THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Multiple Instance Learning for Breast Cancer Magnetic Resonance Imaging

Fahira Afzal Maken

BE (Honours)

A thesis submitted for the degree of Master of Philosophy at

The University of Queensland in 2015

School of Information Technology and Electrical Engineering

Abstract

In this thesis we evaluate the efficacy of multiple instance learning (MIL) as a ‘pure’ machine learning approach for the diagnosis of breast cancer in magnetic resonance images (MRI). The traditional approach for the diagnosis of breast cancer is based on region-of-interest (ROI) based single instance learning (SIL). In the ROI-based SIL, the classification of benign and malignant lesions depends on the features, which are extracted from segmented ROIs. But, an accurate segmentation of a ROI is a challenging task due to poor signal-to-noise-ratio and faint edges due to partial volume effects. Therefore, variations in the segmentation of ROIs can affect the diagnostic outcome. MIL is a relatively new method in the supervised learning, where each sample is represented as a bag of instances. An image in the context of MIL can be considered as a bag of pixels, tiles, or ROIs, which correspond to instances. Here we apply tile-based MIL, where the diagnosis of breast cancer is based on tile-based features which do not require the segmentation of ROIs. Therefore, tile-based MIL has the potential to provide the classification of breast cancer without segmenting the ROIs. This is the motivation for the main objective of this thesis: to estimate the efficacy of tile-based MIL in the detection and the diagnosis of breast cancer MRI.

We initially evaluate the potential of MIL for the detection and the diagnosis of breast cancer in anatomical T2-weighted MRI. In particular, we compare the performance of a MIL-based learner, i.e., citation-kNN (CkNN) against conventional kNN and a Random Forest classifier. We utilise both (generic) tile-based spatial features and (domain specific) ROI-based features. We perform experiments on two datasets consisting of 77 mass-like lesions and 129 both mass-like and non-mass-like lesions. The performance of CkNN as both a diagnostic and screening tool is evaluated using the area under the receiver operating characteristic curve (AUC), estimated over 10-fold cross validation. Results demonstrate that the tile-based CkNN has equivalent performance to the ROI-based classification. However, tile-based MIL has an advantage that it does not require the domain specific ROI-based features typically used in breast MRI. This not only has the potential to make the tile-based classification robust to inaccuracies in the delineation of suspicious lesions, but also makes it suitable for the detection of suspicious lesions prior to segmentation.

Next, we investigate the performance of CkNN for the diagnosis of breast cancer using dynamic MRI. Specifically we use generic tile-based spatio-temporal features derived from T2-weighted MRI and T1-weighted dynamic contrast enhanced MRI. We utilise a discrete cosine transform and contrast enhancement models as feature extraction techniques. We compare the performance of CkNN and kNN against a traditional approach based on bespoke ROI-based features using the 77 mass-like lesions. Empirical results show the equivalent classification performance of both tile-based and ROI-based features. We also develop a MIL-based feature selection technique. We emphasise the importance of MIL-based feature selection criterion for MIL-based classification on an experimental basis. Further, we highlight that generic tile-based spatio-temporal features have improved potential to discriminate benign and malignant lesions as compared to tile-based spatial features.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

1. Fahira A. Maken, Andrew P. Bradley, Yaniv Gal, and Darryl Mcclymont: Multiple Instance Learning for Breast Cancer Magnetic Resonance Imaging, in Proceedings Digital Image Computing: Techniques and Applications (DICTA) 2014, Wollongong, Australia.

2. Fahira Afzal Maken, Andrew P. Bradley: Multiple Instance Learning for Breast MRI Based on Generic Spatio-Temporal Features, International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Brisbane, Australia.

Publications included in this thesis

Refereed conference papers

1. Fahira A. Maken, Andrew P. Bradley, Yaniv Gal, and Darryl Mcclymont: Multiple Instance Learning for Breast Cancer Magnetic Resonance Imaging, in Proceedings Digital Image Computing: Techniques and Applications (DICTA) 2014, Wollongong, Australia. Incorporated as Chapter 3.

Category	Fahira A. Maken	Andrew P. Bradley	Yaniv Gal	Darryl McClymont
Analysis and interpretation of data	70 %	15 %	10 %	5 %
Conception and design	70 %	15 %	10 %	5 %
Drafting and writing	75 %	15 %	10 %	-

2. Fahira Afzal Maken, Andrew P. Bradley: Multiple Instance Learning for Breast MRI Based on Generic Spatio-Temporal Features, International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Brisbane, Australia. Incorporated as Chapter 4.

Category	Fahira A. Maken	Andrew P. Bradley
Analysis and interpretation of data	80 %	20 %
Conception and design	80 %	20 %
Drafting and writing	80 %	20 %

Contributions by others to the thesis

No contribution by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

First, I wish to offer my profound gratitude to my principal advisor, Prof. Andrew P. Bradley, for his continuous support, motivation, and immense knowledge. I thank him for critically proofreading this thesis and papers with patience and providing intellectual comments. I sincerely appreciate his guidance throughout this candidature.

I thank my associate advisors A/Prof Marcus Gallagher and Dr. Yaniv Gal. I especially thank to Dr. Yaniv Gal for his worthy pieces of advice, support, and encouragement at every step in the first eight months of the candidature.

I would like to thank Dr. Darryl McClymont for sharing his work and knowledge with me.

Last, but by no means least, I thank my family, especially my husband, parents, son and daughter for their loving support, prayers, understanding and encouragement. Thank you to my husband, Saad, for proof-reading this thesis.

Keywords

multiple instance learning, breast MRI, feature extraction, DCT, feature selection, classification, T2-weighted, DCE-MRI, tile-based spatio-temporal features.

Australian and New Zealand Standard Research Classifications (ANZSRC)

080109 (Pattern Recognition and Data Mining), 50%

080106 (Image Processing), 50%

Fields of Research (FoR) Classification

FoR code: 0903, Biomedical Engineering, 50%

FoR code: 0801, Artificial Intelligence and Image Processing, 50%

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

2D or 3D	Two dimensional or Three dimensional
ACR	American College of Radiology
APR	Axis parallel rectangle
AUC	Area under receiver operator characteristic curve
BIRADS	Breast Imaging Reporting and Data System
CAD	Computer aided diagnosis (system)
CADe	Computer aided detection
CADx	Computer aided diagnosis
CkNN	Citation *k*-nearest neighbours
CV	Cross validation
DCE	Dynamic contrast enhanced
DCT	Discrete cosine transform
EM-DD	Expectation maximisation diverse density
kNN	*k*-nearest neighbours
MI-SVM	Multiple instance support vector machines
MIL	Multiple instance learning
MRI	Magnetic resonance imaging
PP	Percentage positive
RF	Random forests
ROC	Receiver operator characteristic
ROI	Region of interest
SBF	Sequential backward selection
SFS	Sequential forward selection
SIL	Single instance learning
VOI	Volume of interest

Chapter 1

Introduction

1.1. Breast Cancer

Breast cancer is the most frequently diagnosed cancer in Australian women, which accounts for 15.5 percent of all cancer related deaths in women [1]. Advancements in breast cancer screening have resulted in increasing survival rate among the breast cancer patients [2]. X-ray mammography and ultrasound are commonly used screening methods for breast cancer. However, mammography has poor sensitivity to small cancers. Moreover, it is less effective particularly in dense or augmented breasts. Similarly, poor specificity, and the difficulty of imaging the whole breasts are major shortcomings of ultrasound. MRI has emerged as a promising adjunct imaging modality with the additional benefits of good soft tissue resolution, no ionizing radiation, and improved sensitivity [3]. Since MRI scans do not involve ionization radiation, the use of MRI is well suitable for patients with implants. The potential use of MRI is in the quantification of tumour volume, and multi-focal and multi-centric disease detection [4].

1.2. Breast MRI

Breast MRI scans typically involve the acquisition of anatomical T1-weighted (-w), T2-w MRI and dynamic contrast enhanced MRI (DCE-MRI). In T1-w MRI, fat is the most discernible feature which appears as a region of high image intensity while most of the other breast tissue types depict low to medium image intensity. Due to poor contrast, non-fatty breast tissues particularly carcinomas are indistinguishable from normal breast tissues. On the other hand, although T2-w MRI is more time consuming as compared to T1-w MRI, it provides improved contrast among non-fatty breast tissue [5]. Therefore, the use of T2-w MRI is standard in the clinical interpretation of breast cancer MRI. However, the detection of malignant lesions solely based on spatial signal intensities in the non-contrast enhanced sequences (i.e., T1-w and T2-w MRI) is a challenging task [6,7].

The sensitivity of MRI has greatly been improved with DCE-MRI. In DCE-MRI, a gadolinium based contrast agent is injected into the blood stream which then accumulates in the tissues with high blood flow. Three dimensional T1-w images are obtained both before (pre-contrast) and at several time points after (post-contrast) the injection of the contrast agent. Subtraction images of pre- and post-contrast images are then examined to locate the suspicious (enhancing) regions in the image. These subtraction images particularly highlight the change in signal intensity caused by the uptake of contrast agent. Figure 1.1 shows the pre-, first post-contrast DCE-MRI, and subtraction image. We can clearly see the enhancing lesion in post-contrast image.

Concentration of contrast agent in the tissue depends on the vascularisation, permeability of vessels, and interstitial space of the tissue [8]. Malignant lesions have the ability to develop new blood vessels to get more blood which help them to grow. Increased vascularisation and permeability of malignant lesions make them well distinguishable from the normal tissues in post-contrast images of DCE-MRI. However, some benign lesions also show enhancement in post contrast images [9]. Therefore, it is not a challenging task to detect malignant lesions in DCE-MRI, but to discriminate benign from malignant lesions (low specificity).

One of the important ways to distinguish benign and malignant lesions is to evaluate spatial features along with temporal features (i.e., spatio-temporal) [9,10]. Moreover, there is evidence to prove that the specificity of breast MRI can be improved with the use of un-enhanced T2-w MRI and DCE-MRI [11].

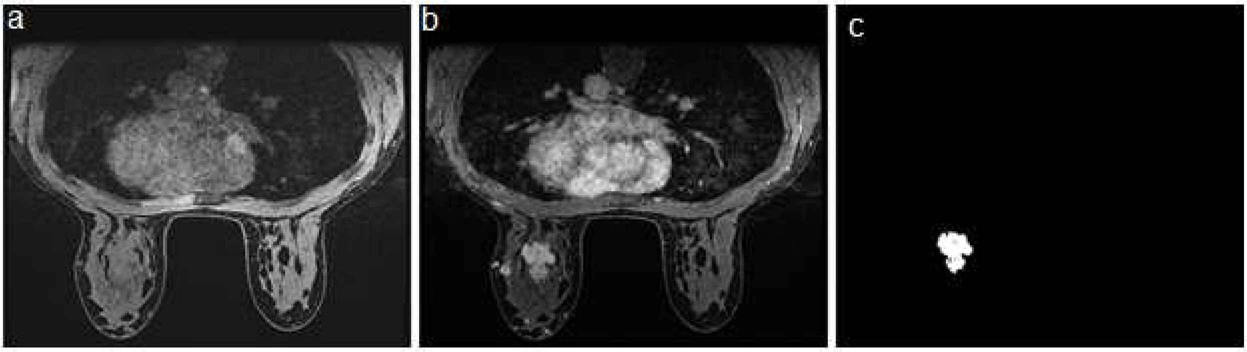


Figure 1.1: DCE-MRI, (a) Pre-contrast image, (b) First post-contrast image, (c) Subtraction image of pre- and post-contrast.

Both T1-w and T2-w images are acquired as three dimensional (3D) spatial data. DCE-MRI is acquired as four dimensional data, which comprise of high resolution stacks at several time points. Each stack typically contains more than a hundred 2D-slice images. Integrating and interpreting all the information from multi-temporal image sequence provided by a DCE-MRI scan is a time consuming, labour intensive and subjective task. Moreover, it is challenging for a human brain to identify subtle differences in the images during a repetitive task. This increases the chances of human error. As a result, lesion detection, the selection of regions-of-interest (ROI), and qualitative assessment of lesion features and margins are subjected to inter-observer variations [12,13]. Further, there is no ‘global standard’ available for the acquisition and interpretation of breast MRI [8,14]. Thus it is difficult to make a definitive statement about the manual interpretation of breast cancer.

1.3. Computer Aided Diagnosis Systems

Lack of standardized interpretation guidelines and both time intensive and subjective interpretation methods limit the benefits of DCE-MRI in routine clinical practice. Therefore, computer-aided diagnosis (CAD) systems have been developed to assist radiologists by providing objective indices of malignancy [15]. CAD systems are software programs that provide various medical image analysis techniques to help radiologists in the efficient interpretation of medical image data. It is important to note that CAD systems provide automated analysis of medical images which is used as a second opinion by radiologists in making diagnostic decisions [16]. Please refer to [16] to find a review of the current status of breast image analysis methods for risk assessment, detection, diagnosis, and treatment of cancer.

The most commonly used commercial breast CAD systems are CADstream (Merge Healthcare, Chicago, IL), DynaCAD (Invivo (Phillips), Gainsville, FL), Aegis Breast (Hologic, Bedford, MA). These software packages can compute time intensity curves, colour coded time intensity curve maps, maximum intensity projections, image subtraction, motion correction, interactive lesion segmentation etc. The comparison of CADstream and dynaCAD can be found in [17].

In the current CAD systems, discrimination between benign and malignant breast lesions is often done using ROI-based methods. Here, the dataset consists of bespoke (i.e., domain specific) features, extracted from each detected and then segmented lesion. In this way, each lesion becomes a labelled instance in the dataset. The feature vector extracted from each lesion is then individually labelled as either benign (negative) or malignant (positive). The features used in traditional single instance learning (SIL) approaches are based on the intensity, texture and morphology of the segmented lesion [11]. However, lesion margins and shape are strongly dependent on an accurate segmentation, which is a challenging task due to poor signal-to-noise-ratio and faint edges due to partial volume effects. Therefore, lesion delineation is affected by variation or uncertainties in the (semi)-automated lesion segmentation process. Clearly, these variations have the potential to lead to variations in the diagnostic outcome. Figure 1.2 illustrates the steps involved in the conventional CAD systems.

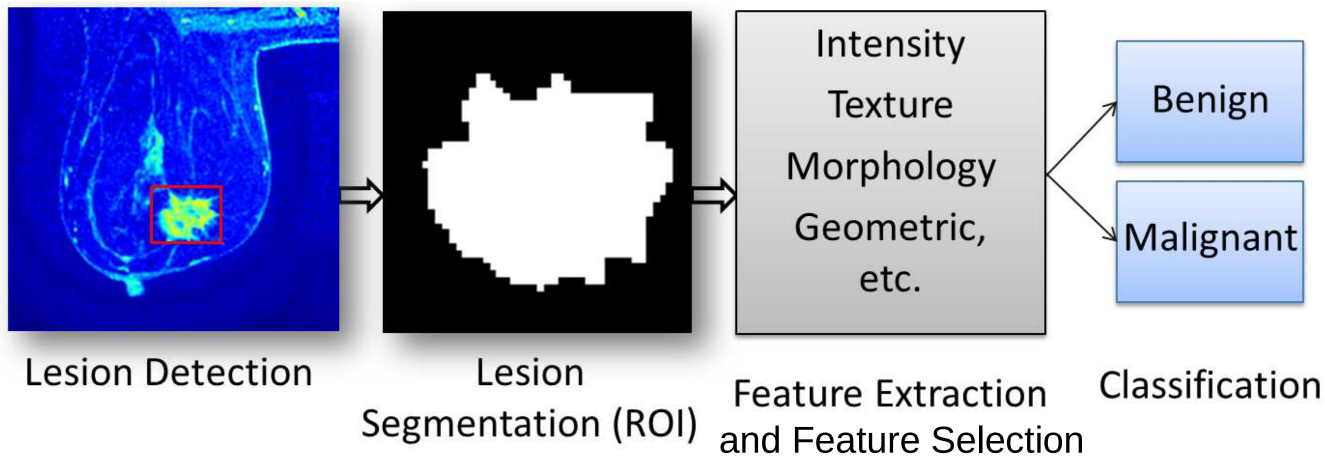


Figure 1.2: Conventional CAD system.

Moreover, in the conventional ROI-based approach, we need prior knowledge about the usefulness of ROI-based features which possibly co-relates with the physiology of the lesion. The importance of ROI-based features may be dependent on the image quality of the image acquisition protocol [16]. We know that the standard protocols for breast MRI acquisition keep on changing over the time to incorporate the latest advances in the hardware and research or to meet the requirement of radiologists [16]. With every change of the protocol, the efficacy of ROI-based features with regard to their diagnostic potential is changed. Therefore, we need the expert's knowledge to analyze the MRI for differentiating benign and malignant lesions. Then based on expert's advice, CAD systems are adapted only for the specific useful (ROI-based) features for the latest acquisition protocol. Thus, there is always an expert in the loop to design the features for the conventional CAD systems.

Multiple instance learning (MIL) [18] is a relatively new paradigm in supervised learning, which appears to be suitable for many CAD related problems, particularly when there is uncertainty regarding the class label given to individual instances. Multiple instance learning is a semi-supervised approach where each labelled sample is represented as a set (or 'bag') of instances. The objective of MIL is then to classify the bag of instances rather than the individual instances. In the context of MIL in image analysis, a bag is a sub-image consisting of multiple instances, where those instances are either individual pixels, square tiles (tile-based MIL) or arbitrary regions of interest

(ROI-based MIL) [19]. Figure 1.3 represents the concept of bag and instances in a tile-based MIL. In a tile-based approach, the features are generic in nature rather than specific to breast cancer. Since these features are extracted from small tiles, not segmented ROIs, classification performance is not affected by the accuracy of the segmented regions. This makes tile-based MIL suitable for both diagnostic applications, which classify already detected lesions, and screening applications, which initially detect suspicious lesions. On the other hand, ROI-based features cannot be used for screening because, segmentation of ROIs itself requires prior detection of lesions. These potential applications of tile-based MIL motivate us to focus our research on the evaluation of efficacy of tile-based multiple instance learning in the classification of breast cancer MRI.

Another advantage of using tile-based MIL is that we do not need prior knowledge in the specific domain for the classification of medical imaging. The reason is that, tile-based MIL is based on generic features which are selected on the basis of their discrimination, rather than prior knowledge. Therefore, the methodology and features remain the same even if MRI acquisition protocol changes.

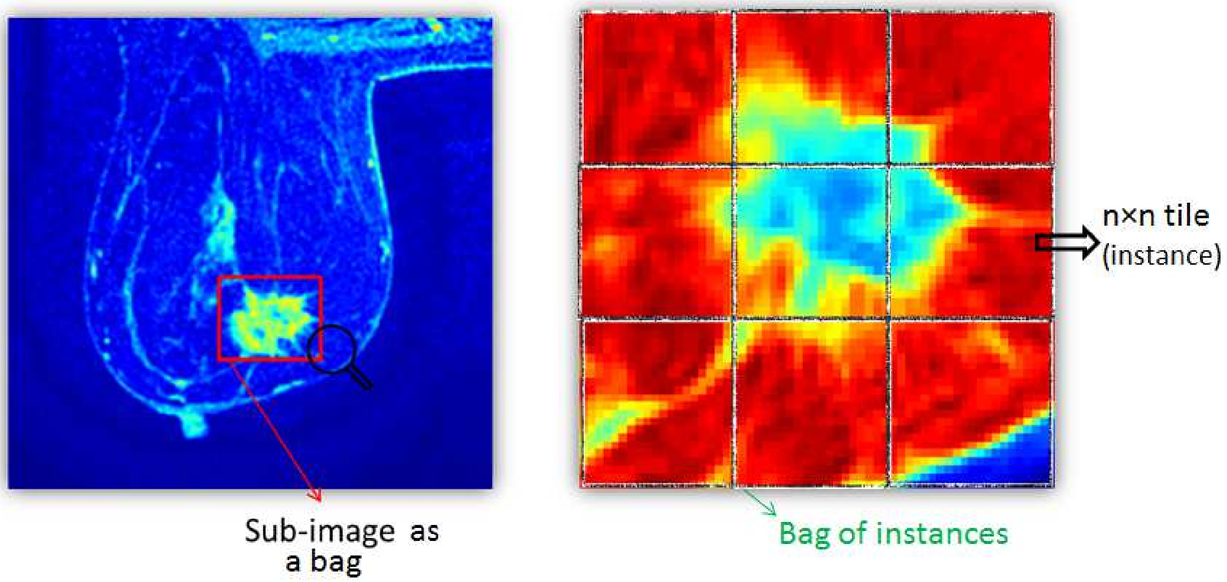


Figure 1.3: Tile-based multiple instance learning.

1.4. Aims and Objectives

The main focus of this research is to investigate the performance of multiple instance learning as a ‘pure’ machine learning method for the classification of breast MRI. In the tile-based MIL, we skip the segmentation step in the CAD and consider a sub-image containing the detected lesion as a bag. Classification of the bag (sub-image containing a detected lesion) is based on tile-based generic features rather than ROI-based domain specific features. This implies that classification of breast cancer MRI based on generic tile-based features would not be affected by inter-observer/ auto-segmentation variations in the lesion segmentation. Moreover, with any change of MRI acquisition protocol, the MIL-based CAD remains un-changed. Based on the potential applications of generic tile-based MIL in CAD, the following are the objectives of this research.

1. To study the performance of multiple instance learning using synthetic datasets.
2. To investigate the performance of multiple instance learning in the identification and classification of breast MRI using generic spatial features from T2-w MRI and to compare the classification performance with the conventional approach (ROI-based single instance learning).
3. To evaluate the efficacy of multiple instance learning for the classification of breast MRI using generic spatio-temporal features (from DCE-MRI and T2-w MRI) and to develop multiple instance learning-based feature selection algorithm.

1.5. Scope

The purpose of this thesis is not to solve breast cancer MRI classification problem, but to evaluate the efficacy of MIL as a ‘pure’ machine learning approach for the diagnosis of breast cancer. When we say ‘pure’ we mean that we use MIL as a generic approach without knowing much about the physiology and domain specific features typically used in the breast cancer MRI. Rather, we use MIL for classification of breast MRI in the same way as it is applied to solve an arbitrary image classification problem. In other words, we utilize generic features based on their level of discrimination as opposed to (bespoke) application specific features selected on the basis of prior knowledge.

Further to this, segmentation of lungs, heart, or breast boundary is not part of this research. We use sub-images that contain manually segmented ROIs. Additionally, we use already registered and bias field corrected images. Moreover, this research is not focussed to develop a new algorithm in the MIL domain. We use either existing tools or modify them to suit our requirement. Further, we estimate the MIL performance for the classification of only breast MRI data. Investigating the MIL for other medical imaging datasets is not part of this work.

1.6. Overview of Thesis

In this chapter, we have introduced the problems associated with the conventional breast MRI CAD systems. Also, we have talked about how multiple instance learning addresses these problems. This thesis begins with the understanding of concepts related to multiple instance learning on an experimental basis (section 3.2). This section provides necessary foundation to use MIL as a classification tool in the later work. In particular, we address our first aim of the research in the beginning of Chapter 3. In the rest of the Chapter 3 and Chapter 4, we address the main aim of the thesis, “to evaluate the efficacy of MIL in the diagnosis of breast cancer MRI”.

In Chapter 2, we introduce the necessary concepts in machine learning which we utilise in the experiments in the subsequent chapters. We begin this chapter with the introduction of MIL-based pattern recognition, followed by the review of MIL and SIL-based learners. Next, we provide reasoning for the experimental methodology which we use to address the aims of this research. Finally, we present the breast MRI data which we utilise in the experiments in the third and fourth chapters.

In Chapter 3, we evaluate the suitability of MIL for both the screening and diagnosis of breast cancer in T2-w MRI. We compare the performance of a MIL-based learner against SIL-based learners. We utilise both (generic) tile-based features and (domain specific) ROI-based features. We perform experiments on two datasets consisting of 77 mass-like lesions and 129 both mass-like and non-mass-like lesions.

