

[52] Sang Ho Lee, Jong Hyo Kim, Jeong Seon Park, Jung Min Chang, Sang Joon Park, Yun Sub Jung, Sungho Tak, and Woo Kyung Moon. Texture analysis of lesion perfusion volumes in dynamic contrast-enhanced breast MRI. In *Biomedical Imaging: From Nano to Macro*, pages 1545–1548, 2008.

[53] P. Hayton. *Analysis of Contrast-Enhanced Breast MRI*. PhD thesis, University of Oxford, 1998.

[54] D. J. Hudson. Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association*, 61(316):1097–1129, 1966.

[55] Yaniv Gal, Andrew Mehnert, Andrew Bradley, Kerry McMahon, Dominic Kennedy, and Stuart Crozier. Denoising of dynamic contrast-enhanced MR images using dynamic nonlocal means. *IEEE Transactions on Medical Imaging*, 29:302–310, 2010.

Appendices

A Parametric models of contrast enhancement

A.1 Hayton model

The Hayton model is a two-compartment model based on a pharmacokinetic model for measurements on the blood-brain barrier using MR, initially proposed by Tofts and Kermode. Unlike the model of Toft and Kermode as well as other traditional pharmacokinetic models, the Hayton model does not attempt to assess enhancement pixels based on absolute enhancement characteristics. Rather than this the model aims at finding regions of tissue that enhances significantly more than healthy tissue, which mimics the way dynamic MR sequences are interpreted by radiologists [53].

Optimally the contrast agent is injected into the bloodstream instantaneously, this would give an exact starting time (t = 0) for the model. However in practice, due to the viscosity of the contrast agent, it is necessary to inject the agent under a short time period. Hayton's model is based on the assumption of instantaneous injection since this injection model has the fewest unknown parameters [53]. The resulting model function is

$$C(t) = \frac{A}{a-b}(e^{-bt} - e^{-at}), \tag{A.1}$$

where A, a, and b are parameters representing the original compartmental variables from the two-compartment model.

Since the model is the sum of two exponential terms, a non-linear least squares algorithm must be used to fit the parameters A, a and b. The model fitting algorithm requires initial parameter values and a limit on the number of iterations allowed. By its iterative nature, these fitting methods are computationally expensive [27]. Recommendations on starting values for these parameters are given by Hayton [53].

A.2 Linear Slope model

The linear slope model has its origins in plant and soil sciences, it's model function is

$$C(t) = \begin{cases} \beta_1 t & \text{if } t \leq \alpha \\ \beta_1 \alpha + \beta_2 (t-\alpha) & \text{if } t > \alpha, \end{cases} \tag{A.2}$$

where α, the abscissa, is the point in time where the two straight lines meet. β_1 and β_2 are the slope of the first and second straight line respectively. Since this is a linear model it can be fitted in a least squares sense.

A.3 Ricker model

The Ricker model comes from literature on fisheries [27], it has the form

$$C(t) = ate^{-bt}, \tag{A.3}$$

where the parameters a and b are real parameters. In order to fit the model in a least squares sense, the model can be rewritten into $y = \alpha t + \beta$ by taking the logarithm of both sides. Here $y = \log(C(t)) - \log t$, $\alpha = -b$ and $\beta = \log a$ [27].

B R program for Stepwise Logistic Regression

```
rm (list=ls())

library(MASS)

raw.data<-read.csv("Features_Labels.csv",header=FALSE)

# Create a model expression of the form V29~V1+V2+....+V27, depending on features
#features.to.keep<-seq(1,27)
#my.formula <- formula(paste("V29 ~",paste("V",features.to.keep,sep="", collapse=" + ")))

#my.glm <- glm(V29~1, family=binomial(link=logit), data=raw.data)
my.glm <- glm(my.formula, family=binomial(link=logit), data=raw.data)
my.step <- stepAIC(my.glm, scope=my.formula, k=log(nrow(raw.data)),direction = "both")
```

C Temporal kinetic features for the Linear Slope, Hayton and Ricker enhancement models

Table C.1: Temporal kinetic features extracted for the Linear Slope model.

Temporal Kinetic Features	**Description**
TL1. Peak value	The maximum value of the curve. Set to zero if the model did not peak within 6.5 minutes.
TL2. Last value of the curve	The last value of the curve.
TL3. Ratio between first and second slope	The first slope divided by the second slope, set to zero if the second slope had zero inclination.
TL4. The area under the model curve (Analytical)	The area under the model curve up to 6.5 minutes.
TL5. Initial Percentage Enhancement (Raw)	The first post-contrast value subtracted and divided by the pre-contrast value, $PE = \frac{I_2 - I_1}{I_1}$.
TL6. Signal enhancement ratio (Raw)	The ratio between the enhancement to first post-contrast value divided by the enhancement to the last, $SER = \frac{I_2 - I_1}{I_{last} - I_1}$. The last post-contrast within the time interval was used.
TL7. First model parameter	The slope from 0 minutes to the abscissa of the joint point.
TL8. Second model parameter	The slope from the abscissa of the joint point to 6.5 minutes.
TL9. Third model parameter	The abscissa of the joint point.

Table C.2: Temporal kinetic features extracted for Hayton and Ricker models.

Temporal Kinetic Features	**Description**
T1. Time to peak	The time for the curve to reach its maximum value, normalised over the time interval. Set to zero if the model did not peak within 6.5 minutes.
T2. Time to reach half of the peak value	The time for the curve to reach half of its maximum value, normalised over the time interval. Set to zero if the model did not peak within 6.5 minutes.
T3. Peak value	The maximum value of the curve. Set to zero if the model did not peak within 6.5 minutes.
T4. Last value of the curve	The last value of the curve.
T5. Maximum value of the derivative (Analytical)	Analytically derived maximum derivative. Set to zero if the value did not occur within 6.5 minutes.
T6. Mean wash-in slope	The value where the curve peaks divided by the time to peak.
T7. Mean wash-out slope	The value where the curve peaks divided by the time to the last curve value.
T8. Ratio between wash-in and wash-out slopes	Wash-in slope divided by wash-out slope.
T9. Numeric approximation of the integral	Numerical approximation of the integral by use of the trapezoidal method.
T10. Numeric approximation of the derivative	The maximum difference between two subsequent curve values divided by the time difference.
T11. Initial percentage enhancement (Raw)	The first post-contrast value subtracted and divided by the pre-contrast value, $PE = \frac{I_2 - I_1}{I_1}$.
T12. Signal enhancement ratio (Raw)	The ratio between the enhancement to first post-contrast value divided by the enhancement to the last, $SER = \frac{I_2 - I_1}{I_{last} - I_1}$. The last post-contrast within the time interval was used.
T13. First model parameter	First model parameter, for Hayton or Ricker model.
T14. Second model parameter	Second model parameter, for Hayton or Ricker model.
T15. Third model parameter	Third model parameter, for Hayton model only.