In Chapter 4, we investigate MIL using generic tile-based spatio-temporal features for the classification of benign and malignant lesions. Specifically, we use T2-weighted MRI and T1-w DCE-MRI. We utilize parametric models and a discrete cosine transform as feature extraction techniques. In particular, we compare the performance of MIL against a traditional approach based on bespoke features extracted from a segmented ROI. We also develop a MIL-based feature selection technique and present the performance of learners with MIL-based feature selection in comparison to SIL-based feature selection.

Chapter 2

Background

In Chapter 1, we have presented the motivation and aims of this research. In this chapter, we provide the necessary concepts in MIL-based pattern recognition which are utilised in the later chapters. In section 2.1, we describe multiple instance learning particularly in the context of image analysis. Next in sections 2.2 to 2.4, we present the reasoning for the selected learners, feature selection, and the choice of performance measure respectively. In section 2.2, we also provide motivation for our first aim of the research. We present the breast MRI data in section 2.5 which we use in Chapters 3 and 4. In this section, we specifically present the imaging protocols of breast MRI data and provide motivation for our second and third aims of the research. Finally in section 2.6, we provide summary of the chapter.

2.1. Multiple Instance Learning

Multiple instance learning (MIL) [18] is a variant of traditional supervised machine learning. MIL differs from supervised learning (i.e., single instance learning) in the nature of learning examples. In single instance learning (SIL), each learning example is represented as an instance of fixed length feature vector. Each instance in supervised learning is associated with a class label. The task of SIL is to learn the model which predicts the class label of unknown instances. In multiple instance learning, a learning example is represented as a set (or ‘bag’) of instances. In other words, a MIL-based learning example comprises of a set of feature vectors. In MIL, each bag is associated with a class label while the instances themselves are not explicitly labelled. In contrast to SIL, MIL aims to find the class label of entire ‘bag’ of instances rather than labelling the individual instances. Therefore, MIL is a type of semi-supervised technique where we know the class label of a set of instances but individual label of each instance is unknown. This makes MIL suitable for semi-labelled (imaging) data, where the class labels of whole images are known but individual labels of instances (i.e., segments of an image) are unknown. According to the *standard* (asymmetric) *MIL assumption*, a bag is labelled positive if at least one instance in the bag is positive, otherwise the bag is negative [18]. The definition of a positive bag in MIL is equivalent to the abnormal breast, which may contain both normal and malignant lesions. Typical applications of MIL include natural scene image classification, web index page recommendation [20], text categorisation [21,22], and stock market prediction [19,23].

To illuminate the concept of multiple instance learning, we present an example of *simple jailer problem* [24]. Suppose there is a locked door. To open this door, we have *N* bunches (bags) of keys (instances). Unlike conventional single instance learning, MIL considers a bunch of keys ‘positive’ if it contains at least one key that can unlock the door. Here learning problem is to build a classifier which can find a ‘positive’ bunch of keys (bag). If we know that a bunch of keys is positive, then we can deduce that there must be at least single key in that bunch which is positive. However, we do not have complete information about which key is positive or how many keys are positive in a positive bunch. The unavailability of labels of individual keys (instances) in a bunch (bag) makes MIL a semi-supervised technique. On the other hand, in conventional SIL, learning aim is to find ‘positive’ keys rather than bunch of keys. Using simple jailer problem, we demonstrate the difference between SIL and MIL-based classification in Figure 2.1.

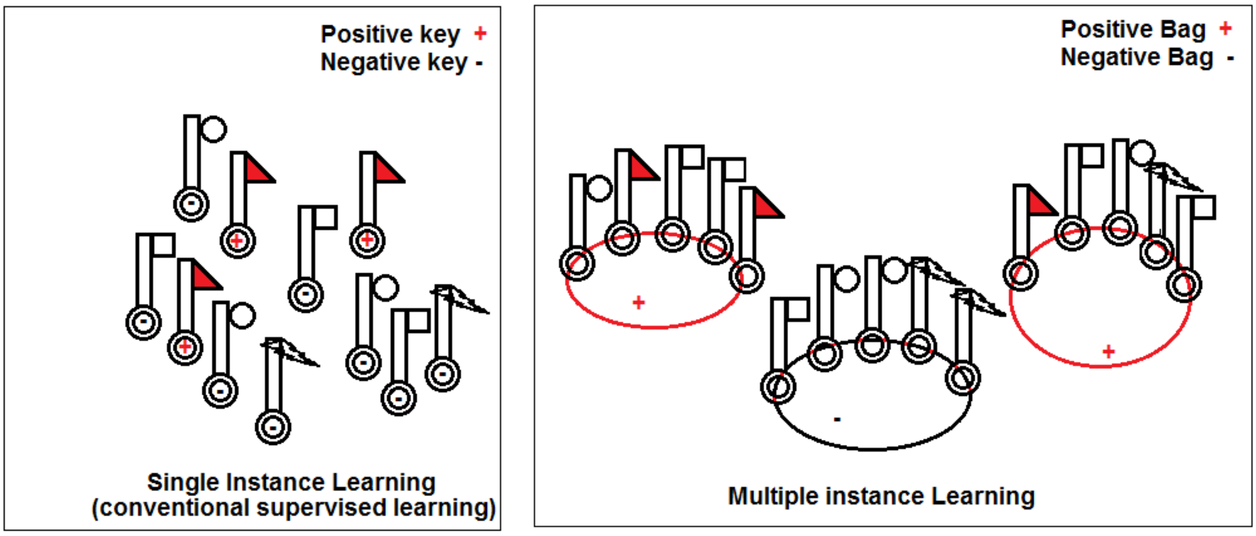


Figure 2.1: Difference between MIL and SIL-based classification.

There are various problems in pattern recognition which are difficult to solve by traditional supervised learning. Consider the example of *drug activity prediction* problem. In the drug activity problem, different conformations (structural variations) of the same molecule exist in nature. There is only one conformation which binds to the target receptor, making it ‘musk’ type [18]. However, for existing molecules, it remains unknown that which specific conformation of a molecule binds to the receptor. Therefore, in learning phase we only know the label of a molecule, while the labels of different conformations of molecule are unknown. In this case, single feature vector may not be sufficient to describe the problem [25]. In other words, traditional supervised methods may not be suitable to solve this problem. This problem can be solved using multiple instance learning, where different conformations of a molecule are considered as a bag and different conformations correspond to instances. According to the standard MIL assumption, a test molecule (bag) is classified as positive if any of its conformations emits musky smell, otherwise it is labelled negative [18,26].

Although, the standard MIL assumption works successfully for the drug activity prediction problem, there are several computer vision tasks including object recognition [21,27], image categorization and content-based image retrieval, which cannot adopt MIL under standard MIL assumption [19]. These tasks learn visual based concepts from labelled image databases, where few parts of the image (i.e., parts containing target object) are required to derive the label of an image. Since, the visual concept (target object) only occupies the small part of the image space, it is reasonable to decompose the image into smaller fragments [28]. Feature vectors from these fragments then become the instances in the bag. These fragments can be pixels, square tiles (tile-based MIL) or arbitrary regions of interest (ROI-based MIL). Since an image is considered as a bag and its segments are considered as instances, the problem of classifying image on the basis of presence of the target concept is a multiple instance learning problem. Thus MIL has vast applications in the problems based on image analysis.

The standard MIL assumption may be applied to the visual concept learning tasks with some success [19]. For example, in the classification of natural scenes of waterfalls [29], a natural waterfall image is classified as ‘positive’ if it contains at least one segment in the image that contains waterfall, otherwise it is negative. However, there are many tasks where standard MIL assumption is not an appropriate choice. Therefore, more general formulations of standard MIL assumption have been proposed for solving those MIL problems which don’t obey strict MIL assumption. For instance, consider a classification problem of natural scene images into *beaches, oceans and deserts* [19]. A *beach image* contains regions of both sand and water. Therefore, an image is classified as a beach image if and only if it contains segments of both sand and water. On the other hand, if it contains segments of sand or water only, it may be classified as a *desert image* or an *ocean image* respectively. This is a type of image classification task where MIL labels the bag on the basis of presence of co-occurrence of certain visual concepts. This alternative MIL assumption is called *presence based MIL assumption* [30].

Another MIL assumption is that a bag is composed of homogeneous group of instances [31,32]. This implies that all instances in a bag belong to the same class. This assumption is utilised in the classification of group of cells in pathology slides where a bag (group) label shows healthy or abnormal slide [32]. In cytological and histological analysis, the specimen (slide) contains a large number of microscopic cells. In the context of traditional supervised learning, each cell is considered as an instance. Since, it is usual practice to label the individual cells manually in the training phase, labelling the individual cells is not only time consuming but also difficult and costly. Therefore, it is reasonable to consider the slide as a homogeneous group of instances. In this case, we need a single label for whole group of cells instead of individual cells. For other generalised MIL assumptions, please refer to [19].

Although MIL has many applications in computer vision problems, and recently in medical imaging [33,34,35], it has not been used for differentiation of benign and malignant lesions in breast cancer MRI. Jianrui Ding, et al., [36] used MIL for the classification of breast ultrasound images. However, their proposed methodology is based on a rough segmentation of ROIs. We have mentioned in section 1.3, and section 1.4 that the motivation for investigating the performance of MIL as a CAD tool is to do screening and diagnosis of breast cancer without segmenting the ROIs. There are number of ways to do this task. In the supervised learning, one way is to extract features from individual pixels and classify each pixel independently as benign or malignant pixel. However, it is both expensive and time consuming for a radiologist to label individual pixels. MIL can solve this issue by providing the label of an image on a coarser level, i.e., a single label for whole screened suspicious region or whole scan.

Another problem with the pixel-wise classification is that we get only first-order intensity information from individual pixels. We do not get any textural information. In order to extract higher-order spatial information from the image, the simplest way is to decompose the image into number of tiles and extract features from each tile. If these tiles (instances) are classified individually, this is called (tile-based) single instance learning. On the other hand, if we classify a set of these instances together, this is called (tile-based) multiple instance learning. Figure 2.2 illustrates the difference between the tile-based SIL and the tile-based MIL for the classification of breast cancer.

In this thesis we evaluate the suitability of tile-based MIL as a CAD tool for the classification of breast cancer MRI. Specifically, we compare the tile-based MIL performance with both conventional (ROI-based) and tile-based SIL. In the context of tile-based MIL, an MRI is a bag composed of square tiles, which correspond to instances. Tile-based features have several advantages over ROI-based features. We discuss them in Chapter 3. To our knowledge this is the first time that a MIL-based CAD (which does not require segmentation of ROIs for feature extraction) has been proposed for the detection and diagnosis of cancers in breast MRI.

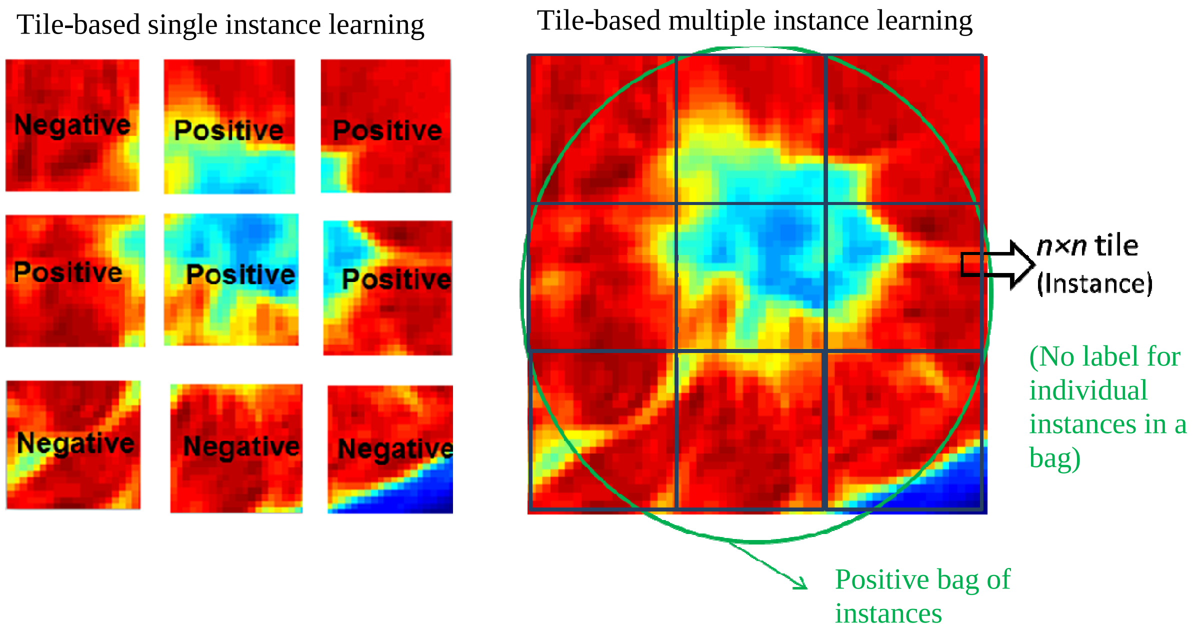

Tile-based single instance learning
Negative
Positive
Positive
Positive
Positive
Positive
Negative
Negative
Negative
Tile-based multiple instance learning
n×n tile
(Instance)
(No label for individual instances in a bag)
Positive bag of instances

Figure 2.2: Difference between the tile-based SIL and the tile-based MIL.

2.2. Algorithms

In any pattern recognition problem where we do not know the underlying probability density distribution function or it is difficult to estimate, it is desirable to use a non-parametric technique for classification [37]. k-nearest neighbour (kNN) is a well-established and an attractive non-parametric classification technique in pattern recognition literature [38]. In order to investigate the classification performance of MIL on breast MRI data in Chapter 3 and 4, we evaluate the performance of a MIL-based kNN algorithm called **Citation-kNN** (CkNN) [39]. We have chosen CkNN because it is a non-parametric technique which involves the optimization of only two parameters (i.e., reference neighbours ‘*k*’ and citer’s rank ‘*c*’). Moreover, CkNN has been used for solving various MIL problems with high accuracy [19].

CkNN is an adaption of conventional kNN to a MIL problem. There are two main differences between conventional kNN and CkNN. The first difference is of a distance function. In conventional kNN, nearest neighbours are usually decided according to Euclidean distance, which measures the distance between individual instances. However, MIL is concerned with the classification of bag of instances which cannot be represented by a single point. Therefore, Euclidean distance cannot be used for measuring the distance between set of instances as required by a MIL dataset. This distance measure problem has been resolved by other distance functions [40,41], i.e., minimum Euclidean distance (Minimum), maximum distance, Hausdorff distance, and centroid distance, etc. These (MIL-based) distance functions measure the distance between sets of points rather than individual points. The effectiveness of distance function depends on the nature of dataset. Given the two bags (set of instances): $X = \{x1, x2, \ldots, xn\}$, and $Y = \{y1, y2, \ldots, ym\}$, above distance functions can respectively be defined as:

$$D_{\mathrm{Min}}(X,Y)=\min_{x\in X}\{\min_{y\in Y}\{d(x,y)\}\}$$

$$\mathrm{D}_{\mathrm{Max}}(X,Y) = \max_{x\in X}\{\max_{y\in Y}\{d(x,y)\}\}$$

$$D_{\mathrm{H}}(X,Y) = \max\{h(X,Y), h(Y,X)\}, where$$

$$h(X,Y) = \max_{x\in X}\min_{y\in Y}\{d(x,y)\}$$

$$\mathrm{D}_{\mathrm{cen}}(X,Y) = d\left(\frac{1}{n}\sum_{i=1}^{n} xi, \frac{1}{m}\sum_{j=1}^{m} yj\right)$$

Graphical illustration of Hausdorff distance and minimum Euclidean distance is given in Figure 2.3. In Figure 2.3, we have two bags X and Y, each containing multiple instances. h(X,Y) represents the distance from the maximum of the bag X (the farthest instance in the bag X relative to the bag Y) to the minimum of the bag Y (the closest instance in the bag Y relative to the bag X). Similarly, h(Y,X) represents the distance between the maximum of bag Y and the minimum of bag X. Hausdorff distance between these two bags is h(Y,X), which is the maximum of h(X,Y) and h(Y,X). Minimum Euclidean distance is represented by E_{min}, which is the distance between the minimum of bag X and the minimum of bag Y. We can see that in each bag we have multiple instances, and thus we cannot use Euclidean distance to find the distance between two bags. The reason is that Euclidean distance can only measure the distance between two points, but cannot be used to calculate the distance between two bags (sets of points).

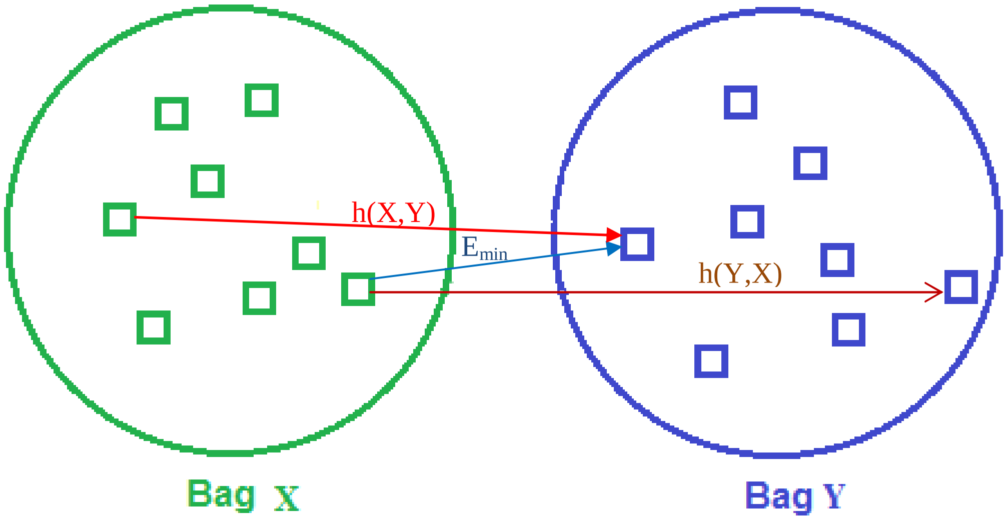


Figure 2.3: Graphical illustration of MIL-based distance functions**.**

The second difference between the conventional kNN and CkNN is of classification approach. In conventional kNN, the label of a test example is determined by the majority vote of k-reference neighbours. However, this is not sufficient criterion for the classification of bags in a MIL dataset. The assumption is that a positive bag in the feature space may contain large number of negative instances. In such a case, it is likely that a negative test bag considers positive training bags as its neighbours. This scenario may result in false positive outcome [39]. Therefore, to obtain an optimal performance (which is robust to the noise), Wang and Zuker [39] devised another method to combine the nearest bags for predicting an unseen bag, called citation-kNN. In CkNN *citers* (‘c’) in addition to the bag’s *reference* neighbours (‘k’) are sought to predict the label of an unseen bag. k-reference neighbours are the nearest bags to the test bag and the term citers refer to those bags which consider the test bag as their neighbour.

Please note that, both CkNN and conventional kNN can be utilised to do classification in both MIL and SIL paradigms. For example, we can use citation-kNN for single instance learning if it is utilised with Euclidean distance. Similarly, conventional kNN can perform MIL-based classification if we use Hausdorff distance or any other MIL-based distance function. In other words, distance function is the main difference between conventional kNN and MIL-based kNN. The notion of *citation* in CkNN is only used to get an optimal classification of MIL datasets.

k-nearest reference neighbours for a particular bag are defined by a MIL-based distance function (i.e., maximum Hausdorff distance) and c-citers are ranked according to the similarity with the test bag. In CkNN, both reference’s labels and citer’s labels contribute equally towards the classification of an unknown bag. Suppose there are kp and kn positive and negative bags in k-nearest reference neighbours and cp and cn, respectively denote the number of positive and negative bags in c-nearest citers. Then, a test bag is labelled positive if the number of the positive bags in the mutual reference and citer decision is greater than the negative bags, i.e., $kp + cp > kn + cn$, otherwise, the bag is negative.

To compare the performance of CkNN in a SIL paradigm we also select **kNN** [42]. kNN is a simple, effective and non-parametric supervised technique which has been used extensively [43]. In kNN, most common class of k-nearest (reference) neighbours of a test example determines its class label. Choice of k is not straightforward. A small value of k is vulnerable to the noise, while a large value of k increases the computational complexity of the algorithm. For a two-class classification problem, the value of k is usually a small odd integer to avoid the tie [44]. For a fixed length feature vector, the reference neighbours are usually decided according to Euclidean distance.

Both kNN and CkNN are lazy learners which store all the training examples until testing. They do not construct an explicit description of the target function by generalising the training data.

In order to predict the class of a new (test) example, these learners first calculate the distance of a test example with every sample in the training set. Then, the class of a test example is decided according to the majority among the k-nearest neighbours (and c-citers in case of CkNN). As a member of lazy learning family, they have a disadvantage of time complexity in making predictions. For example, kNN requires $O(Nd)$ time to predict the class of a test example, where N is the number of training examples in d dimensions [44]. Similarly, the computational complexity of CkNN is $O(n^2Nd)$, where n is average number of instances in a bag, and $O(n^2)$ presents the time complexity for pair-wise distance measurement between instances across the bags [45]. In addition, for a fixed length feature vector (d), CkNN computes c-citers by performing a sort operation $O(NlogN)$ for every training example [46].

Before investigating the efficacy of CkNN using breast MRI dataset, it is important to study how CkNN works with different distance functions, and how different parameter values (k, and c) affect the classification performance of CkNN. In particular, we want to study the performance of CkNN with the two most extensively used distance functions, i.e., minimum Euclidean distance and maximum Hausdorff distance. This is the motivation for our first aim of the research (section 1.4). We address this aim in the beginning of Chapter 3 by performing an experiment using synthetic datasets.

Many MIL algorithms follow the standard MIL assumption, for example, multi instance support vector machine (MI-SVM), expectation maximization diverse density (EM-DD), axis parallel rectangle (APR) etc. Please refer to [48] to find the survey of these MIL learners. These are purpose-built MIL learners which were developed specifically to solve the MIL problems. On the other hand, there are several other MIL algorithms which don’t obey strict the MIL assumption [47]. These learners are based on generalised MIL assumptions, where the bag label is decided by considering most or all instances in a bag [47,48]. MIL learners where the bag level label is derived from instance labels are categorised as instance based approaches [46]. Instance based approaches assume that instances have hidden class labels. These approaches first estimate the function to determine the class labels at instance level. Then, they use that function to predict the class at bag level.

Another category of MIL learners is metadata-based approaches [46]. These learners assume that some meta-level information determines the class of a bag. In metadata-based approaches, a sample (bag) is embedded into a single instance vector space. In this way, they allow the supervised learners to make prediction in the MIL domain while dealing with the set (bag) of instances as a single unit [46]. In other words, these approaches determine the label of a bag without considering the labels at instance level [25]. CkNN belongs to this category of MIL learners. In Chapter 3, we compare the performance of CkNN, where metadata information is used to predict the class of a bag, to other instance based MIL learners, for example, MI-SVM, APR, and EMDD, which follow the standard MIL assumption.

2.3. Feature Selection

Classification performance of a learner depends upon the size of training data, number of features and the learner's complexity. For a given size of training data, if the number of features is increased, the performance of classifier is improved. But this does not imply that, for accurate classification we can add as many features as possible. Because, classification performance is increased with increasing dimensionality up to certain point, after which it begins to degrade. This process is called *curse of dimensionality* or *peaking phenomenon* [49]. The reason for peaking phenomenon is that if we keep on adding features to get accurate classification, at certain point the classifier over-fits the given data at the cost of poor generalisation. This means that, although it gives accurate performance on the given (seen) data but due to poor generalisation it will not give good performance on an un-seen datum. Also, the feature space becomes sparser as the dimensionality of feature space grows and given samples become less representative of their class. Jain A. & Duin, P. [49] provide a rule of thumb about the optimal number of features, which states that, "at least ten times as many training samples per class as the number of features is a good practice". To avoid the curse of dimensionality, we utilise a feature selection technique to get the required number of features. Feature selection is not only important to improve the generalisation accuracy, but also to learn a more compact representation of a pattern [50]. Moreover, feature selection helps us to eliminate the non-useful features and reduces the dimensions of feature space [16].

There are several feature selection techniques available for traditional supervised learning, where each instance is associated with a class label. As mentioned in section 2.1, MIL differs from the conventional SIL in the nature of learning examples. In MIL, a set of multiple instances constitute a learning example, where we know the class label of each set of instances, but not for the individual instances. Since, feature selection algorithms used in the supervised learning have been developed to address the learning tasks where the label of each instance is known; these algorithms are not directly applicable for MIL applications [50]. Due to the difference of classification approach of both MIL and SIL methods, we believe that MIL-based feature selection criterion is important to reduce the dimensionality of a MIL dataset. Since in MIL, not much work has been done on the development of feature selection techniques; we develop a MIL-based feature selection algorithm (as part of our third aim) in Chapter 4.

2.4. Performance Estimation

A. *True and apparent performance*

True performance of a classifier is based on evaluation of the classifier on the ‘true’ distribution of samples. Here ‘true’ refers to the unseen samples which have not been used to design the model. If a classifier is designed and evaluated on the same dataset, the estimated performance is called apparent performance. In general, the apparent performance of a classifier is optimistically biased and tends to have high variance. *Bias* is a measure of how much the expected prediction of a classifier varies from the correct value. “*Variance* measures how much, on average, *model prediction* vary around the expected value (going from one subset to another)” [51]. If the bias is kept low, the learning model is subjected to the risk of having high variance. Similarly, if the variance is kept low, the model does not learn the particular characteristics of the data and may have high bias. Finding an optimal model is to get the best trade-off between the bias and the variance.

Graphical representation of bias and variance is given in Figure 2.4. In this figure, the central Red cross represents the value to be estimated (correct value). Green circles represent several estimates over different samples. If a classifier gives all predictions close to each other and to the correct value, it has low bias and low variance. On the other hand, if there is a scatter among the predicted values and all are away from the correct value, the model has high bias and high variance as illustrated in Figure 2.4.

The presence of bias indicates that the given model does not contain the solution. This is called *under-fitting*. If there is variance, it indicates that the model has learnt the particular characteristics of the given data. This is called *over-fitting*. This means that, the model gives good classification performance only on the samples that were used to design it. In case of a new sample, such classifier tends to fail to predict the class. In other words, the classifier has been designed or modeled only to classify a particular dataset. The true performance is almost invariably lower than the apparent performance. The reason of relatively better apparent performance is the over-fitting of the classifier to the particular characteristics of the sample data [52]. Designing and evaluating a classifier on the same dataset is useless because it tends to give the poor performance on new samples. Since the focus of machine learning is rarely to replicate the training data but the prediction for new cases [51], to design a model which possibly gives the correct prediction for new cases, we split the data into independent training and test sets.

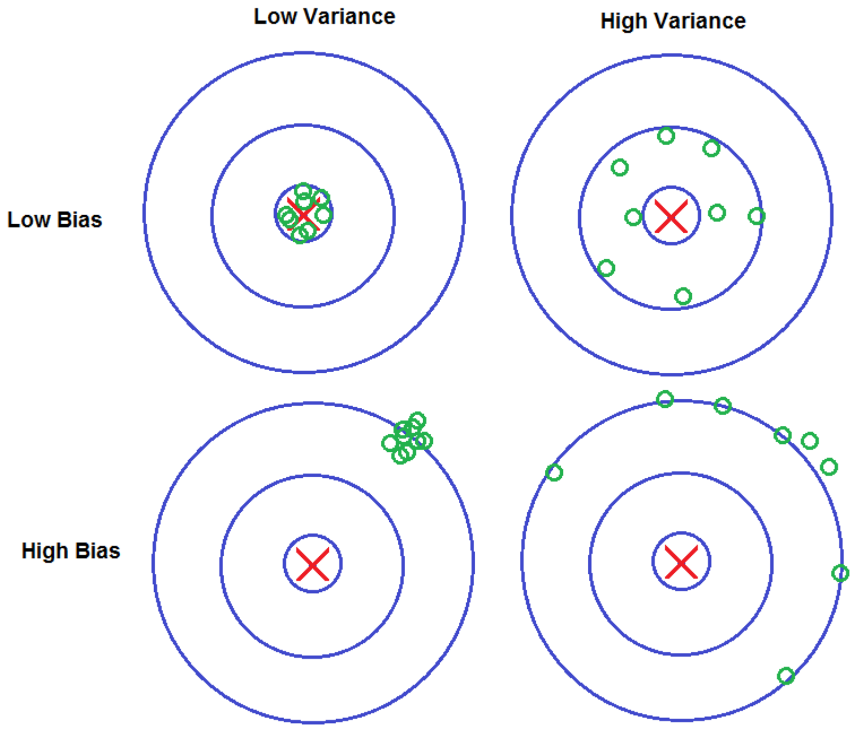


Figure 2.4: Graphical illustration of bias and variance.

B. Stratified 10-fold cross validation

To split the data into training and test sets, we utilise stratified 10-fold cross validation (CV) [51]. In stratified CV, we split the data into training and test sets by roughly preserving the class proportions in each fold. We use bag level CV for MIL datasets to maintain the bag structure consisting of multiple instances. In each fold of 10-CV, dataset are partitioned into an independent 90% training data and 10% test data. We have chosen 10-CV because; it gives an estimate which has low variance as compared to single hold-out estimation. For example, if data are split into 90% training set and 10% test set in a single hold-out estimator, it is likely that there is a lot of variation in estimated performance obtained with different partitions of data for the training and test sets. However, 10-fold CV reduces this variation by averaging over 10 different partitions of data which makes the performance estimate less sensitive to the partitioning of the data.

C. Performance measure

The most commonly used measure of the performance of a classifier is error rate [53]. Error rate is a measure of misclassification rate of a classifier, calculated as follows,

$$Error\ rate = \frac{Number\ of\ misclassified\ samples}{Total\ number\ of\ samples}$$

However, error rate summarises the overall performance of a classifier by assuming equal misclassification cost for both false positive and false negative. In many applications, we need to distinguish different type of errors. But, we do not have enough information to associate different costs functions for different types of errors. In this thesis, we summarise the performance of a classifier using a receiver operating characteristics curve (ROC), i.e., area under ROC curve (AUC) [54]. Unlike error rate, the measurement of AUC is not based on equal misclassification cost assumption. Classification accuracy ($1 - error$) is another commonly used performance measure. But, the estimate of accuracy depends on the prior distribution of samples. Suppose a dataset contains high proportion of samples in positive class as compared to negative class. In this case, even if classifier totally fails to distinguish negative samples from positive samples, the estimated accuracy can still be very high because of high proportion of positive samples in the dataset. Unlike accuracy, the AUC gives an estimate of a classifier's performance which is independent of the prior distribution of samples in the dataset.

We have chosen AUC as a performance measure because it is independent of the prior class density distribution and misclassification costs. Moreover, AUC is more discriminative between classifiers [55], and more suitable for MIL problems [56]. ROC curve is a plot of true positive rate (sensitivity) against false positive rate (1- specificity) across varying cut-offs. The equations for sensitivity and specificity are given below.

$$Sensitivity = \frac{True\ Positives}{Total\ no.of\ positive\ samples = (True\ Positive + False\ Negative)}$$

$$Specificity = \frac{True\ Negative}{Total\ no.of\ negative\ samples = (False\ Positive + True\ Negative)}$$

where

True Positive are positive samples which are correctly classified as positive,

True Negative are negative samples which are correctly classified as negative,

False Positive are negative samples which are incorrectly classified as positive, and

False Negative are positive samples which are incorrectly classified as negative.

ROC curve presents the performance of a two class classifier over a range of discrimination threshold. Using ROC curve analysis, we can select an optimal threshold value for clinical use [57]. The AUC sums up the whole ROC curve irrespective of the threshold value [58,59]. Thus AUC is an effective measure of performance which is not affected by the decision criterion. An AUC of 1 corresponds to the perfect discrimination between two classes given the correct threshold, while an AUC of 0.5 indicates that for any threshold value, classifier discriminates between two classes at random.

2.5. Breast MRI Database

MRI data used in the experiments presented in Chapters 3 and 4 are derived from breast MRI database, which consists of 165 clinical bilateral breast MRI investigations of 150 subjects. MRI investigations were performed by Queensland X-ray using an 8-channel breast coil on 1.5 T GE Signa HDxt scanner with the patient lying in prone position. The examinations included screening cases (high risk category), problem-solving cases, and cases that had previous surgery. In each case at least one suspiciously enhancing lesion was identified by the reporting radiologist. All lesion findings were subsequently verified by cyto- or histopathology. Each lesion was individually biopsied under either MRI or ultrasound guidance. The detail of lesion pathology can be found in [92,111]. We use the T1-w anatomical, T2-w anatomical and the DCE-MRI data in the experiments.

Registered non-fat suppressed (using 2D tailored radio frequency sequence) T1-w anatomical MRI and registered fat suppressed (using 2D short time inversion recovery sequence) T2-weighted anatomical MRI are used. Both T1-w and T2-w MR images were acquired axially with pixel spacing of 0.5859 mm to 0.6250 mm. The acquisition matrix for non-registered T1-w and T2-w MRI is 512×512 and 320×224 respectively. In both cases the field of view and slice thickness is 32cm and 4mm respectively. Example images of T1-w and T2-w MRI are shown in Figure 2.5. We can see the dominance of fat in non-fat suppressed T1-w MRI.

We use non-contrast enhanced (T1-w and T2-w) MRI in Chapter 3. The use of T2-w MRI in the conventional CAD is evident from the literature. Various studies suggest that high signal intensities in lesions on T2-w MRI cannot differentiate benign and malignant lesions [6,60,61,62,63,64,65,66]. However, numerous other studies indicate that the high signal intensity in lesions on T2-w MRI is distinctive in the discrimination between benign and malignant lesions [67,68,69,70,71,72]. Few other studies conclude that T2-w features are promising diagnostic features for the classification of benign and malignant lesions [73,74,75]. Above all, it is standard clinical practice to use T2-w MRI for the diagnosis of cancer in breast MRI. This is the motivation for the second aim of our research: to evaluate the generic tile-based spatial features from T2-w MRI for the discrimination of benign and malignant lesions.

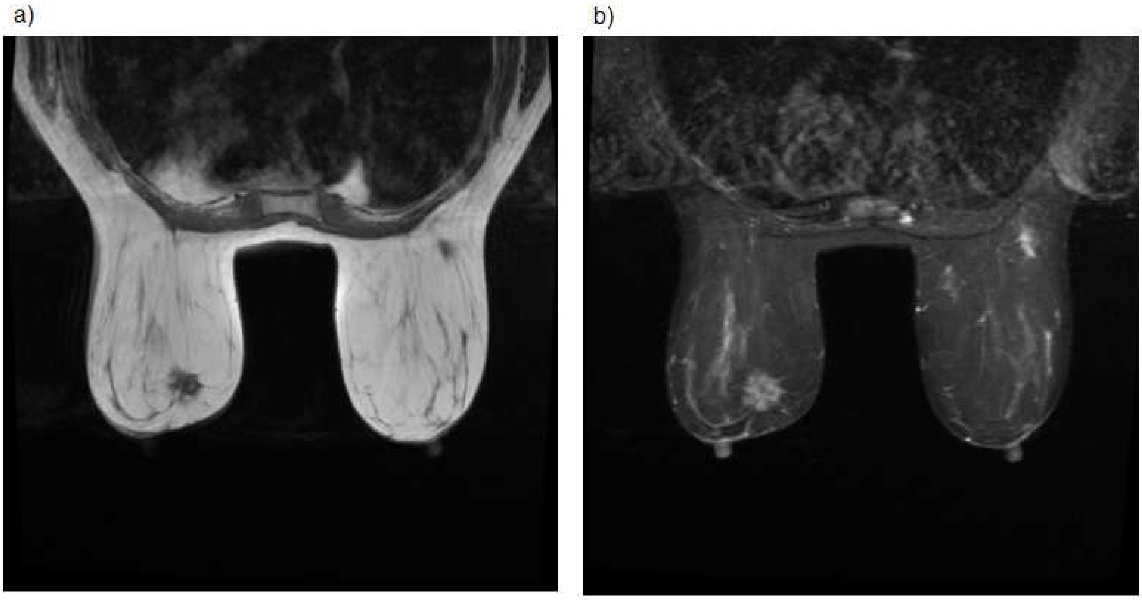


Figure 2.5: Example images of non-contrast enhanced MRI. (a) Non-fat suppressed T1-w MRI, (b) Fat suppressed T2-w MRI.

In DCE-MRI, the fat suppressed T1-weighted images were acquired as either four or five stacks using a 3-dimensional (3D) fast spoiled gradient-echo (FSPGR) sequence (Echo time = 3.4ms, Repetition time = 6.5ms and flip angle of 10°). The first stack corresponds to baseline pre-contrast images and the remaining stacks comprise of post-contrast images. Acquisition time of each stack was around 90 seconds with a 45 second delay between the pre-contrast and the first post-contrast stack. The second last stack was acquired in a sagittal orientation. All of the remaining stacks were acquired axially with a field of view of 32cm, a 360×360 acquisition matrix, and slice thickness of 1mm. Each stack contained more than hundred two dimensional slice images ranged from 116 to 182 with a median of 150. We use only four stacks which were acquired axially. The example images of DCE-MRI are shown in Figure 2.6.

Several studies favour the combined use of DCE-MRI and T2-w MRI in conventional CAD for improved discrimination of benign and malignant lesions [11,76,77,78]. Moreover, there is considerable evidence to support the use of both spatial and temporal (i.e., spatio-temporal) features for the diagnosis of breast cancer [10,15,79]. Therefore, to address the third aim of this research, we use both DCE-MRI and T2-w MRI in Chapter 4. In particular, we investigate the importance of tile-based spatio-temporal features for the discrimination of benign and malignant lesions.

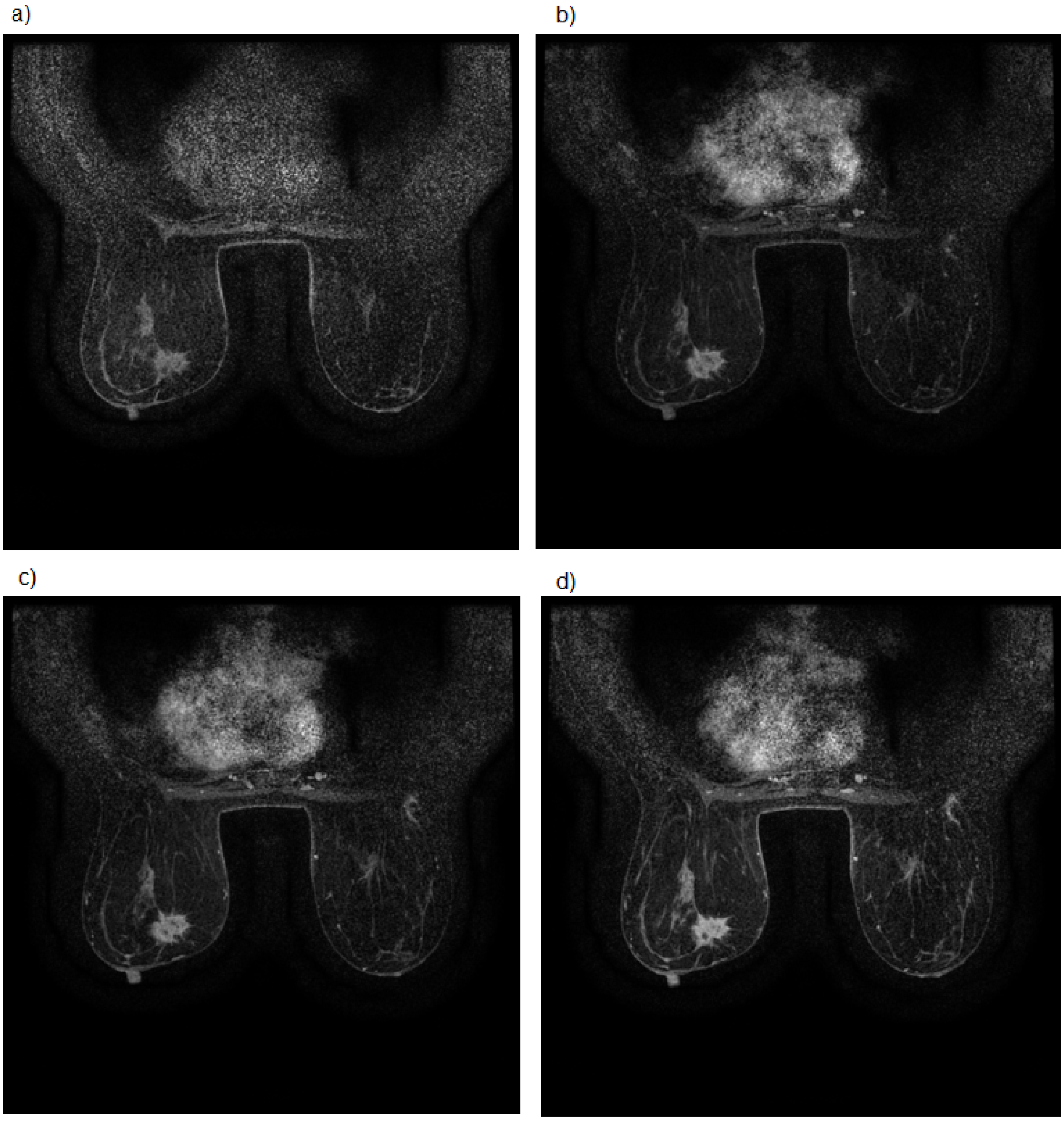


Figure 2.6: Example images of DCE-MRI. (a) Pre-contrast, (b) First post-contrast, (c) Second post-contrast, (d) Third post-contrast DCE-MRI.

We use manually segmented ROIs to ground truth the tiles in MIL dataset. Manual segmentation of each lesion was performed in 3D by a radiologist with more than 12 years experience in breast MRI. Each lesion in the subtraction volume (first post-contrast stack minus the pre-contrast stack) from the DCE-MRI stack was manually delineated using CAD tools such as region-growing tool in OsiriX. In case of over-segmentation (due to inclusion of surrounding enhancing tissue or blood vessels), a 3D maximum intensity projection was used to view the lesion and the 3D region-growing mask was cropped using the freehand scissor tool. Small holes were filled using a morphological closing operation by a 3×3×3 cube on each 3D connected component.

The resulting 3D volumes-of-interests (VOI) were used as the ground truth. Moreover, the radiologist provided the mass-like or non-mass-like description of each lesion, as defined by the American College of Radiology (ACR) Breast Imaging Reporting and Data System (BIRADS) lexicon [80].

For simplicity, to discriminate benign and malignant lesion, we utilise only one slice that contains the maximum cross-section between the subtraction image of pre- and post-contrast MRI and the original MRI. This reduces the three dimensional anatomical MRI data and four dimensional DCE-MRI data to two dimensional and three dimensional data respectively. Additionally, we do not perform heart, or breast boundary segmentation. Rather, we utilise sub-images that contain manually segmented ROIs of suspicious lesions. To differentiate malignant lesions from normal tissues, we also utilise sub-images from contralateral side of the breast which contains normal tissue. We discuss this in more detail in Chapter 3. Please note that we use the manually segmented ROIs to ground truth the data in the training phase. However, we do not need segmentation of ROIs to classify an unseen image (sub-image). The approval of the Human Research Ethics Committee at the University of Queensland was obtained for this research. The Ethics numbers are 2007001313 and 2013000563.

2.6. Conclusion

In this chapter, we have presented the necessary background on machine learning concepts to support our experiments in the later chapters. Specifically, we have presented an introduction to the statistical pattern recognition in images based on multiple instance learning. We have provided reasoning for the methodology, which we adopt in the experiments in the next chapters. In addition, we have provided the motivation for the aims of this research. In the next chapter, we evaluate the performance of CkNN in the classification and identification of non-contrast enhanced breast MRI.

Chapter 3

Multiple instance learning for breast MRI using spatial features alone

In this chapter, we investigate the potential of multiple instance learning (MIL) [18] in the identification and classification of breast cancer MRI. In particular, we compare the performance of MIL against traditional single instance learning (SIL) using both generic tile-based and domain specific ROI-based features. As mentioned in Chapter1, sensitivity of MRI has been improved with the use of contrast agent in DCE-MRI. However, the injection of contrast agent is an expensive and invasive procedure. Therefore, here we initially investigate the performance of MIL for the classification of only T2-w MRI, which provides intensity based features in spatial domain alone. In the next chapter, we investigate the importance of both temporal and spatial generic features for the diagnosis of breast MRI using both DCE-MRI and T2-w MRI. The material presented in this chapter is expanded version of [81]. Please note that section 3.2 is not part of [81]. Further, we have added an additional experiment to justify our choice of selected MIL learner. This paper was presented at the Digital Image Computing: Techniques and Applications (DICTA), Wollongong, 2014.

The rest of the chapter is structured as follows. In Section 3.1, we present a brief introduction to the CAD tools used for the identification and classification of breast cancers. In section 3.2, we perform an experiment to understand the working of CkNN using synthetic datasets. Next in section 3.3, we present an experimental methodology to evaluate the MIL performance using the breast MRI dataset. Here we specifically talk about datasets, learners, and features for the classification of breast MRI. In section 3.4 and 3.5, we present the results and discussion. We finalise this chapter with conclusion in section 3.6.

3.1 Introduction

As mentioned in Chapter 1, CAD systems are used to assist radiologists in doing non-subjective analysis of medical imaging data. There are two types of CAD systems used in clinical practice: CADx systems are for the diagnosis of breast cancer and CADe systems are used as screening tools to assist radiologists with data visualization and to highlight the suspicious regions which may be indicative of disease [82]. In breast MRI, cancer screening of patients who are at high risk of developing breast cancer is mostly performed with the dynamic contrast enhanced (DCE) MRI [16]. However, we know that the injection of contrast agent in DCE-MRI is not only an invasive procedure but also expensive. Therefore, we initially use T2-w MRI for the detection of suspicious lesions.

Once the suspicious lesion is highlighted in the screening phase, the next phase involves the diagnosis of breast cancer, i.e., the discrimination of benign and malignant lesions. CADx systems assist radiologists in the discrimination of benign and malignant lesions by providing computerised characterisation of the lesion. However, the efficacy of the computerised characterisation is dependent on the accurate segmentation of lesion from the surrounding breast tissue. As mentioned in Chapter 1, the accurate segmentation of the lesion is a challenging task due to poor signal-to-noise ratio and faint edges due to partial volume effects. Therefore, lesion segmentation is subjected to uncertainties in the automatic segmentation process or to inter-observer variations (in case of manual segmentation). Since the aim of CADx systems is to increase the efficacy of interpretation as well as to reduce the inter-observer variations [16], we use tile-based characterisation of the suspicious regions by using tile-based multiple instance learning. Clearly, tile-based features do not require delineation (segmentation) of the lesion and are independent of the variations in the lesion segmentation process, thus can increase the efficacy of interpretation.

Based on the CADx and CADe applications, we perform two studies to investigate the classification performance of MIL on breast cancer MRI. The aim of first study is to investigate the efficacy of tile-based MIL as a CADx diagnostic tool using T2-weighted breast cancer MRI and its comparison with traditional SIL approaches that require segmented lesions for the extraction of ROI-based features. The aim of second study is to analyse the performance of MIL as a CADe breast cancer screening tool. Here, the algorithms aim to classify malignant lesions as positive and (either) benign lesions or normal tissue as negative.

To evaluate the efficacy of MIL for screening and diagnosis of breast cancer MRI, we utilise 60×60 sub-images (blocks) that contain either manually segmented ROIs of benign/malignant lesions or (unsegmented) normal tissue. An instance is then a feature vector extracted from the (non-overlapping) tiles within these sub-images. We select the size of square tiles on an experimental basis.

Before evaluating the performance of CkNN on breast MRI dataset, we first study the performance of CkNN using synthetic datasets. We use synthetic datasets because they give us the flexibility to change the characteristics of data which is not possible in case of real data. We also use synthetic datasets to optimise the parameters of CkNN which can then be used in the classification of breast MRI.

3.2 Understanding the Performance of CkNN Using Synthetic Datasets

Here, we create two synthetic datasets which could reflect the characteristics of breast MRI data to some extent. Following is the detail of experiment which helps us to understand how CkNN works with varying characteristics of data over a range of parameter values.

A. Methodology

For simplicity, we assume that features extracted from tiles in breast MRI dataset has Gaussian distribution in the feature space; therefore, we construct two common types of datasets based on *I-Λ* and *I-4I* datasets [83]. We change the difficulty (overlap between classes) of each dataset by modifying the variance in the *I-Λ* and *I-4I* datasets. For a more detailed description of setting difficulty of synthetic datasets, please refer to [84]. In dataset 1 (*I-Λ)*, instances are centred at different mean values, i.e., positive instances are centred at 1 and negative instances are centred at 0. The variance of instances in positive and negative bags is 1.8 and 2.3 respectively. On the other hand, dataset 2 (*I-4I)* contains instances both in positive and negative bags centred at 0, with the variance of 1.2 and 0.8 respectively. Initially, we generate each dataset consisting of 300 bags (with equal priors), with a bag size of 120 instances in 15 dimensions.

To optimize the parameters of CkNN for different characteristics of datasets, we create each dataset with the flexibility of changing its properties, such as, number of bags, no. of instances, mean, variance, dimensions, etc. In order to evaluate the working of CkNN for different bag compositions, we vary the percent positives (PP) in positive bags, i.e., from 1 to 99 PP. For each bag composition, we evaluate the performance of CkNN using two most commonly used distance functions i.e., maximum Hausdorff (Hausdorff) distance and minimum Euclidean (Minimum) distance. The performance measure utilized is mean AUC. To understand the effect of both small and large values of '*k* reference neighbours' and '*c* citer's rank' on the performance of CkNN, we also search the parameter space for 72 combinations of *k* and *c*. Table A (Appendix A) shows the [*k*, *c*] values utilised in this experiment.

B. Results

Surf plots (Figure A.1 to A.4 in Appendix A) illustrate the performance of CkNN using both Hausdorff distance and Minimum distance over 72 [*k*, *c*] combination values (as mentioned in Table A) for ten bag compositions of both dataset 1 and dataset 2. Colour bar maps the performance of CkNN in mean AUC. For simplicity, we also present the results only with those parameter values (*k* and *c* values) which could represent the overall performance of CkNN using two distance functions on each dataset.

For dataset 1, Figure 3.1 illustrates the performance of CkNN using Hausdorff distance in comparison to Minimum distance with *k* = 5 and *c* = 11. We can see that performance of CkNN using Hausdorff distance has an increasing trend with increasing PP, while there is no clear trend with PP using Minimum distance. Also, CkNN gives better performance with Minimum distance for 10 to 40 PP after which, results with Hausdorff distance are dominating. However, this trend of performance dominance in relation to PP on the basis of distance functions is variable over a range of *k* and *c* combination values (Figure A.1 and Figure A.2). Overall, Minimum distance gives robust performance irrespective of the bag composition, particularly with a large *k* or *c*.

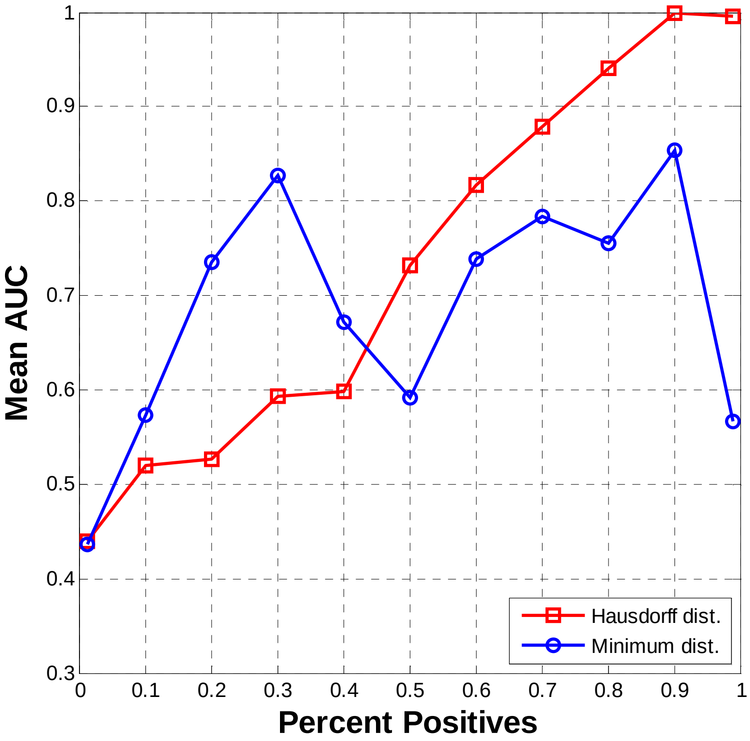


Figure 3.1: Performance of CkNN with Hausdorff distance in comparison to Minimum distance using Dataset 1.

For dataset 2, Figure 3.2 shows the performance of CkNN using Hausdorff distance and Minimum distance with $k = 3$, $c = 32$. It can be seen that, the performance of CkNN is not improved with increasing PP, rather remains unstable using either distance function. Overall, CkNN performs better with Hausdorff distance as compared to Minimum distance.

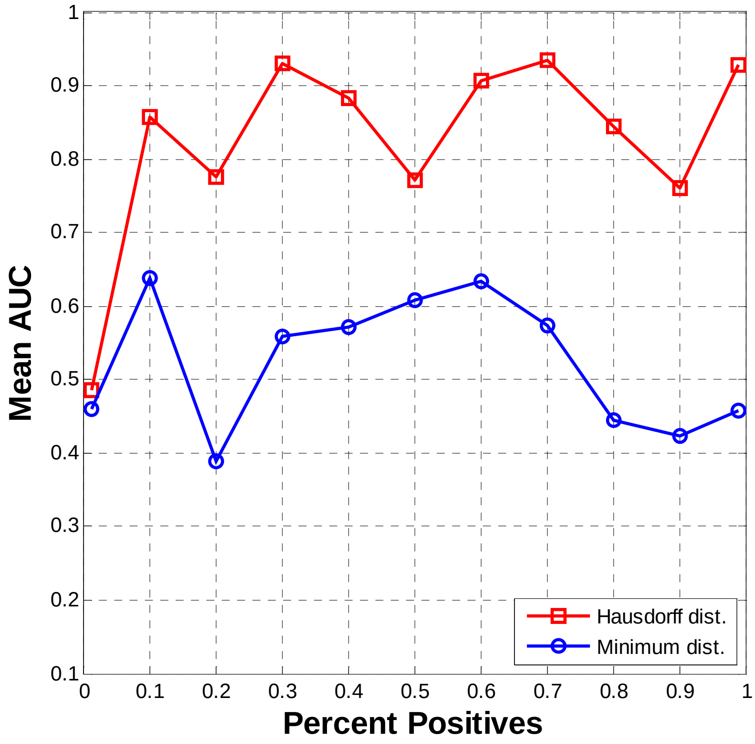


Figure 3.2: Performance of CkNN with Hausdorff distance in comparison to Minimum distance using Dataset 2.

In addition, we present the mean performance of CkNN over 10 different bag compositions using dataset 1 and dataset 2 in Figure A.5 and Figure A.6 (Appendix A) respectively. These box plots illustrate the performance of CkNN with Hausdorff distance in comparison to Minimum distance on each dataset. On each box, the central mark is the median of the performance, while the edges are 25th and 75th percentiles. Outliers are plotted individually, beyond the extent of whiskers. Figure A.5 suggests over all better mean performance of CkNN (using dataset 1) with Hausdorff distance as compared to Minimum distance particularly for small *k* and *c* combination values. For dataset 2, CkNN performs better with Hausdorff distance as compared to Minimum distance as shown in Figure A.6.

Discussion

We have presented the classification performance of CkNN with two distance functions using two synthetic datasets in Figures A.1 to A.4, Figure 3.1 and Figure 3.2. In addition, we have presented the mean performance of CkNN over 10 different bag compositions for 72 combinations of *k* and *c* values in Figure A.5 and Figure A.6. The performance estimation of CkNN on synthetic datasets indicates that CkNN shows robust performance over a wide range of [*k, c*] values with suitable distance function. However, large values of *k* compromise the basic assumption behind the nearest neighbours based learners. Also big values of *k* or *c* make CkNN computationally expensive. Therefore, we utilise only small values of *k* and c to optimise the parameters of CkNN for the identification and classification of breast MRI. Moreover, CkNN shows increasing trend with PP when utilised with Hausdorff distance using dataset 1. Steven D. Brossi, et al., [84] find the same results in their work. They present the classification performance of CkNN using *I-Λ* dataset with Hausdorff distance for different bag compositions and bag sizes (no. of instances per bag). Although characteristics of dataset 1 (bag size, class priors, and number of features) in our experiment is different from [84], the performance trend is almost identical.

The performance evaluation of CkNN on synthetic datasets gives us an insight into the working of CkNN with the two types of distance functions for different bag compositions over a range of parameter values. This experiment helps us to evaluate the performance of CkNN on breast MRI dataset. The mean performance of CkNN over different bag compositions indicates that CkNN performs better with Hausdorff distance as compared to Minimum distance. On the basis of this result, we use CkNN with Hausdorff distance for the diagnosis of breast cancer in MRI in the next section.

3.3 Experimental Methodology for the Classification of Breast MRI

A. Datasets

To analyse the relative classification performance of learners in MIL and SIL domain on different nature of breast MRI datasets (presented in section 2.5), we divide the data (presented in section 2.5) into only mass-like (dataset A) and both mass-like and non-mass-like lesions (dataset B). Table 3.1 highlights the detail of datasets used in the experiments. Dataset A contains 77 mass-like lesions out of 165 lesions and dataset B consists of 129 (both mass-like and non-mass-like) lesions from 129 subjects containing single lesion. In total, there are 53 malignant and 24 benign lesions in dataset A and 77 malignant and 52 benign lesions in dataset B. Each lesion in the subtraction images of DCE-MRI was manually delineated by radiologist using CAD tools such as region growing tool in OsiriX [85]. These delineated lesions are used to estimate the ground truth for labelling the tiles in MIL dataset.

Table 3.1: Detail of datasets

Dataset	**Size of Dataset**	**No. of Malignant Lesions**	**No. of Benign Lesions**
Dataset A	77 mass-like lesions	53	24
Dataset B	129 mass-like and non-mass-like lesions	77	52

For tile-based experiments, we use segmented blocks of size 60×60 pixels from T2-w and bias field corrected T1-w MRI (T1-w MRI is used only for the selection of tile size). This block size covers most of the lesions in the database and the same size of blocks for all the subjects allows consistency in the experimental evaluation. There are two sets of segmented blocks. The block set overlapping roughly with the manually segmented ROI, contains either malignant or benign lesions, while the other set consists of non-lesion (normal) tissue. Due to inhomogeneous intensities in MRI, we normalize T1- and T2-w MR images before processing by dividing T1 and T2 pixel wise by the sum of T1, T2 and ε.

$$Ti = \frac{Ti}{T1 + T2 + \varepsilon}$$

where *i*=1 and 2.

We ground truth the datasets using histopathology reports. For MIL, we consider a block of 60×60 pixels of T2-weighted MRI overlapping with the ROI as a bag and feature vector from a tile of $n \times n$ pixels as an instance. Malignant cases constitute the positive bags while image blocks containing benign lesions make up the negative bags. Moreover, in experiments where normal tissues are used, they are also considered as negative. We ground truth the tiles (instances) in the bags containing malignant lesions using the difference between the pre and post-contrast images. Specifically, a tile is considered as positive if at least one pixel in the tile overlaps with a malignant

ROI, otherwise it is negative. In addition, we consider all tiles in the bags containing benign lesions or normal tissue as negative.

To evaluate the performance of MIL as a diagnostic tool in CADx study, we use both datasets A and B. For performance estimation of MIL as a screening tool in CADe study, we only use dataset B with both normal and suspicious lesions. We have selected dataset B for CADe study to train the classifiers on a variety of breast tissue types necessary for the screening purpose.

B. *The Learners*

In order to investigate the classification performance of MIL on the breast MRI data, we compare CkNN with traditional (single instance) kNN. In addition, we use random forest (RF) both for feature selection and to benchmark the performance of variants of kNN. RF is a commonly used robust classifier which can handle large amount of features without being affected by the curse of dimensionality due to inherent internal randomised distribution of data among decision trees [86]. Moreover, it provides an estimate of the importance of individual features [87,88]. In the experiments, we select ROI-based features on the basis of their importance returned by RF.

We use *Multiple Instance Learning Toolbox*[I] [89] an add-on to PRTools toolbox written in MATLAB®, and *PRTools toolbox* 4.2.0[II]. For random forest classification, we use R Version 3.0.148[III], [90] and the randomForest package [91]. Each random forest classifier is constructed using 2000 trees and the parameter '*mtry*,' which controls the number of variables randomly sampled as candidates at each split, is set to 3.

We optimise the parameters of two versions of kNN on the synthetic datasets *I-Λ* and *I-4I* [83]. We make certain modifications in the creation of *I-Λ* and *I-4I* to achieve breast cancer MRI data characteristics in terms of difficulty, number of features and number of bags. For more details on the creation of synthetic dataset, modification in associated characteristics and optimisation of parameters on synthetic datasets, please see section 3.2. The selected parameter values are tabulated in Table 3.2. We appreciate that the performance of learners in the classification of breast cancer MRI may be pessimistically biased due to optimisation of parameters on the synthetic datasets rather than on the breast MRI dataset. However, we are interested in the relative performance of the learners and have limited breast MRI data.

As mentioned in Chapter 2, CkNN does not follow the explicit MIL assumption. It classifies the bags on the basis of majority vote of the neighbouring bags. There are other MIL algorithms

[I] http://prlab.tudelft.nl/david-tax/mil.html
[II] http://prtools.org/software
[III] http://www.r-project.org

which strictly follow the standard MIL assumption, such as, EM-DD, MI-SVM, and APR. These are the types of eager learners which generalise the training data to construct a concept space. In contrary, CkNN is a type of lazy learner which does not generalize the training data. Rather, it stores all the data until testing. Here, we also compare the performance of CkNN with the other MIL-based eager learners which follow the strict MIL assumption.

Table 3.2: Optimised parameter values

Dataset	Learner	Parameter Values
A	CkNN	k=7, c=10
	kNN	k=7
B	CkNN	k=3, c=2
	kNN	k=13
B (for malignant versus benign or normal)	CkNN	k=7, c=5
	kNN	k=13

C. Feature sets

To avoid over-fitting to the data, we use different feature sets for selection of tile size and for comparison of MIL performance with SIL. Specifically, we use generic tile-based features using both T1- and T2-w MRI to select the tile size. Moreover, we utilise these features to compare the performance of CkNN with other MIL algorithms. A list of generic tile-based features used for both, the selection of tile size and comparison of MIL-based algorithms is given in Table 3.3.

To estimate the performance of CkNN in CADx and CADe studies, we use modified intensity and edge intensity features which have either been used in image database or in medical image analysis. Modification of the features is made to acquire information relevant to the lesion by extracting directional features. These features are generic in nature, which represent the intensity statistics in a tile (e.g., mean, variance, edge strength) rather than specific to breast cancer. Table 3.4 presents the generic tile-based features extracted from T2-w images alone.

Table 3.3: List of tile-based features for the selection of tile size

1. Mean intensity in T1-w MRI
2. Mean intensity in T2-w MRI
3. Standard deviation in T1-w MRI
4. Standard deviation in T2-w MRI
5. Mean gradient (magnitude) T1-w MRI
6. Mean Laplacian in T1-w MRI
7. Mean Laplacian in T2-w MRI

Table 3.4: List of tile-based features used in CADx and CADe studies

1. Mean intensity
2. Maximum of standard deviation along axial direction
3. Maximum of standard deviation along transverse direction
4. Mean absolute gradient in axial direction
5. Mean absolute gradient in transverse direction
6. Mean of maximum Laplacian in axial direction
7. Mean of maximum Laplacian in transverse direction

For the ROI-based traditional classification, we use T2-w specific features from manually segmented lesions (ROIs). The ROI-based T2 features from literature are listed in Table 3.5. In CADx study, we perform feature selection for the ROI-based T2 features using RF. The importance coefficient for each feature is taken to be its mean decrease in Gini coefficient. After performing feature selection we obtain: (1) for dataset A, a set of first 10 features (2) for dataset B, a set of middle ten features from 2 to 11 listed in Table 3.5. Here we briefly describe the ROI-based features presented in Table 3.5. Further for more details on ROI-based features please refer to the relevant references.

Table 3.5: List of ROI-based features used in CADx Study

1. Mean in VOI relative to fat [60,92]
2. Standard deviation in VOI relative to fat [60,92]
3. 20th percentile in VOI [93]
4. 90th percentile in VOI [93]
5. Presence of oedema (92nd percentile in 2mm shell) [92,93]
6. Presence of oedema in non fatty tissue [92,93]
7. Homogeneity [11]
8. Maximum correlation coefficient [11]
9. Sum average [11]
10. Radial gradient index [11]
11. Presence of oedema (98th percentile in 20mm shell) [92,93]
12. Mean of eroded VOI [92,93]

Features 1 to 4 and 12 represent the signal intensity of the lesion in normalised T2-w images (normalisation is done with respect to the intensity of fat) in VOI. The details of normalisation of T2-w images with respect to fat can be found in [92]. Features 5 and 11 are oedemas which are computed from a thick margin (shell) defined around the VOI. Please note that oedema is the swelling caused by the accumulation of watery fluid in the body cells or tissues which appears as hyper-intensity in T2-w MRI. The presence of oedema can be indicator of carcinoma, or associated with inflammatory breast cancer. However, the occurrence of oedema does not necessarily indicate malignancies. Heart disease may also cause development of oedema [94]. These above mentioned features are essentially proposed by van Aalst et al. [93]. In feature 6 (which is a variation on feature 11), we do not include fat pixels (in the shell) in the feature computations. Features 7 to 9 are grey-level co-occurrence textural features, and feature 10 is radial gradient index, derived from normalised T2-w images. The radial gradient index correlates with the degree of peripheral hyper-intensity, and has been shown to exhibit the early peripheral enhancement in DCE-MRI [95].

In CADe study, we compare the performance of learners using tile-based features alone. We do not compare the performance of CkNN against ROI-based classification because in non-contrast enhanced MRI (e.g., T2-w MRI), segmentation of ROIs itself requires prior detection of lesions. The dataset are split into training and validation subsets using bag level stratified 10-fold cross validation (CV) [51]. Mean area under receiver operating characteristics (ROC) curve (AUC) [54] over 10-fold CV is used as a performance measure. For both CADx and CADe studies, we perform a t-test to evaluate the significance of the difference between the AUC obtained with CkNN and single instance kNN and RF.

D. Selection of tile size

Tile size is clearly an important parameter in the tile-based image classification because tile represents the feature vector. We select the tile size by evaluating the performance of CkNN against different sizes of tiles. To select a suitable tile size, we use dataset B with both suspicious and normal tissue types to get a size of tile robust to the nature of dataset. Features, listed in Table 3.3, are extracted from non-overlapping tiles of size ranging from 2×2 pixels per tile to 30×30 pixels per tile. We use non-overlapping tiles to have minimum redundant information from tiles. CkNN with Hausdorff distance is used to classify the dataset into malignant and normal tissue/benign lesion. The performance of CkNN for different tile sizes is shown in Figure 3.3.

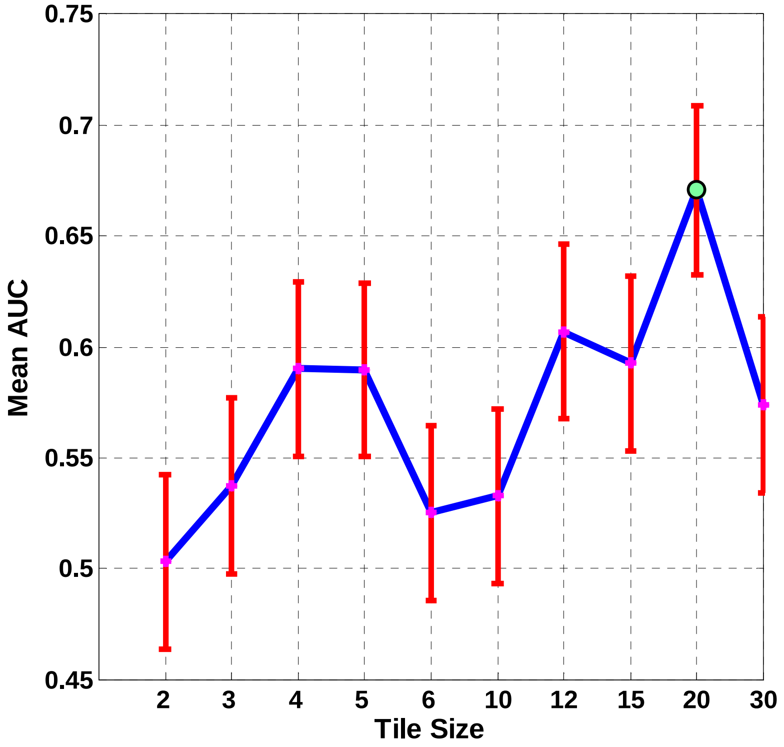


Figure 3.3: Mean AUC for square tile sizes.

From Figure 3.3 it can be seen that, CkNN performs best with the tile size of 20×20 pixels. This corresponds to nine non-overlapping instances in a bag of size 60×60 pixels. Therefore, we use this tile size of 20×20 pixels in both the CADx and CADe studies. Further, we compare the performance of CkNN with other MIL learners with the selected tile size. A comparison of CkNN with other MIL algorithms is shown in Figure 3.4. Figure 3.4 shows that CkNN performs better as compared to other MIL algorithms on anatomical breast MRI data. This supports our choice of lazy learner.

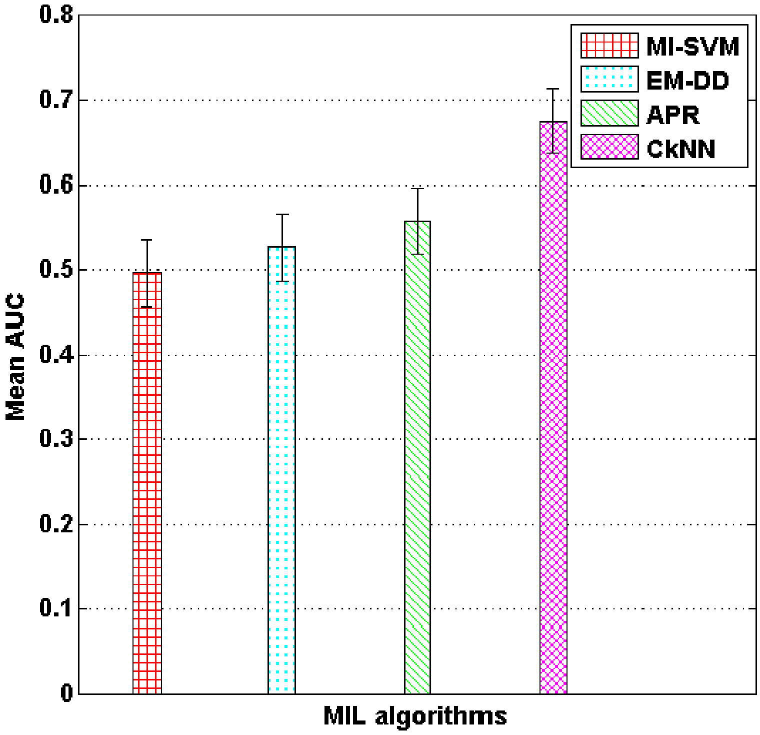


Figure 3.4: Comparison of MIL algorithms.

3.4 Results

A. CADx study: MIL as a diagnostic tool

For dataset A, Table 3.6 and Figure 3.5 summarise the comparison of classification performance of CkNN against kNN and RF. The corresponding average ROC curves are presented in Appendix B (Figure B.1).

Table 3.6: Performance of CkNN in comparison to SIL using dataset A in CADx study

Dataset	Technique	Learner	Mean AUC
Dataset A (77-mass-like lesions)	MIL (Tile-based)	CkNN	0.727 ± 0.058
	SIL (Tile-based)	kNN	0.549 ± 0.025
		Random Forest	0.696 ± 0.023
	Traditional SIL (ROI-based)	kNN	0.683 ± 0.062
		Random Forest	0.771 ± 0.053

From Table 3.6 and Figure 3.5 we can see that, CkNN acquires a mean AUC value of 0.727. Tile-based kNN and RF achieve mean AUC values of 0.549 and 0.696 respectively. The ROI-based kNN and RF achieve mean AUC of 0.683 and 0.771 respectively.

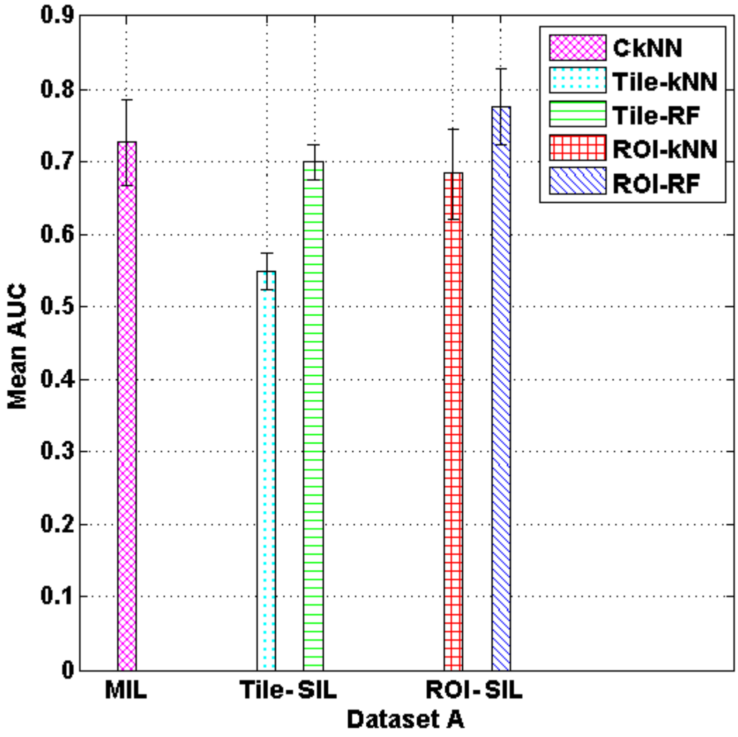


Figure 3.5: Comparison of CkNN with the tile-based and the ROI-based SIL using dataset A in CADx study.

Figure 3.5 suggests the marked difference between mean AUC yielded by CkNN and the tile-based kNN. A t-test with a *p* value <0.05 confirms statistically significant difference between CkNN and the tile-based kNN. On the other hand, the difference between performance of CkNN and the tile-based RF is not statistically significant. However, there is statistically significant difference between the performance of the tile-based SIL and the ROI-based SIL. The performance of CkNN is statistically similar to the ROI-based kNN and RF. Apparently, the best mean AUC is achieved with the ROI-based RF. But, a *p* value (0.6843) suggests that the difference between the performance of CkNN and the ROI-based RF is not statistically significant.

For dataset B, the classification performance of CkNN in comparison to the tile-based and the ROI-based kNN and RF is presented in Table 3.7 and Figure 3.6. Figure B.2 (Appendix B) presents the corresponding average ROC curves.

Table 3.7: Performance of CkNN in comparison to SIL using dataset B in CADx study

Dataset	Technique	Learner	Mean AUC
Dataset B (129 mass-like and non-mass-like lesions)	MIL (Tile-based)	CkNN	0.592 ± 0.050
	SIL (Tile-based)	kNN	0.532 ± 0.018
		Random Forest	0.617 ± 0.018
	Traditional SIL (ROI-based)	kNN	0.486 ± 0.052
		Random Forest	0.651 ± 0.048

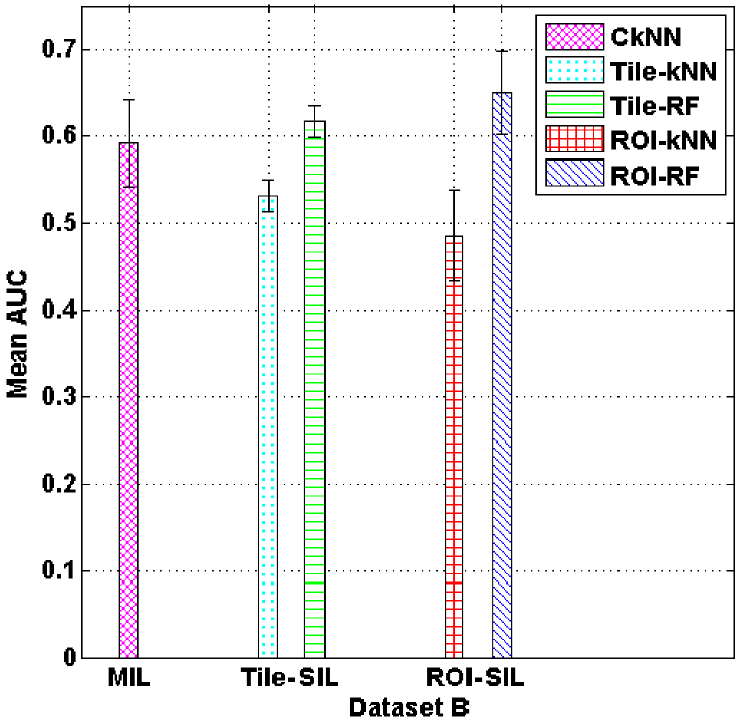


Figure 3.6: Comparison of CkNN with the tile-based and the ROI-based SIL using dataset B in CADx study.

Table 3.7 and Figure 3.6 show that we achieve mean AUC values of 0.592 with CkNN and 0.532, 0.617 with the tile-based kNN and RF respectively. The mean AUC values obtained with the ROI-based kNN and RF are 0.486, 0.651 respectively. For dataset B, we observe almost identical trends to those of dataset A but with decreased mean AUC values. This demonstrates that on this dataset the classification performance of kNN is equivalent to a random labelling of the instances, i.e., it cannot reliably discriminate the two classes.

B. CADe study: MIL as a screening tool

A comparison of the performance of CkNN as a screening tool for discrimination of malignant from benign lesions/normal tissue types with the tile-based single instance kNN is illustrated in Table 3.8. Figure B.3 (Appendix B) provides the average ROC curves for CADe study.

Table 3.8: Performance of CkNN in comparison to kNN in CADe study

Learner	**Mean AUC**
CkNN	0.621 ± 0.039
kNN	0.593 ± 0.017

The comparison of CkNN with the tile-based kNN indicates that CkNN yields a mean AUC value of 0.621 while, kNN achieves a mean AUC value of 0.593. A *t*-test shows that there is not a statistically significant difference between the performance of CkNN and the tile-based kNN on this dataset.

3.5 Discussion

Figure 3.3 shows the performance of CkNN against different tile sizes. We observe from Figure 3.3 that performance of CkNN is improved as the tile size is increased (or the number of instances per bag is decreased). The decreasing mean AUC with decreasing the tile size can be explained by two factors: 1) noise is being added to the feature vector when the size of tile is decreased, and 2) CkNN performs better with a smaller bag size [39]. As the tile size is further increased beyond 20×20, the performance of CkNN is reduced. This is because features extracted from tiles larger than 20×20 pixels provide more generalised and spatially irrelevant information that compromises the classification performance.

Figure 3.4 represents the comparison of CkNN with other MIL algorithms using breast MRI data. In particular, Figure 3.4 compares the performance of CkNN, where label of bag is determined without using explicit MIL assumption on the relationships of instances and bag labels, to the other MIL algorithms which follow strict MIL assumption. These experimental results indicate that, methods which strictly follow standard MIL assumption (e.g., MI-SVM and EM-DD, and APR) do not give comparable results to CkNN on the breast MRI data. Lin Dong draws a similar conclusion from his research. He presents a comparison of MIL algorithms on 16 datasets in [48]. He finds that learners which discard MIL assumption give better results on many datasets. Moreover, we know that APR algorithm is specially designed to solve the drug discovery problem [39]. Also, it involves a lot of parameters, which are hard to set properly for other datasets. Poor generalizability of APR to the breast MRI data could be another possible reason for the poor classification result.

We have presented the comparison of CkNN with the tile-based and the ROI-based SIL using datasets A and dataset B in Figure 3.5 and Figure 3.6 respectively. We observe three main trends from Figure 3.5 and 3.6. The first is related to the comparison of the performance of CkNN and the tile-based single instance kNN and RF. The performance of CkNN is better than the tile-based single instance kNN but equivalent to RF. Average ROC curves in Figure 3.7 also demonstrate the similar trend. Moreover, classification performance of the tile-based single instance RF is better than the tile-based single instance kNN. The reason for this is likely the robustness of RF and its inherent internal feature selection process.

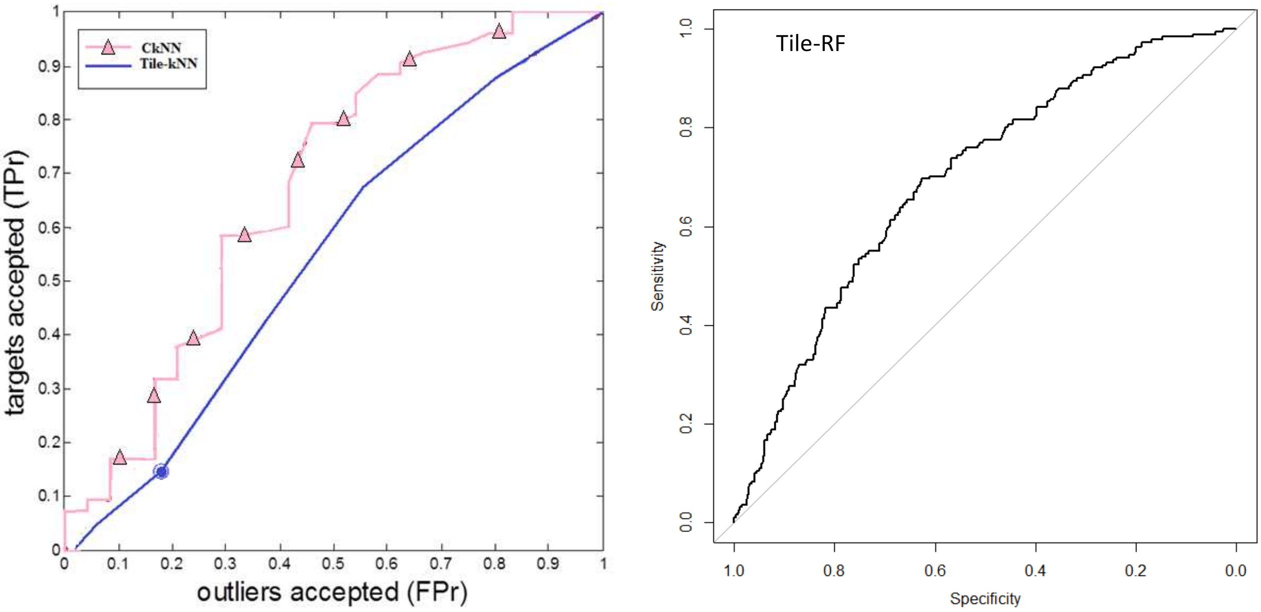


Figure 3.7: Average ROC curve of CkNN in comparison to the tile-based SIL using dataset A.

The second trend is based on the relative classification performance of the tile-based and the ROI-based SIL. From Figure 3.5, we observe that, the ROI-based SIL gives better classification performance as compared to the tile-based SIL. We also observe similar trends in Figure 3.6 but with the reduced mean AUC values. Also, ROI-based classification using kNN gives an apparent random labelling of samples. The reason for the relatively poor performance of the tile-based SIL as compared to the ROI-based SIL can only be the difference of the features in both cases. For the ROI-based SIL, we have used state-of-the-art T2 features, which are application dependant; while the tile-based SIL is based on simple, generic features.

The third trend is associated with the comparison of the performance of CkNN with the ROI-based kNN and RF. On dataset A, CkNN has equivalent performance to the ROI-based kNN while, on dataset B, the performance of CkNN is significantly better than the ROI-based kNN. On the other hand, the ROI-based RF gives slightly better performance as compared to CkNN. The better performance of the ROI-based single instance RF as compared to CkNN can be explained by two possible reasons: difference of the features used in the performance estimation of CkNN and the ROI-based single instance RF and the difference of classifier. RF is a robust classifier as compared to kNN [96] with an internal feature selection capability.

Statistically, there is not a significant difference between the performance of CkNN and the ROI-based single instance RF. Thus MIL, as a ‘pure’ machine learning technique, has equal potential for diagnostic classification for breast cancer MRI even when using generic tile-based features. The tile-based generic features have four main advantages over the ROI-based features. First, they are computationally easy to calculate. Second, they do not depend on an accurate segmentation of the lesion. Third, the extraction of generic features does not require extensive knowledge of breast cancer MRI. And fourth, the extraction of generic features allows the screening for suspicious lesions without delineating the lesions. On the other hand, the ROI-based features depend on an accurate delineation of lesions, thus, cannot be used for screening purpose.

Furthermore, the improved classification performance of all the learners on dataset A as compared to dataset B indicates that dataset containing mass-like lesions (dataset A) is easier to classify into benign and malignant lesions as compared to the dataset containing both mass-like and non-mass-like lesions (dataset B).

For the tile-based classification, we have not considered the curse of dimensionality. The nearest neighbour learners are particularly sensitive to the curse of dimensionality [39] especially with small *k* values. High dimensionality is not a problem with RF due to distribution of randomised subsets of features and samples among decision trees [86]. Therefore, it is reasonable to ask that if we do feature selection before classification, is it likely that we will get better results? To analyse this question, we utilize the tile-based features selected by RF based on the importance coefficients returned by RF. For simplicity, we use these SIL features for MIL classification. We acknowledge that the results may be pessimistically biased as we select the subset of features in the SIL domain for the performance estimation of MIL domain (CkNN). However, for the variants of kNN, the results are not classifier biased, as important features are selected using RF and tested using kNN.

We determine the optimal number of features by applying the rule of thumb as recommended by Jain, et al., [49]. For dataset A, we have 53 positive and 24 negative samples. After dividing the data into training and test set using stratified 10-fold CV, we get a training set consisting of $(53 - 5\ =)$ 48 positive and $(24 - 2 =)$ 22 negative samples. Therefore, the optimal number of features using the rule of thumb is determined as:

$$Optimal\ No. of\ features = \llcorner \frac{22}{10} \lrcorner\ = 2$$

Similarly, for dataset B, the optimal number of features for CADx and CADe studies is 5 and 7 respectively. For simplicity, we select $\ulcorner \frac{2+5+7}{3} \urcorner\ = 5$ features for each study. The selected features for datasets A and B are presented in Table 3.9.

Table 3.9: Five important tile-based features

Dataset	**5 Selected Features from Table** 3.4
A (mass-like lesions)	1,2,4,5,7
B (both mass-like and non-mass-like lesions)	1,2,4,5,6

From Table 3.9 we observe that, the tile-based feature set after feature selection for dataset B contains features mainly in the axial direction. One possible reason is that many malignant lesions are directed parallel to the duct system in the axial direction. If lesions have bigger diameter in axial direction as compared to transverse direction, the lesion is more likely to be malignant [97,98]. This indicates that MIL with generic features can find physiological relevant information as well.

After features selection, we re-evaluate the classification performance of CkNN and the tile-based SIL using kNN and random forest in CADx study. Figure 3.8 and Figure 3.9 present a comparison of the classification performance of CkNN and the tile-based single instance kNN and RF before and after feature selection for datasets A and B respectively. From Figure 3.8 and Figure

3.9, we observe that there is little improvement in the performance of the tile-based single instance kNN after feature selection. The performance of the tile-based single instance RF before and after feature selection is almost the same (as expected).

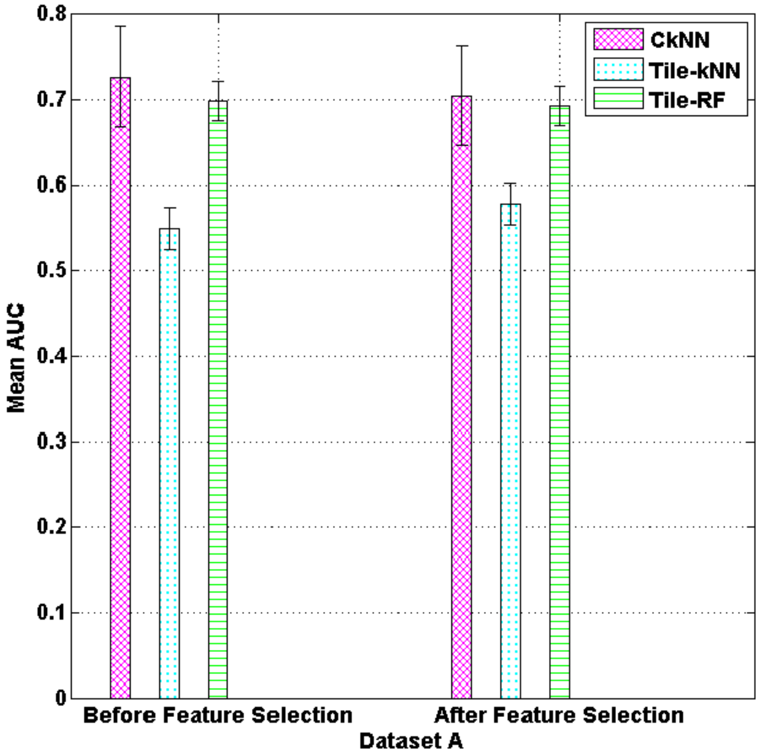


Figure 3.8: Comparison of the learners before and after feature selection using dataset A.

From Figure 3.8, a comparison of the performance of CkNN for the mass-like lesions before and after feature selection indicates that performance of CkNN is degraded slightly after feature selection. But, this decrease in performance is not significant. To explain the possible reason for the decrease in the performance of CkNN after feature selection, we believe that, generic features in both axial and transverse direction are equally important for the classification of mass-like breast lesions. However, Figure 3.9 shows a significant improvement in the performance of CkNN after feature selection, which goes from mean AUC value of 0.592 to 0.701. The improved performance of CkNN is evidence for its efficacy in discriminating benign and malignant lesions even for a more challenging dataset containing both mass-like and non-mass-like lesions.

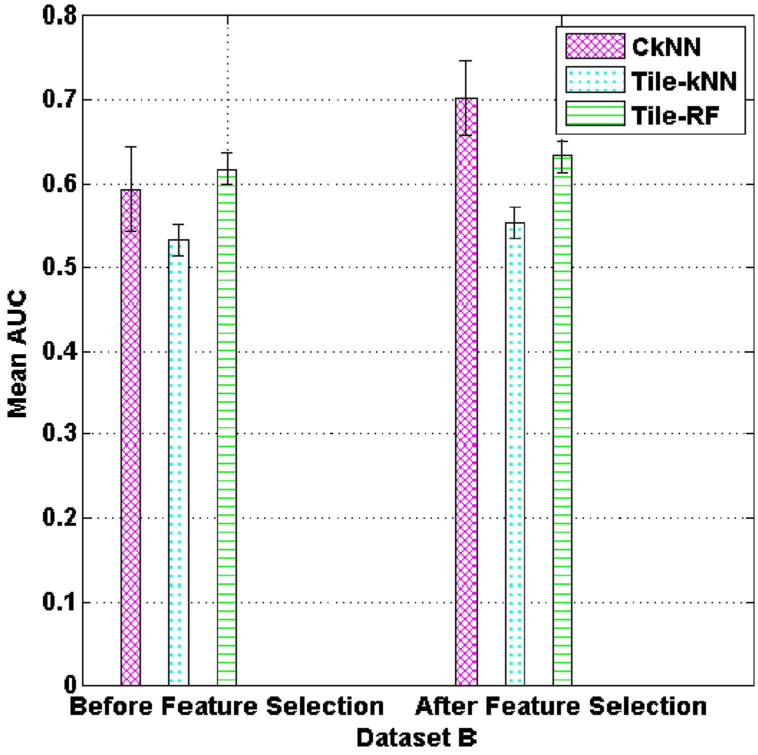


Figure 3.9: Comparison of the learners before and after feature selection using dataset B.

Encouraged by the improved performance of CkNN after feature selection in CADx study, we repeat CADe study with the new set of features. The comparison of results before and after feature selection, as illustrated in Figure 3.10, confirms improved performance of CkNN after feature selection. The results after feature selection indicate that, if a feature selection step is added before classification, MIL can also be a suitable choice to assist radiologists in breast cancer screening. Thus, MIL has potential to highlight the suspicious regions without delineating the ROI if it is trained over malignant, benign and normal tissue types. The improved performance of CkNN after feature selection shows that feature selection is an important step in MIL to avoid the curse of dimensionality. So far, very few feature selection techniques in MIL have been proposed. Therefore, we develop a MIL-based feature selection algorithm in Chapter 4.

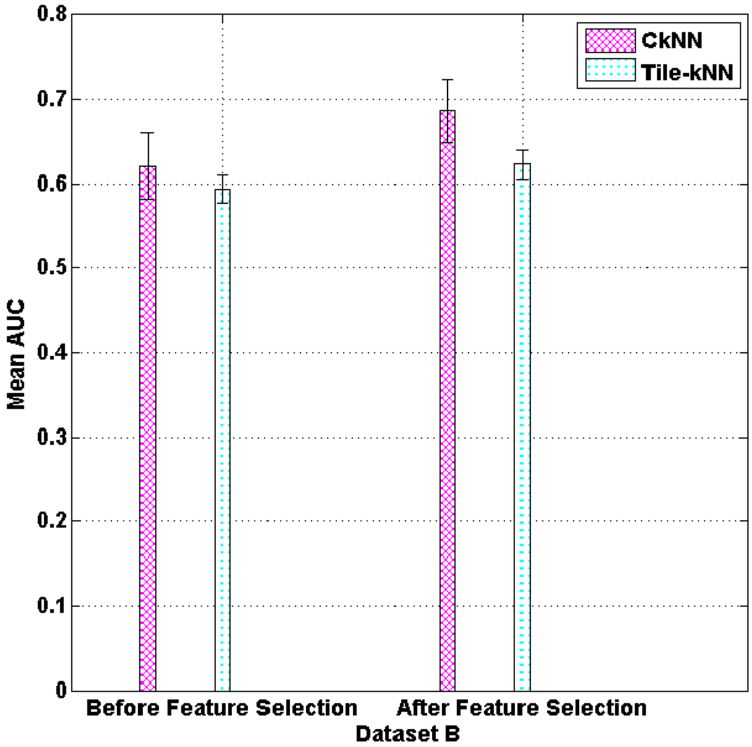


Figure 3.10: Comparison of the learners before and after feature selection in CADe study.

3.6 Conclusion

In this chapter we have first developed an insight into the working of CkNN using synthetic datasets on an experimental basis. We have also used the similar methodology to optimise the parameters of CkNN for the classification of breast MRI. Empirical results indicate that the performance of CkNN is robust to a wide range of [*k*, *c*] values. After studying working of CkNN on synthetic datasets, we have evaluated the classification performance of CkNN using breast cancer MRI in both CADx and CADe applications. The initial experiment for the selection of tile size shows that the performance of CkNN is affected by the tile size. A small tile size gives a noisy feature vector while a big tile size represents spatially irrelevant features which compromise the performance of CkNN. A comparison of the performance of CkNN with other MIL learners indicates the better performance of CkNN. This demonstrates the greater classification efficacy of CkNN on breast MRI data and supports our choice of lazy learner.

The empirical results of the CADx study indicate that CkNN performs significantly better than the tile-based kNN. Moreover, the ROI-based SIL based on state-of-the-art T2-w features is better than the tile-based SIL. Also, performance of CkNN is statistically equivalent to the ROI-based SIL. However, tile-based classification using CkNN does not require domain specific features and is robust to the inaccuracies in the segmentation of suspicious lesions. Therefore, CkNN is a suitable choice for the classification of T2-w breast cancer MR images. Results of CADe study indicate that CkNN also has the potential to be used as a screening tool in breast MRI by eliminating the need for the delineation of suspicious lesions. The improved performance of CkNN after feature selection emphasizes the continued importance of feature selection in multiple instance learning.

Here we have evaluated the performance of CkNN using generic spatial features alone from T2-w MRI. In the next chapter, we estimate the performance of CkNN using both spatial and temporal (spatio-temporal) generic features derived from both T2-w and DCE-MRI.

Chapter 4

Multiple instance learning for breast MRI using generic spatio-temporal features

In this chapter we evaluate the performance of multiple instance learning (MIL) as a ‘pure’ machine learning approach for the classification of breast cancer. Specifically, we use generic tile-based features derived from T2-weighted MRI and T1-weighted DCE-MRI. Since, DCE-MR images are obtained at several time points, so we not only get spatial image information but also information in time domain (i.e., temporal). Therefore, we particularly investigate the importance of both spatial and temporal (spatio-temporal) features for the differentiation of benign and malignant lesions. We utilize parametric models of contrast enhancement and a discrete cosine transform (DCT) as feature extraction techniques. We compare generic tile-based features against ROI-based features. We also compare the results with those of Chapter 3, where we have used MIL for the classification of non-contrast enhanced MRI using generic tile-based spatial features alone. The core material presented in this Chapter forms the basis of a conference paper [99] which is published in the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

Remainder of the chapter is organised as follows. In Section 4.1, we present a brief introduction to the types of contrast enhancement curves in model fitting. Here, we talk specifically about the difference of the trends that malignant and benign lesions show in model fitting. In section 4.2, we present the experimental methodology to evaluate spatio-temporal features. Here we talk about datasets, learners, and features for the classification of breast MRI. We also perform an initial experiment to determine the suitable tile size for the evaluation of spatio-temporal features. In section 4.3 and 4.4, we present the results and discussion. Finally, we conclude this chapter in section 4.5.

4.1. Introduction

As mentioned in Chapter 1, the sensitivity of breast MRI has been improved with the use of contrast agent in DCE-MRI. DCE-MRI has an advantage that it provides a good contrast between suspicious lesions and normal tissue types. Also, the use of T2-w MRI is standard in clinical interpretation of breast cancer MRI. Therefore in this chapter, we use both DCE-MRI and T2-w MRI to empirically evaluate the performance of multiple instance learning. Specifically, we use generic tile-based spatio-temporal features for the discrimination of benign and malignant lesions.

As described in Chapter 1, DCE images are acquired before and at several time points after the injection of contrast agent. An important criterion for the diagnosis of breast cancer in DCE-MRI is a signal intensity change over time [100]. Fitting parametric models (empirical or pharmacokinetic models) of contrast enhancement to these time points result in contrast enhancement curves (time-intensity curves). In Figure 4.1, three basic curve shapes have been shown. These time-intensity curves can be summarised in two phases. Early upslope phase (post-contrast rise) depicts wash-in characteristics, which are often categorised as slow, inter-mediate and rapid. Delayed phase can be categorised as persistent (or curved) identified by a positive slope, plateau (zero slope), and washout (negative slope). Most malignant lesions in the post-contrast images depict an early stronger enhancement (initial fast upslope) with a rapid washout. Moreover, lesions which depict a plateau in the delayed phase are mostly considered suggestive of malignancy too. On the other hand, a slow enhancement in the initial phase followed by a persistent or curved enhancement in the delayed phase is a characteristic of most benign lesions [101,102].

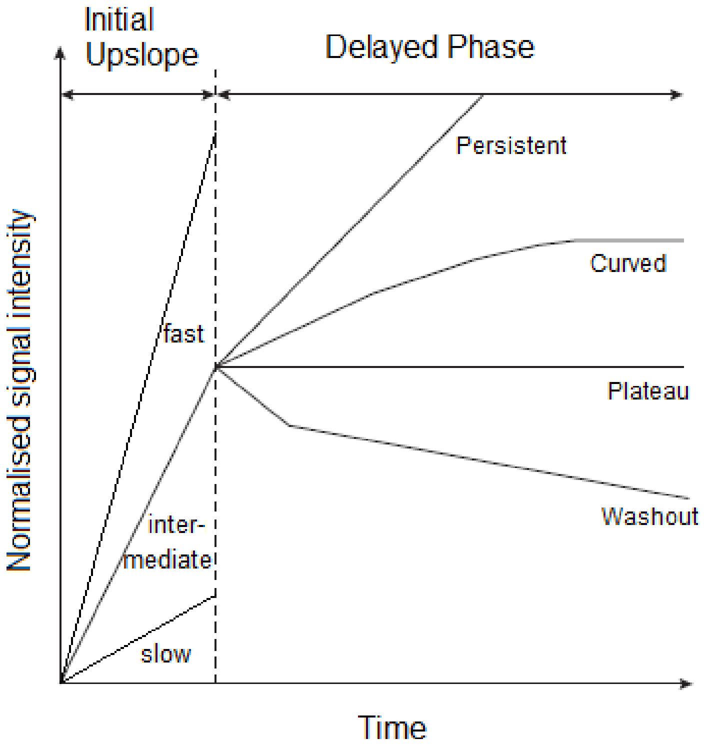


Figure 4.1: Types of contrast enhancement curves [101].

We can extract features from time-intensity data using model free approaches, empirical models and pharmacokinetic models. In model-free approaches, we extract information (time to peak, intensity of peak etc.) from time-intensity data without fitting any model to the data [103]. Empirical models are simple models which describe the time-intensity curves typically with between two and four parameters. Linear slope [112], simplified gamma variates [104], Ricker [105], Gal [113], and Agliozzo [106] models are examples of empirical models. Pharmacokinetic models are based on compartmental flow assumption, i.e., these models assume that the contrast agent is distributed between two main tissue compartments (the intra-vascular plasma volume space and the extra-vascular / extra-cellular volume). Examples of these include Tofts model [107], Brix model [108] and Hayton model [109].

Model-free approaches do not make any assumptions about the underlying signal but require high temporal resolution to get accurate parameters. Pharmacokinetic models involve a large number of fitting parameters and require high temporal resolution. Empirical models, on the other hand, involve few parameters and are suited to low temporal resolution acquisitions [92]. In this chapter, we utilise generic empirical models of contrast enhancement to extract temporal features from DCE-MRI.

In addition, we utilise a discrete cosine transform (DCT) [110] as a feature extraction technique. We know that good performance of a classifier is dependent on high discriminatory power of the features representing an image [83]. Discrete cosine transform is a generic technique which transforms a signal (i.e., image here) from spatial domain to frequency domain. DCT has the property to concentrate the visually significant information in just few low frequency coefficients. Since low frequency coefficients preserve the important information, we assume that these low frequency coefficients can provide high discrimination between benign and malignant lesions.

4.2. Experimental Methodology

A. *Dataset*

Here we use ‘dataset A’ previously used in our work on the diagnosis of breast cancer MRI in Chapter 3. This dataset consists of 53 malignant and 24 benign mass-like lesions. In Chapter 3, we have considered a block of 60×60 pixels from T2-weighted (T2-w) MRI overlapping with a manually segmented ROI as the bag and the features extracted from $n \times n$ tiles as instances. Here, we extend this work to both T2-w MRI and DCE-MRI.

B. Algorithms

In order to estimate the classification performance of MIL on breast MRI data using spatio-temporal features, we evaluate the performance of CkNN in comparison to kNN. In addition, we compare the performance of CkNN with the results of a more conventional approach described in [92,111], because this study was performed on the same dataset. This study investigates the discriminatory power of state-of-the-art ROI-based features from multi-modal MRI using a Random Forest (RF) classifier. In particular, we compare the results with the relevant results of study 2 based on DCE-MRI alone and DCE-MRI combined with T2-w MRI.

In Chapter 3, we have used CkNN for the selection of tile size. But to avoid over-fitting to the data, we have utilised different feature sets for both the selection of tile size and the performance estimation of CkNN. Here to avoid over-fitting to the data, we select the tile size on a SIL-based learner, i.e. a Random Forest classifier, rather than CkNN. If we use CkNN for the selection of tile size and then estimate its performance using the same dataset, the performance of CkNN would be optimistically biased. Hence, we select a SIL-based RF for the selection of tile size. We know that RF is a robust classifier which is not affected by the curse of dimensionality [86]. Therefore, we select the tile size without reducing the dimensionality. We acknowledge that, selecting the tile size with a SIL-based learner may add negative bias to the performance of a MIL-based learner. However, due to limited data we cannot split the data independently for both the selection of tile size and the performance estimation of CkNN.

C. Features

In the previous chapter, we have proposed a generic tile-based MIL approach for the identification and classification of non-contrast enhanced breast cancer MRI. In particular, we have used modified generic features from image database to discriminate benign and malignant lesions. Here we extend this approach to DCE-MRI and include both spatial and temporal features. We utilise a DCT to extract generic tile-based features. DCT has the advantage that it allows us to extract generic tile-based features in multi-dimensions. To extract generic tile-based spatio-temporal features, we first expand the 60×60 sub-image by reflecting boundaries to make it 64×64 pixels block in case of T2-w MRI and 64×64×4 block in case of DCE-MRI. We then decompose 64×64 blocks of images into independent n×n tiles for T2-w MRI and n×n×4 cubes for DCE-MRI. These tiles/cubes represent the instances in each 64×64 or 64×64×4 bag.

To obtain this spatio-temporal information, we evaluate two approaches: 1) using a 3D-DCT on DCE-MRI alone, which gives spatial information combined with temporal i.e., spatio-temporal.

2) using a 2D-DCT on T2-w MRI to obtain spatial features plus three parameters each from a linear slope model [112] and an empirical model of contrast enhancement (Gal model) [113] to obtain the temporal features from DCE-MRI.

We have chosen these above mentioned parametric models because they are generic in nature. These models are not derived from pharmacokinetics, i.e. these models do not make assumptions about the relation of concentration of contrast agent and intensity (two compartment flow). Also, the parameters of these models are independent of the density and nature of tissue type [113].

The equations and initial estimates of parameters for each of the above mentioned model are given in Table 4.1. Variable t and y respectively show the time and intensity since the injection of contrast agent. Both of the selected models are not linear in their parameters and hence linear least squares algorithm cannot be used to fit these models. We use MATLAB implementation of Trust-Region algorithm, a non-linear fitting algorithm, to fit both models to the enhancement curves of each tile. The three parameters from linear slope model are initial slope $(\beta 1 = y/t)$, delayed slope (β2), and the time point where these two slopes meet (α). The three parameters from Gal model are $a,\ b, and\ c$, which are positive, real parameters [113].

Table 4.1: Parametric models of enhancement

Parametric Model	**Equation**	**Initial Parameter Estimates**	**Bounds**
Linear Model	$f(t) = \begin{cases} \beta 1 t & if\ t \leq \alpha \\ \beta 1 \alpha + \beta 2(t - \alpha) & if\ t > \alpha \end{cases}$	$\beta 1 = y2/t2$ $\beta 2 = 0, and\ \alpha =$ $t2\ (the\ first\ post -$ $contrast\ time$ [114]	$-\infty < \beta 1 < \infty$ $-\infty < \beta 2 < \infty$ $-\infty < \alpha < \infty$

Gal Model	$f(t) = a.t.e^{-\frac{t^c}{b}}$	A = 1.2, b = 3.2,
		c = 0.8
			$a \in (0{,}200)$
			$b \in (0{,}100)$
			$c \in (0{,}3)$

We want to fit the model to the data with an approach which is both computational efficient and robust to the noise. We know that each tile contains n^2 pixels. Fitting a non-linear model to every pixel in a tile is time consuming and involves extensive computations. Pixel-wise model fitting is not only computationally expensive, but also sensitive to the noise, specifically to the motion artefacts [114]. Moreover, when we fit the model on a pixel-by-pixel basis, we get n^2 values of each parameter in a tile. To summarise each tile by a single feature vector as required by a MIL dataset, we need to average each parameter value over n^2 values to represent an instance. Summarising the tile with an average value of each feature may reduce the effect of noise in a tile, but not the computational complexity.

To reduce the computational complexity, we do not fit the enhancement models pixel-wise, but rather to the relative enhancement based on the mean of each $n \times n$ tile from DCE-MRI stack. Relative enhancement ensures that we get the starting point of enhancement curves at (0,0). The equation for relative enhancement for DCE-MRI stack is given below.

$$R(i) = \frac{y(i) - y(1)}{y(1)}$$

where

$R(i) = \text{Relative enhancement for the i}^{\text{th}}\ \text{DCE} - \text{MRI stack},$

$y(i) =$ mean of $n \times n$ tile intensity in the i$^{\text{th}}$ stack, and

$y(1)$ = mean of $n \times n$ tile intensity in the pre-contrast stack.

Averaging the pixel values in a tile also averages out the noise in the tile. We can also fit the model to the median, maximum, or some percentile of the pixel values in a tile. But, taking the average of a tile has the advantage that it reduces the effect of noise in a tile by averaging over all pixels. Please note that, fitting the model over n^2 points (not on pixel-by-pixel basis) is also a computationally efficient method. In this case, model would not fit exactly to the data points but to the average of the points, which overcomes the effect of noise in a tile. However, we believe that, model fitting to the mean of tile is the simplest way and is no worse than fitting to n^2 points.

Example curves of both models, when fitted to the mean of a tile, are shown in Figure 4.2. The curves for benign and malignant tiles resemble to the corresponding trends in Figure 4.1. Therefore, we believe that these generic models can give domain specific temporal features. Thus, the use of selected generic models for the tile-wise (mean of tile) rather than the pixel-wise fitting is justified.

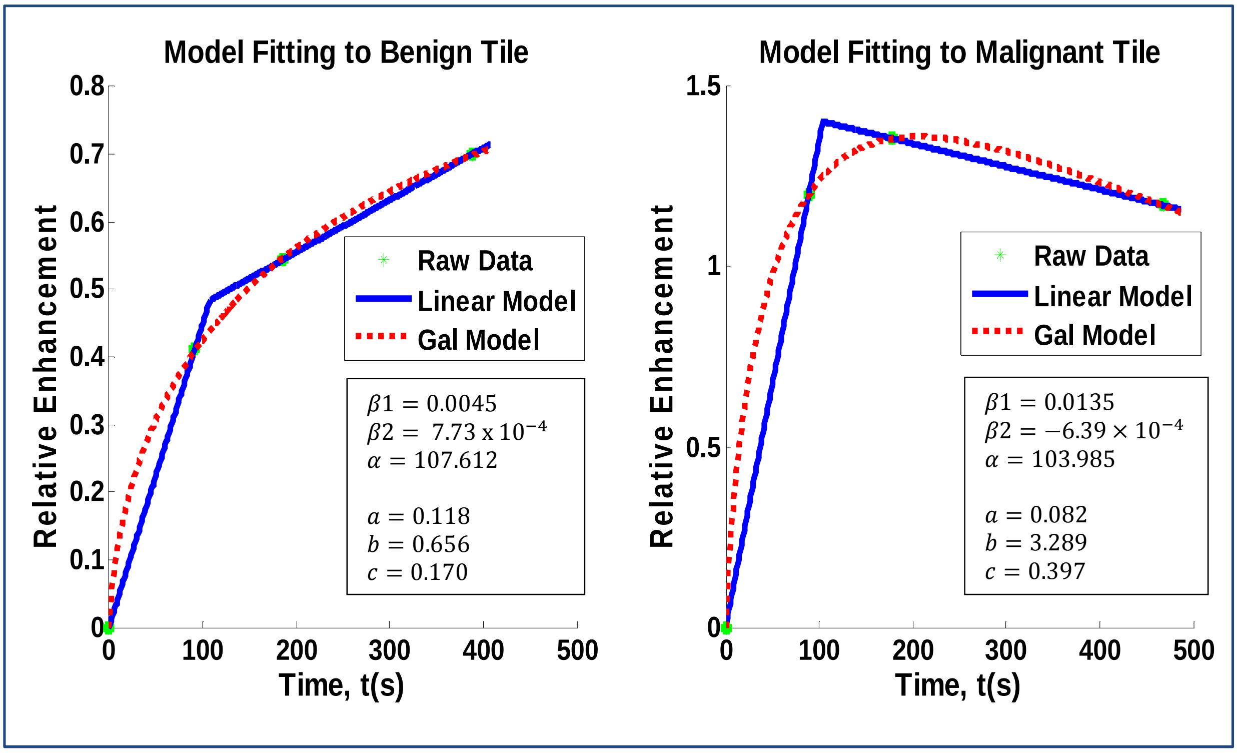


Figure 4.2: Example curves of model fitting to the mean of a tile.

In Figure 4.2, we also provide parameter values we get, after fitting these models to a benign and a malignant tile. We observe that the parameters of a linear slope model are quite interpretable. For example, the product $\alpha\beta1$ shows the height at the joint point. This product gives measure of degree of enhancement in the initial phase [114]. Initial slope $\left(\beta1 = \frac{y}{t}\right)$ has positive value which is indicative of an enhancing lesion. $\beta1$ for a malignant tile is greater than that of a benign tile. For a given time, greater value of $\beta1$ depicts a faster upslope, which is the characteristic of a malignant lesion. Similarly, $\beta2$ indicates the nature of the enhancement in the delayed phase. For example, positive and negative values of a delayed slope ($\beta2$) indicates the continued uptake and wash-out characteristics respectively. We know that, a washout (negative slope) or plateau (zero slope) slope is commonly observed in the malignant lesions, while positive slope characterises the benign lesions.

We can interpret the parameter values from Gal model as well. Gal model is a product of three factors. The first factor is a. It is the scaling parameter. The second factor is t, which ensures zero value of the model when $t = 0$. The third factor is $e^{-\frac{t^c}{b}}$, which determines the shape of the curve. Parameter b influences the slopes in initial and delayed phase by determining the width of Gaussian function, while parameter c influences the difference between two slopes [115]. Gal model approximates a straight line when parameter $c \approx 0$. The peak of the curve (which represents the point of maximum intensity) lies where the derivative of model is zero [115]:

$$f'^{(t)} = a.e^{-\frac{t^c}{b}}\left(1 - e^{\frac{t^c.c}{b}}\right) = 0$$

$$tpeak = \left(\frac{b}{c}\right)^{\frac{1}{c}}$$

The value at the peak is:

$$f(tpeak) = a\left(\frac{b}{c.e}\right)^{\frac{1}{c}}$$

Example curves of both models when fitted to the median and maximum of a benign and malignant tile are shown in Figure 4.3. For a comparison purpose, model fitting to the mean of the tiles is also presented. We can observe that, model fitting to the maximum of a tile does not present reasonable characteristics of benign and malignant lesions. For example, curves fitted to the benign tile depict the characteristics of a malignant lesion (i.e., a fast upslope in the initial phase and a washout in the delayed phase). Similarly, enhancement curves fitted to the maximum of the malignant tile show slow wash-in characteristics as compared to when fitted to the mean or median of the tile.

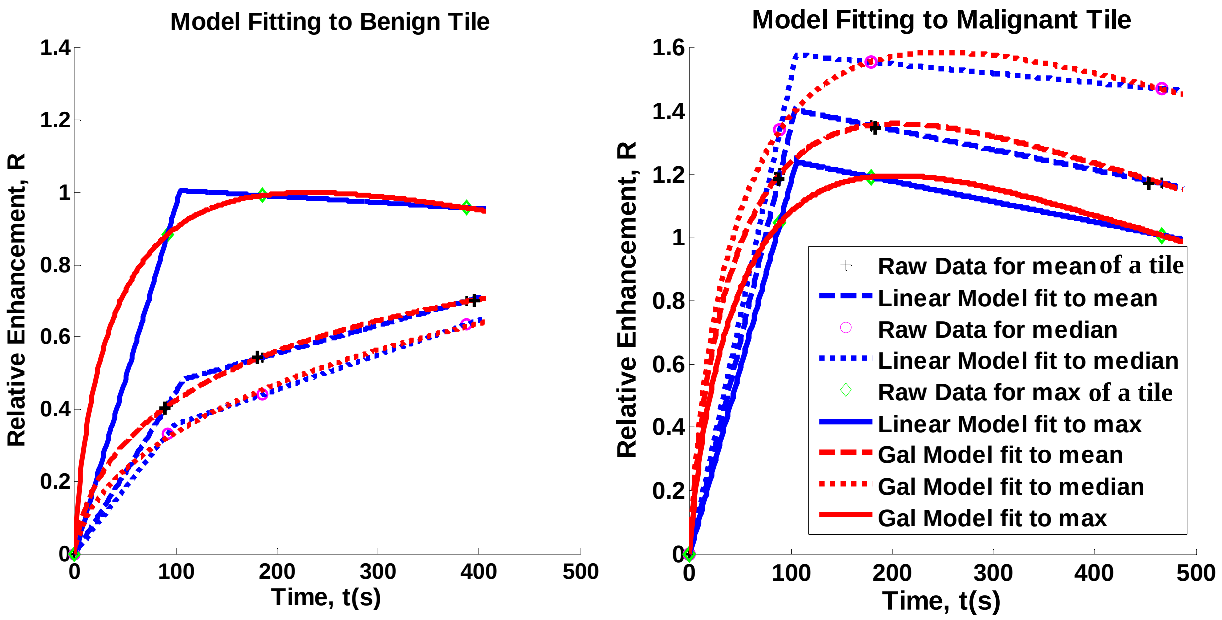


Figure 4.3: Example curves of model fitting to the mean, median, and maximum of a tile.

There is not much difference between curve fitting to the mean or median of the tiles. However, model fitting to the mean of the malignant tile is slightly better than that to the median of the tile. For example, we get a steep washout slope for malignant tile when fitted to the mean as compared to when fitted to the median. Also, we have already mentioned that fitting the model to the mean of tile has the advantage of reducing the effect of noise in a tile by averaging pixel values in a tile.

While using the DCT as a feature extraction technique, we apply a 3D-DCT to each *n*×*n*×4 cube of DCE-MRI and a 2D-DCT to each *n*×*n* tile of T2-w MRI. Since, a DCT concentrates the energy of the original data in only a few low frequency coefficients; we select 25 coefficients in a 3D-zigzag traversal [116]. A 3D-zigzag traversal is an extension of 2D-zigzag traversal, where coefficients are scanned in increasing order based on similar sum of their indices $i.e., (x + y + z)$; where *x* and *y* represents the spatial indices and *z* represents the third dimension (temporal here). By extracting coefficients in a zigzag manner, the correlation between the coefficients is minimized and we extract the coefficients in increasing order of frequency; the smaller the sum, the lower the frequency [117]. The location of 25 extracted 3D-DCT spatio-temporal coefficients (features) in a 3D-zigzag traversal for a standard 8×8 tile size in the four time points is shown in Figure 4.4.

For spatio-temporal features from DCE-MRI and T2-w MRI, we extract 25 features by combining the first 19 2D-DCT coefficients extracted in a zigzag transversal with the six temporal features extracted from DCE-MRI. Please note that here in feature set, the first 19 features are 2D-DCT spatial features while the rest are temporal features. A 2D-zigzag traversal for a standard 8×8 tile is shown in Figure 4.5. The distribution of extraxted 19 2D-DCT spatial features are also shown.

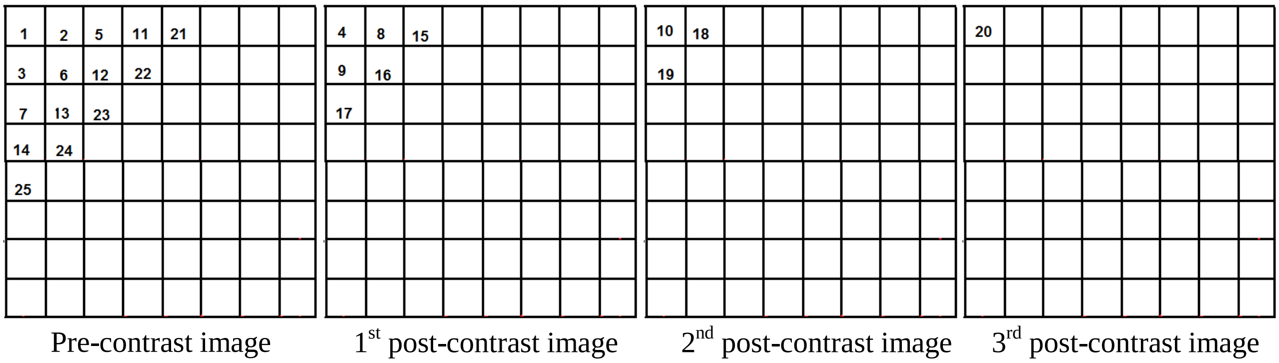


Figure 4.4: 25 extracted 3D-DCT spatio-temporal features in a 3D zigzag traversal.

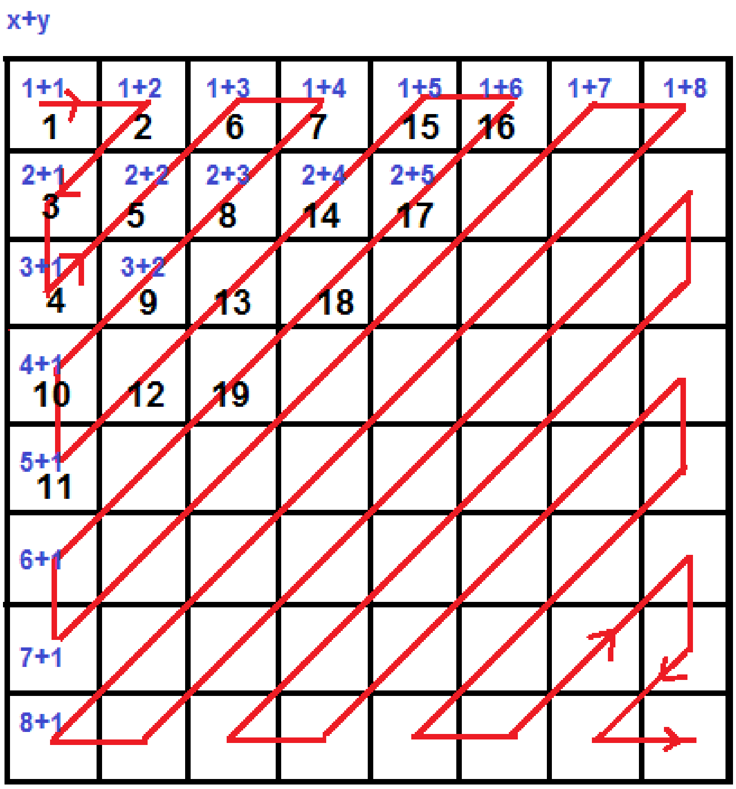


Figure 4.5: A 2D-zigzag traversal.

D. Tile size selection

In a tile-based approach, tile size represents a feature vector. Inappropriate tile size can give noisy feature vector which can compromise the classifier's performance. Therefore, we perform an initial experiment to determine the suitable tile size. We evaluate the performance of RF for independent tile sizes of order 2^n (where n = 2, 3, 4, 5, and 6), using 25 coefficients from 3D-DCT. For simplicity, we also use the selected tile size to compute 2D-DCT plus temporal features from T2-w and DCE-MRI. The performance of RF for square tile sizes is shown in Figure 4.6.

From Figure 4.6 we can see that with a tile size of 4×4 pixels, RF performs best to classify benign and malignant tiles with a mean AUC of 0.949. However, this tile-size corresponds to 256 independent (non-overlapping) instances in a bag of size 64×64 pixels. As mentioned in the section 2.2, the computational complexity of CkNN is $O(m^2Nd)$, where m is the average number of instances in a bag, N is the number of training examples (bags), and d is the number of dimensions. This means that for a fixed size of training set (N) and a fixed length of feature vector (d), CkNN has an order of m^2 time complexity. This bag size of 256 instances is not only computationally expensive for the performance estimation of CkNN but also for model fitting. Therefore, for computational efficiency, we use a tile size of 8×8 pixels for the performance estimation of the variants of kNN.

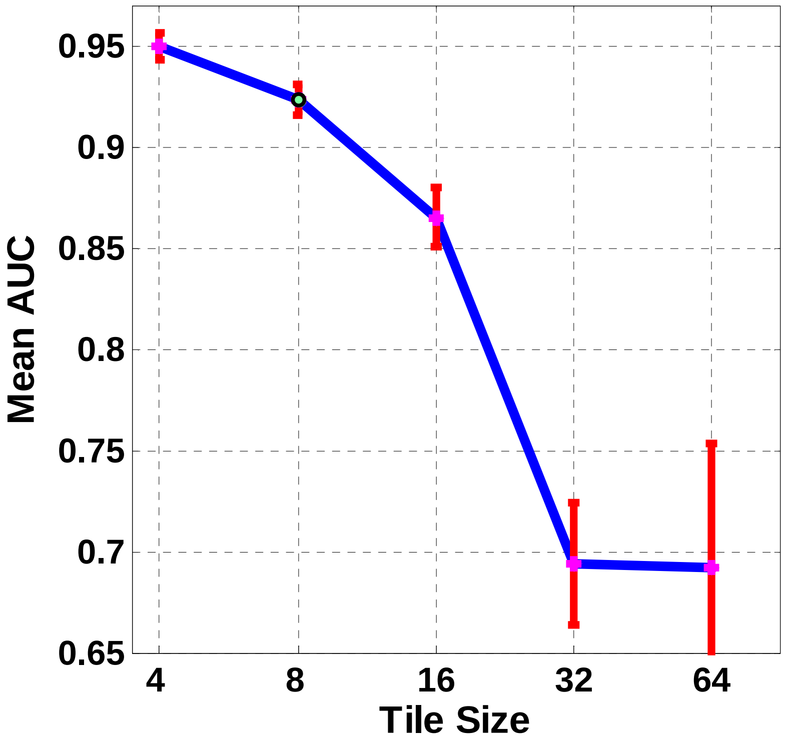


Figure 4.6: Mean AUC with RF for square tile sizes.

With an 8×8 tile size, RF achieves a mean AUC of 0.923. Although, the performance of RF with an 8×8 tile size is not as good as that with a 4×4 tile size, but this tile size requires fewer computations for the performance estimation of CkNN. Also, choosing the 8×8 tile size involves model fitting to relatively less number of instances (64 instances) in a bag as compared to the tile size of 4×4 (256 instances). Moreover, since we fit the model to the mean of a tile, we believe that averaging over 8×8 pixels reduces more effect of noise in a tile as compared to averaging over 4×4 pixels.

E. Feature selection

For fair comparison with the results of Chapter 3, we select 5 features only. In order to reduce the dimensionality from 25 features to just 5 features, we utilise a *plus-l-take-away-r* algorithm [118]. We do not select features using an *exhaustive search*. The reason is that, exhaustive search involves computation of combinatorial subsets of features. For example to select 5 features out of 25 features, an exhaustive search would evaluate $\frac{25!}{(25-5)!\times 5!} \approx 53$ thousands of feature subsets. Moreover, plus-l-take-away-r algorithm has advantage over other feature selection methods, i.e., sequential forward selection (SFS) and sequential backward selection (SBS) algorithms. For example, *SFS* successively adds one feature at a time which in combination with the selected features maximises the criterion function. However, once the feature is selected, it cannot be discarded. Similar is the problem with *SBS* method in bringing back the feature into the optimal subset after deleting it. *Plus-l-take-away-r* method resolves this problem of feature subset nesting by adding *l* features using *SFS* and deleting *r* features using *SBS* [49].

We know that kNN gives good classification performance when the size of training dataset is infinitely large. For example, nearest neighbours based learners give probability of error which is very close to (less than twice) the Bayes probability of error [119]; which is the lowest possible error rate for a given classification problem. Due to the limited data, we do not select features on the basis of learner's (CkNN's) performance. Rather, we select features in the both MIL and SIL-based distance functions. Specifically, we use minimum Hausdorff distance for the MIL-based feature selection and minimum Mahalanobis distance for the SIL-based feature selection. We select features on the training data and estimate the performance on independent test data via a 10-fold cross validation (CV) scheme [51]. In other words we treat the selected feature set as a parameter of kNN-based learners in each fold. The performance measure utilized is mean AUC [54]. Due to the limited amount of data, we optimize the parameters of learners used in the 3D-DCT by randomly selecting coefficients from the 2D-DCT on T2-w MRI. We believe that, this results in an unbiased estimate of parameter values because parameters are tuned solely on T2-w MRI which is entirely independent from T1-w DCE-MRI. However, this will bias the estimated performance of 2D-DCT plus temporal features based classification. The tuned parameters are presented in Table C (Appendix C).

4.3. Results

The selected five 3D-DCT spatio-temporal features are presented in Table 4.2. Table 4.3 and Figure 4.7 show a comparison of the classification performance of CkNN, kNN using the selected 3D-DCT spatio-temporal features, and traditional ROI-based classification with RF using just DCE-MRI [92,111]. The corresponding mean ROC curves using 3D-DCT features are presented in Figure D.1 (Appendix D).

Table 4.2: The five selected 3D-DCT features

Validation Fold	5 MIL-based Selected Features from Figure 4.4	5 SIL-based Selected Features from Figure 4.4
1st to 10th except 4th	1, 2, 3, 5, and 6	2, 3, 4, 5, and 6
4th	1, 2, 3, 6, and 7	

Table 4.3: Performance of the 3D-DCT tile-based features and traditional ROI-based features

Technique	Learner	Mean AUC
Tile-based MIL	CkNN	0.816 ± 0.047
Tile-based SIL	kNN	0.838 ± 0.010
ROI-based SIL [92,111]	RF	0.824 ± 0.046

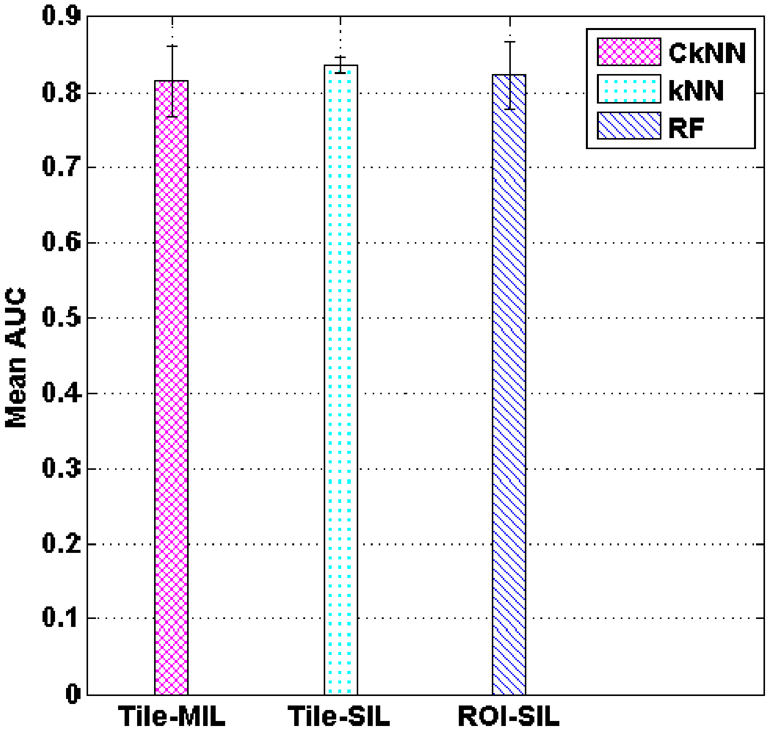


Figure 4.7: Performance of the 3D-DCT tile-based features and traditional ROI-based features.

Figure 4.8 presents the performance of CkNN and kNN with MIL-based feature selection in comparison to SIL- based feature selection using the 3D-DCT features. It can be seen that MIL-based feature selection is important for MIL-based classification (with CkNN). Similarly, SIL-based feature selection is important for SIL-based classification.

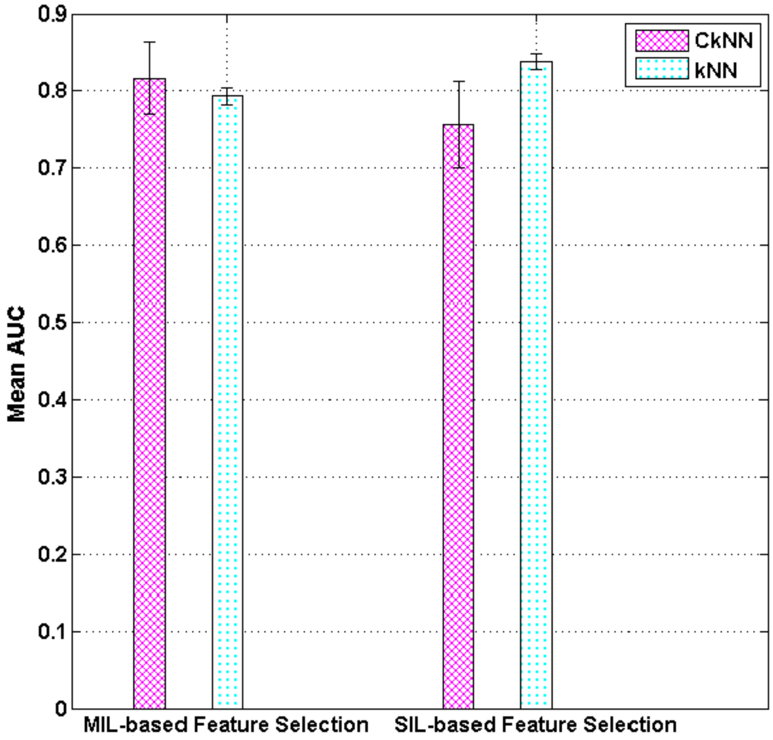


Figure 4.8: Performance of the learners with MIL-based feature selection in comparison to SIL-based feature selection using the 3D-DCT features.

The 2D-DCT plus temporal features selected in all validation folds are presented in Table 4.4. Table 4.5 and Figure 4.9 compare the classification performance of CkNN, kNN using the selected 2D-DCT plus temporal features against the traditional ROI-based classification using both DCE-MRI and T2-w MRI with RF [92,111]. Figure D.2 (Appendix D) shows the corrosponding mean ROC curves using 2D-DCT plus temporal features.

Table 4.4: The five selected 2D-DCT plus temporal features

5 MIL-based Selected Features	5 SIL-based Selected Features
1, 20, 23, 24, and 25	2, 3, 4, 5, and 6

Table 4.5: Performance of the 2D-DCT plus temporal features against traditional ROI-based features

Technique	Learner	Mean AUC
Tile-based MIL	CkNN	0.778 ± 0.052
Tile-based SIL	kNN	0.702 ± 0.012
ROI-based SIL [92,111]	RF	0.838 ± 0.045

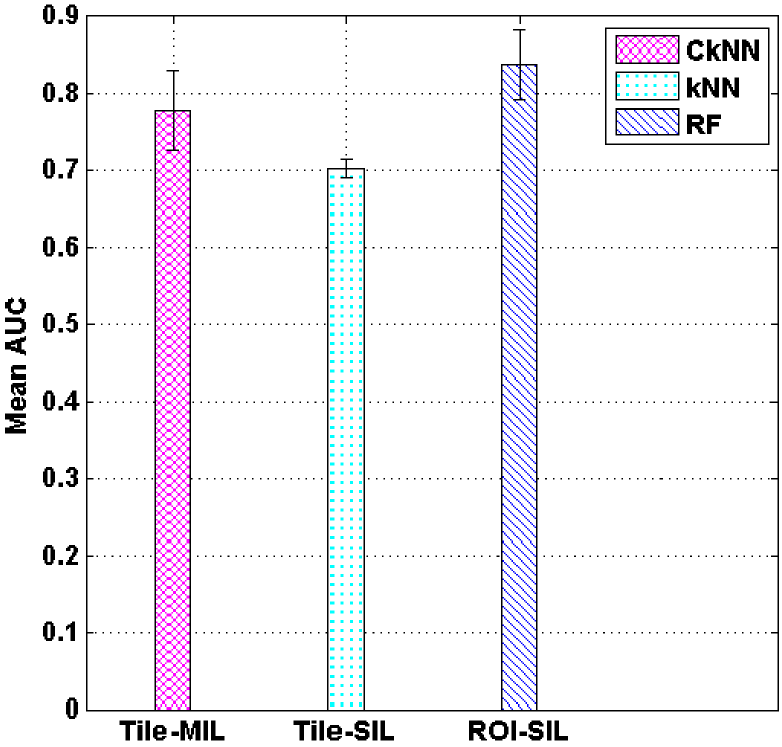


Figure 4.9: Performance of the 2D-DCT plus temporal features against traditional ROI-based features.

Figure 4.10 shows the classification performance of CkNN and kNN with MIL-based feature selection in comparison to SIL-based feature selection using the 2D-DCT and temporal features. Figure 4.10 shows that the MIL-based learner (CkNN) performs best with the MIL-based feature selection. However, here SIL-based classification (with kNN) also gives better performance with MIL-based feature selection as compared to SIL-based feature selection.

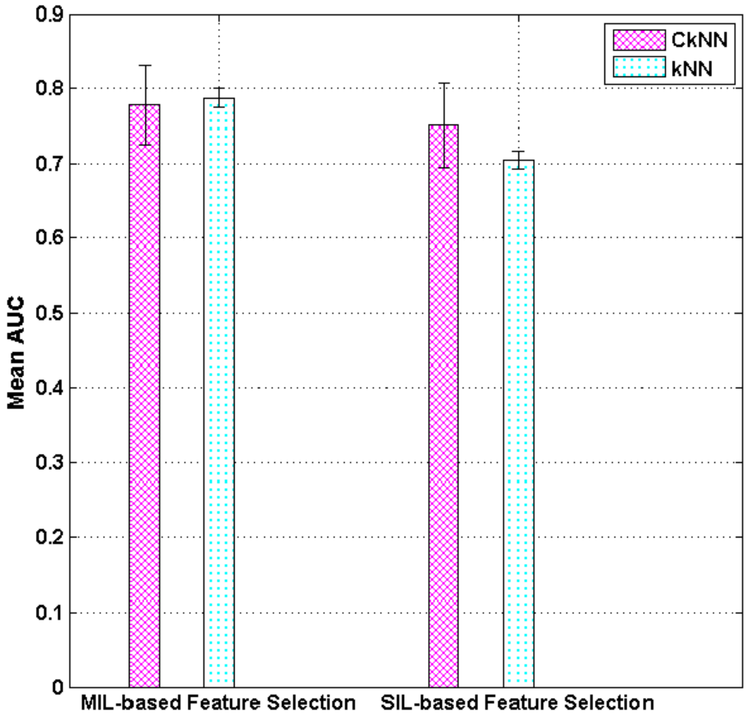


Figure 4.10: Performance of the learners with MIL-based feature selection in comparison to SIL-based feature selection using the 2D-DCT plus temporal features.

4.4. Discussion

Figure 4.6 presents the performance of the tile-based single instance RF for independent tile sizes using the 3D-DCT spatio-temporal features. This figure shows that RF yields the best performance with a tile size of 4×4. But, using a tile size of 4×4 gives computaionally expensive estimates of both model fitting and performance evaluation of CkNN. Therefore, for computational efficiency, we have used an 8×8 tile size. Here, we compare the results of Figure 4.6 (with the 8×8 tile-size alone) against the results presented in Table 4.3.

The first comparison is between the tile-based RF and the tile-based kNN. A *t-test* confirms that, RF gives significantly better classification of benign and malignant tiles as compared to kNN. Here, both algorithms are evaluated with the same features. Thus, we can say that this difference of the results is due to the difference of algorithms. We know that RF is a robust classifier as compared to kNN [96], which is not affected by bias and variance dilema. Here, we do not compare the SIL-based RF with the MIL-based CkNN. Because, we already have analysed the reason for the poor performance of the tile-based kNN as compared to the tile-based RF. Based on the better performance of the tile-based RF in comparison to the tile-based kNN, we believe that if we had used a MIL-based RF, we could get better MIL-based classification.

The second comparison is between the tile-based RF and the ROI-based RF. The tile-based RF performs significantly better as compared to the ROI-based RF. In this case, the classifier is same. The only reason for the differenec of the results is the difference of features. Based on the significantly better performance of the tile-based spatio-temporal features as compared to the ROI-based features, we believe that the tile-based spatio-temporal features have strong potential to classify benign and maliganat lesions.

In Table 4.3 and Table 4.5, we have presented the classification of benign and malignant lesions using the generic tile-based spatio-temporal features and the ROI-based features. A *t-test* indicates that the performance of CkNN is equivalent to the performance of kNN. Also, the generic tile-based classification using CkNN is equivalent to the traditional ROI-based classification. Thus, the tile-based classification using MIL is a viable option for the classification of benign and malignant breast lesions with the additional advantages mentioned in Chapter 3. This indicates that the generic tile-based features allow us to use MIL as a 'pure' machine learning method to solve the breast MRI classification problem with the equivalent performance to the traditional ROI-based classification.

Based on these results, we can say that MIL is a suitable choice for CAD systems. However, a 'pure' machine learning approach has a disadvantage that it produces an unintelligible 'black-box' model to the clinicians. That is, the learning process and generic features are not easily interpretable by the clinicians. However, while traditional ROI-based features may be interpretable by clinicians (as they relate to the physiological properties of the lesion), the learning process (i.e., RF) still results in a black-box model.

Figure 4.11 shows a comparison of the spatio-temporal features from Table 4.3 and Table 4.5 against the spatial features from Chapter 3. The perfromance of spatio-temporal features in comparison to spatial features is also summarised using mean ROC curves in Figure 4.12. Figure 4.11 and Figure 4.12 confirm that classification performance is improved with the tile-based spatio-temporal features compared to the spatial features alone. Moreover, mean ROC curves in Figure 4.12 indicate that we obtain a specificity of 0.37, 0.42, and 0.52 at the clinically useful sensitivity of 0.9, with CkNN using spatial (chapter 3), 2D-DCT plus temporal and 3D-DCT spatio-temporal features respectively. Improved specificity further confirms the better classification performance of spatio-temporal features compared to the spatial features alone. These ROC curves also suggest that CkNN can be used to differentiate benign and malignant lesions with a specificity of 0.45, 0.53, and 0.6 at clinically applicable sensitivty of 0.82, using spatial, 2D-DCT plus temporal and 3D-DCT spatio-temporal features respectively.

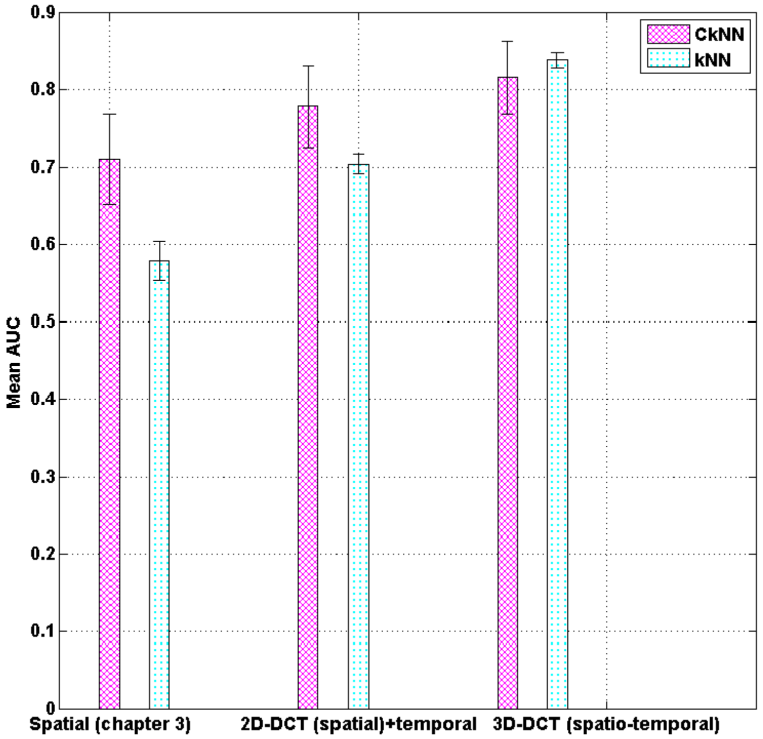


Figure 4.11: Performance of the spatio-temporal features in comparison to the spatial features (Chapter 3).

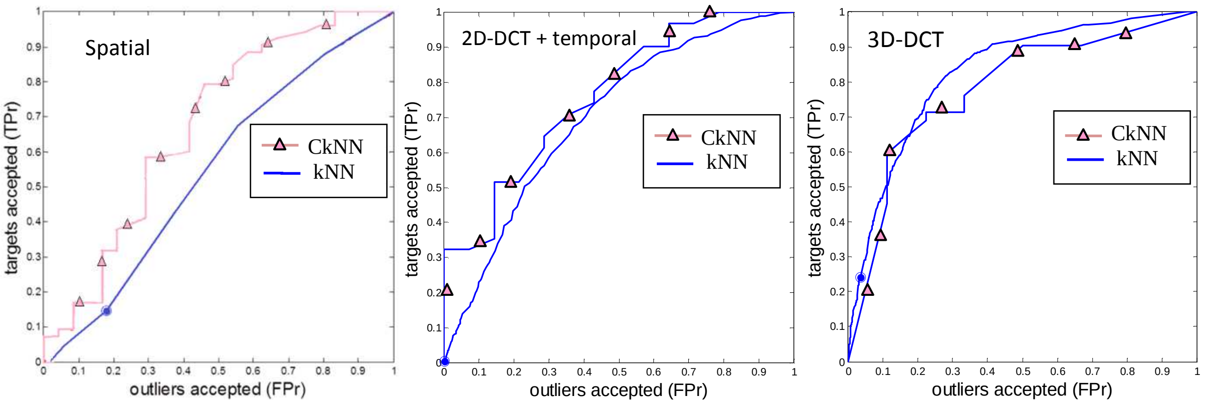


Figure 4.12: Mean ROC curves using spatio-temporal features in comparison to spatial features.

Next, we analyse the classification performance of the learners from Table 4.3 in comparison to Table 4.5. A t-test indicates that the performance of CkNN with the 3D-DCT spatio-temporal features is equivalent to that with the 2D-DCT plus temporal features. However, kNN performs significantly better with the 3D-DCT features as compared to the 2D-DCT plus temporal features. We know that, Table 4.3 presents an unbised estimate of the performance using the 3D-DCT features, while the performance of CkNN and kNN is biased in Table 4.5 due to parameter optimization on the same dataset. Therefore, the 3D-DCT spatio-temporal features would appear better than the 2D-DCT plus temporal features. Moreover with the 2D-DCT plus temporal features, we extract temporal information using parametric enhancement models. Although these models are generic, they are domain specific to some extent. Thus we can say thatthe 3D-DCT spatio-temporal features are more ‘pure’ (generic) than the 2D-DCT plus temporal features.

Average ROC curves using the 3D-DCT spatio-temporal features from DCE-MRI are shown in Figure D.1 (Appendix D). We can see that at clinically useful sensitivity of 0.9, the specificity using the tile-based 3D-DCT spatio-temporal features is 0.6, 0.52 with kNN and CkNN respectively, while we get a specificity of 0.55 using ROI-based state-of-the-art features with RF. There is not much difference between the performance of tile-based and ROI-based features. Rather, we get even better specificity using the tile-based spatio-temporal features with kNN at a sensitivity of 0.9. This encourages the use of generic tile-based spatio-temporal features for the classification of benign and malignant lesions in breast MRI.

From Figure 4.8 and Figure 4.10, we can see that CkNN performs better with MIL-based feature selection as compared to SIL-based feature selection. This demonstrates the importance of a MIL-based criterion for dimensionality reduction in MIL-based classification. Similarly, a SIL-based criterion is important for SIL-based classification.

After doing feature selection, we assess the relative importance of low and high frequency coefficients selected from the DCT. We first arrange DCT coefficients into frequency groups based on the similar sum of their indices in the same way as is done in [117]. That is, coefficients which share similar sum of indices are gathered into a single frequency group. In this way we get five groups for 25 3D-DCT coefficients and six frequency groups for 19 2D-DCT coefficients. Next, we divide DCT frequency groups equally into low and high frequency. We also evaluate the relative importance of horizontal, vertical and diagonal coefficients by counting the number of occurrence of each selected coefficient w.r.t its position.

In MIL-based feature selection, all selected features belong to low frequency group. Moreover, we get equivalent counts of horizontal, vertical and diagonal coefficients. A similar trend was identified for the SIL-based feature selection. This demonstrates an approximately equal importance of horizontal, vertical and diagonal low frequency DCT features for the classification of mass-like lesions. This statement is in accordance with the Chapter 3 where we have analysed the relative importance of horizontal and vertical generic tile-based spatial features for the mass-like lesions. For the 2D-DCT plus temporal features, the MIL-based feature selection returns only the DC coefficient from the 2D-DCT and four temporal model features while, with SIL-based feature selection, we get all low frequency DCT features and no model features. The high occurrence of model features in MIL-based feature selection confirms the importance of temporal information from DCE-MRI for the classification of benign and malignant lesions.

4.5. Conclusion

In this chapter we have evaluated the efficacy of MIL as a ‘pure’ machine learning technique for the discrimination of benign and malignant lesions in breast MRI. Experimental results indicate that performance of CkNN using the tile-based spatio-temporal features is statistically equivalent to the ROI-based classification. However, the tile-based approach does not require any domain specific features and is robust to inaccuracies in the segmentation of suspicious lesions. Therefore, CkNN may be a suitable choice for the classification of benign and malignant lesions. Also, spatio-temporal features have improved discrimination compared to spatial features alone. Moreover, 3D-DCT spatio-temporal features are better than 2D-DCT plus temporal features. Further, we highlight that MIL-based feature selection is important for MIL-based classification.

Chapter 5

Summary and conclusions

In the previous chapters (Chapters 3 and 4), we have investigated the potential of multiple instance learning in the classification of benign and malignant lesions in breast cancer MRI. In this chapter, we provide a short summary of the chapters in section 5.1. Next, we highlight the key contributions and outline the limitations of this research in sections 5.2 and 5.3 respectively. Based on the key contributions of this research, we propose the future work in section 5.4.

5.1 Summary of Chapters

In the 1st **Chapter**, we have outlined the motivation for investigating the efficacy of multiple instance learning in breast cancer MRI CAD systems. Particularly, we have highlighted the issues of inter-observer variations or uncertainties in (semi)-automated lesion segmentation linked to the conventional CAD systems. Later, we have introduced MIL as a key to address these issues. The aims of the research are stated below.

1. To study the performance of multiple instance learning using synthetic datasets.
2. To investigate the performance of multiple instance learning in the identification and classification of breast MRI using generic spatial features from T2-w MRI and to compare the classification performance with the conventional approach (ROI-based single instance learning).
3. To evaluate the efficacy of multiple instance learning for the classification of breast MRI using generic spatio-temporal features (from DCE-MRI and T2-w MRI) and to develop multiple instance learning-based feature selection algorithm.

In **Chapter 2**, we have provided the necessary background on the MIL-based pattern recognition. Specifically, we have talked about MIL in general, followed by MIL and SIL-based learners. We have presented the justification for the experimental methodology, we use in the subsequent chapters. In addition, we have provided the motivation for the aims of this research. Finally, we have presented the breast MRI data, which are used in the experiments in the 3rd and 4th Chapters.

In the 3rd **Chapter,** we have first provided the experiment to understand the working of MIL using synthetic datasets. This experiment addresses the first aim of the research. The results of this experiment give us insight to utilise MIL as a classification tool for the diagnosis of breast cancer. Later in **Chapter 3,** and **Chapter4**, we have addressed the main aim of this thesis, “to evaluate the efficacy of MIL in the diagnosis of breast cancer in MRI”.

We have estimated the performance of MIL for the detection and classification of benign and malignant lesions using spatial features in **Chapter 3**. Particularly, we have used T2-w MRI. We have utilised both (generic) tile-based features and (domain specific) ROI-based features. We have performed experiments on two datasets consisting of mass-like lesions and both mass-like and non-mass-like lesions. Results show that there is not significant difference between the classification potential of the tile-based and the ROI-based features.

In **Chapter 4,** we have presented the performance of MIL for the diagnosis of breast MRI using generic tile-based spatio-temporal features. Specifically, we have used both T1-w DCE-MRI and T2-w MRI. We have utilised a discrete cosine transform and enhancement models as feature extraction techniques. We have compared the performance of MIL against a traditional approach based on bespoke features extracted from a segmented ROI. We also have developed a MIL-based feature selection technique and presented the performance of the learners with MIL-based feature selection in comparison to SIL-based feature selection. We have demonstrated the importance of spatio-temporal features for the classification of benign and malignant lesions. Results demonstrate the equivalent performance of the tile-based generic spatio-temporal features and the ROI-based domain specific features.

5.2 Key Contributions and Findings

In this work, we have used MIL as a generic machine learning approach to diagnose breast cancer in MRI. We have estimated the performance of MIL as a both CAD detection and diagnosis tool. Specifically, we have introduced tile-based MIL as a ‘pure’ machine learning approach to address the problem of variations, associated with the segmentation of ROIs, in conventional CAD system.

To address the first aim of this thesis, we have analysed the performance of CkNN over a range of parameter values (*k* and *c*) using both Hausdorff distance and minimum Euclidean distance. Specifically, we have used *I-Λ* and *I-4I* synthetic datasets in ten bag compositions. Empirical results show that the performance of CkNN has different trends with two distance functions on each dataset. Overall, CkNN performs better with Hausdorff distance as compared to Minimum distance on both datasets. For *I-Λ* dataset, the performance of CkNN increases with increasing the percent positives when utilised with Hausdorff distance. On the other hand, there is not clear trend with Minimum distance. Overall, for high percent positives, CkNN gives better performance with Hausdorff distance particularly with small *k* or *c* combination values. For *I-4I* dataset, CkNN shows unstable performance for different bag compositions with both distance functions.

We have utilised segmentation-free classification approach to address the second and the third aim of this work. We have provided an empirical evidence to suggest that segmentation-free identification and classification of breast MRI dataset using generic tile-based MIL is a viable option.

We have proposed generic tile-based spatial features using T2-w MRI for the characterisation of breast lesions. We also have experimentally proved the importance of generic tile-based features for the classification of breast cancer MRI (medical imaging). To classify both mass-like and non-mass-like lesions as benign or malignant, we have achieved an AUC of 0.7 ± 0.046 using generic tile-based features, while we have acquired an AUC of 0.651 ± 0.048 using ROI-based bespoke features. Moreover, we have provided an empirical evidence to suggest that the tile-size is an important parameter in the tile-based learning.

As part of the third aim of this research, we have explored the importance of 3D-DCT based spatio-temporal features in the diagnosis of breast cancer. We have provided an empirical evidence to suggest that spatio-temporal features have improved discrimination as compared to spatial features alone. We also have empirically shown the better discrimination of 3D-DCT spatio-temporal features as compared to 2D-DCT plus temporal features. Moreover, we have developed a new MIL-based feature selection technique in a MIL-based distance function. In addition, we have provided an experimental evidence to demonstrate the importance of MIL-based feature selection criterion for MIL-based classification. Further, we have proposed a model where we treat features as parameters of model after doing feature selection in every fold of cross validation.

5.3 Limitations

We must acknowledge the limitations of this work. In this thesis, we have analyzed the efficacy of tile-based MIL for only breast MRI data using CkNN alone. We have mentioned in Chapter 4 that nearest neighbour based learners give good classification performance if the size of dataset is very large. The small size of data is a major limitation of this work. Due to the limited amount of data, we could not split the data independently for parameter tuning, feature selection, and performance estimation. Please note that CkNN has shown good performance (identical to the tile-based or ROI-based SIL) with parameter tuning on the synthetic datasets in Chapter 3 and on other dataset in Chapter 4 (we have optimised the parameters used in 3D-DCT by randomly selecting coefficients from 2D-DCT on T2-w MRI). However, CkNN could give even better performance if parameters were tuned on the subset of the same dataset.

5.4 Future Work

Tile-based spatio-temporal features have been found promising in the diagnosis of breast cancer in MRI (Chapter 4). Future research should consider investigation of multiple instance learning as a screening tool using spatio-temporal features. Moreover, the generic tile-based spatio-temporal features give an AUC of 0.95 with a 4×4 tile size and 0.92 with an 8×8 tile size with the tile-based SIL using RF for the classification of benign and malignant tiles (subsection D of section 4.2). This performance is significantly better than the tile-based SIL using kNN. Based on the significantly better performance of the tile-based RF as compared to the tile-based kNN, future research should focus on investigating a tile-based MIL using RF for the classification of medical imaging and compare its performance with the conventional CAD (ROI-based SIL using RF).

Moreover, we know that a tile-based MIL gives labels to the bag, while a tile-based SIL provides labels to individual tiles. Label of tiles can be used to get the rough segmentation of the detected lesions. Future research should consider tile-based spatio-temporal features using SIL-based RF to get the rough segmentation of ROIs; which can be used in ROI-based MIL.

Jianrui Ding et al. [36] have demonstrated the improved classification of benign and malignant lesions in breast ultrasound images using ROI-based MIL. Their proposed methodology extracts local textural features of the sub-regions in roughly segmented ROIs to characterize the lesions. Investigation of ROI-based MIL for the diagnosis of breast cancer in MRI and its comparison with the conventional approach (ROI-based SIL) could be a potential avenue for further research especially when data is semi-labelled, i.e., when we get labels only for histopathology verified suspiciously enhancing ROIs out of many ROIs detected in the screening process.

In summary, a tile-based approach does not require any domain specific features and is robust to inaccuracies in the segmentation of suspicious lesions. Thus, MIL using generic tile-based features permits the fully automatic detection and segmentation-free classification of suspecious lesions.

Bibliography

[1] Australian Institute of Health and Welfare & Cancer, Australia, “Breast cancer in Australia: an overview,” Cancer series, no. 71, Cat. No. CAN 67, Canberra: AIHW, 2012.

[2] Tabar L., Dean P.B., “Mammography and breast cancer: the new era,” *Int J Gynaecol Obstet* (PubMed: 14499978), 82, no. 3, pp. 319–326, 2003.

[3] Rankin, S.C., “MRI of the breast,” *Br J Radiol*, 73, pp. 806–818, 2000.

[4] P.L. Davis, M.J. Staiger, K.B. Harris , M.A. Gannot , J. Klementaviciene , K.S. MacCarty , Jr., H. Tobon. “Breast cancer measurements with magnetic resonance imaging, ultra sonography, and mammography,” *Breast Cancer Res. Treatment*, vol 37, pp. 1–9, 1996.

[5] Santyr G.E., “MR imaging of the breast: imaging and tissue characterization without intravenous contrast,” *Breast Imaging*, MR Clinics of North America (Philadelphia: Saunders), no. 2:4, pp. 673–690, 1994.

[6] Kerslake R.W., Carleton P.J., Fox J.N., Imrie M.J., Cook A.M., Read J.R., Bowsley S.J., Buckley D.L., Horsman A., “Dynamic gradient-echo and fat-suppressed spin-echo contrast-enhanced MRI of the breast,” *Clin Radiol*, 50, no. 7, pp. 440–454, 1995.

[7] Heywang S.H., Bassermann R., Fenzl G., Nathrath W., Hahn D., Beck R., Krischke I., Eiermann W., “MRI of the breast-histopathologic correlation,” *European Journal of Radiology*, 7, no. 3, pp. 175–182, 1987.

[8] M.Van Geothem, W. Tjalma, K. Schelfout, I. Verslegers, I. Biltjes, and P. Parizel, “Magnetic resonance imaging in breast cancer,” *European Journal of Surgical Oncology*, 32 (9), pp. 901–910, 2006.

[9] J.S. van den Brink, Y. Watanabe, C.K. Kuhl, T. Chung, R. Muthupillai, M.V. Cauteren, K. Yamada, S. Dymarkowski, J. Bogaert, J.H. Maki, C. Matos, J.W. Casselman, and R.M. Hoogeveen, “Implications of SENSE MR in routine clinical practice,” *Eur J Radiol*, 46 (1), pp. 3–27, Apr 2003.

[10] Y. Gal, A. Mehnert, A. Bradley, D. Kennedy, S. Crozier, “New spatiotemporal features for improved discrimination of benign and malignant lesion in dynamic contrast-enhanced-magnetic resonance imaging of the breast,” *Journal of Computer Assisted Tomography*, 35 (5), pp. 645-652, 2011.

[11] N. Bhooshan, M. Giger, L. Lan, H. Li, A. Marquez, A. Shimauchi, and G.M. Newstead, “Combined use of T2-weighted MRI and T1-weighted dynamic contrast-enhanced MRI in the automated analysis of breast lesions,” *Magnetic Resonance in Medicine* 66, no. 2, pp. 555–564, Aug 2011.

[12] Mussurakis S., Buckley D.L., Coady A.M., Turnbull L.W. & Horsman A., “Observer variability in the interpretation of contrast enhanced MRI of the breast”, *The British journal of radiology,*69 (827), pp. 1009–16, 1996.

[13] Kinkel K., Helbich T.H., Esserman L.J., Barclay J., Schwerin E.H., Sickles E.A. & Hylton, N.M.,“MR imaging of suspicious breast lesions: diagnostic criteria and interobserver variability”, *American Journal of Roentgenology* (July), pp. 35–44, 2000.

[14] Ikeda, D.M., Hylton, N.M., Kinkel, K., Hochman, M.G., Kuhl, C.K., Kaiser, W., Weinreb, J.C., Smazal, S.F., Degani, H., Viehweg, P., Barclay, J. & Schnall, M.D., “Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies,” *Journal of magnetic resonance imaging* : *JMRI* 13 (6), pp. 889–95, 2001.

[15] Weijie Chen, Maryellen L. Giger, Li Lan, and Ulrich Bick, “Computerized interpretation of breast MRI: investigation of enhancement variance dynamics,” *Med. Phys,* 31, no. 5, pp. 1076–1082, 2004.

[16] Maryellen L. Giger, Nico Karssemeijer, and Julia A. Schnabel, “Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer,” *Annu. Rev. Biomed*. Eng. 2013. 15: pp. 327–57, May 13, 2013.

[17] Pan, J., Dogan, B. E., Carkaci, S., Santiago, L., Arribas, E., Cantor, S. B., Wei, W., Stafford, R. J. & Whitman, G. J., “Comparing performance of the CADstream and the DynaCAD breast MRI CAD systems : CADstream vs. DynaCAD in breast MRI,” *Journal of digital imaging*, 2013.

[18] Thomas G. Dietterich, Richard H. Lathrop and Tomas Lozano-Perez, “ Solving multiple instance problem with axis-parallel rectangles,” *Artificial Intelligence,* 89, no. 1-2, pp. 31–71, 1997.

[19] James Foulds, Eibe Frank, “A review of multi- instance learning assumptions,” *Knowledge Engineering Review* (Cambridge University Press) 25, no. 01, pp. 1–25, 2010.

[20] A. Zafra, S. Ventura, C. Romero, and E. Herrera-Viedma, "Multi-instance genetic programming for web index recommendation," *Expert System with Applications*, vol. 36, pp.11 470–11 479, 2009.

[21] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in NIPS'02: Proceedings of Neural Information Processing System, Vancouver, Canada, pp. 561–568, 2003.

[22] Zhou, Z., Sun, Y., Li, Y., "Multi-instance learning by treating instances as non-IID samples," In: Proc. 26th ICML, pp. 1249–1256, 2009.

[23] J. Yang, *Review of multiple instance learning and its applications*, Technical report, School of computer science, Carnegie Mellon University, 2005.

[24] Y. Chevaleyre, J.D. Zukar, "Solving multiple instance and multiple part learning problems with decision tree and rule sets," Application to the mutagenesis problem, In proceedings of the 14th biennial conference of the Canadian Society for Computational Studies of Intelligence, Springer, pp. 204-214, 2001.

[25] Veronika Cheplygina, David M. J. Tax, Marco Loog, "On classification with bags, groups and sets," arXiv:1406.0281v2 [stat.ML] , 7 Oct 2014.

[26] Fu G, Nan X, Liu H, Patel R, Daga P, Chen Y, Wilkins D, Doerksen R, **"**Implementation of multiple-instance learning in drug activity prediction**,"** *BMC Bioinformatics*, 13 (Suppl 15)**:**S3, 2012.

[27] R. Rahmani, Sally A. Goldman, Hui Zhang, Sharath R. Cholleti, and Jason E. Fritts, "Localized content based image retrieval," *IEEE Trans Pattern Anal Mach Intell,* 30 (11), pp. 1902–1912, 2008.

[28] M.C. Burl, M. Weber, P. Perona, "A probabilistic approach to object recognition using local photometry and global geometry," In proceedings of the 5th European conference on Computer Vision, Springer, pp. 628–641, 1998.

[29] O. Maron, A.L. Raton, "Multiple instance learning for natural scene image classification," In proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, pp.341–349, 1998.

[30] N. Weidmann, Two-level classification for generalised multi-instance data, Master's Thesis, Albert Ludwig University of Freiburg, 2003.

[31] Noor A. Samsudin, Andrew P. Bradley, “Group-based meta classification,” In International Conference on Pattern Recognition, pp. 2256–2259, 2008.

[32] Noor A. Samsudin, Andrew P. Bradley, “Nearest neighbour group-based classification,” *Pattern Recognition***,** Volume 43, Issue 10, pp. 3458–3467, October 2010.

[33] Jinbo Bi, Jianming Liang, “Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure,” In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[34] Xu, Y., Zhang, J., Chang, E., Lai, M., Tu, Z., “Context-constrained multiple instance learning for histopathology image segmentation,” MICCAI, pp. 623–630, 2012.

[35] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V. Hajnal, Daniel Rueckert, “Multiple instance learning for classification of dementia in brain MRI,” *Medical Image Analysis* 18, 808–818, 2014.

[36] Jianrui Ding, H.D. Cheng, Jianhua Huang, Jiafeng Liu, and Yingtao Zhang, “Breast ultrasound image classification based on multiple-instance learning,” *Journal of Digital Imaging*, 25, pp. 620–627, 2012.

[37] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, John Wiley and sons, Inc., New York, 2001.

[38] B.V. Dasarathy, *Nearest neighbours norms: NN pattern classification techniques*, IEEE computer society press, 1991.

[39] J. Wang, J.D. Zucker, “Solving the multiple-instance problem: a lazy learning approach,” In Proceedings of the 17th International Conference on Machine Learning. San Francisco, CA, pp. 1119–1125, 2000.

[40] Edgar, Gerald, *Measure, Topology, and Fractal Geometry*, (3rd print), Springer-Verlag, 1995.

[41] D.J. Peuquet, “An algorithm for calculating minimum Euclidean distance between two geographic features,” *Comput Geosci*, 18 (8), pp. 989-1001, 1992.

[42] FIX, E. and HODGES, J.L., JR, Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4, 1951.

[43] Dhurandhar A., Dobra A., “Probabilistic characterization of nearest neighbour classifier,” *Int J Mach Learn and Cyber* 4, no. 4, pp. 259-272, 2013.

[44] C. Elkan, “Nearest neighbour (kNN) classification method,” 2011, Jan., Available: http://cseweb.ucsd.edu/users/elkan/250Bwinter2011/.

[45] Ranga Raju Vatsavai, “Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery,” Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, Illinois, USA, p.p. 1419-1426 August 11-14, 2013, [doi>10.1145/2487575.2488210].

[46] Dánel Sánchez Tarragó, Chris Cornelis, Rafael Bello, Francisco Herrera, “A multi-instance learning wrapper based on the Rocchio classifier for web index recommendation,” *Knowledge-Based Systems*, Volume 59, p.p.173–181, March 2014.

[47] Xu,X, Statistical learning in multiple instance problems, Master’s thesis, University of Waikato, 2003.

[48] Lin Dong, A comparison of multi-instance learning algorithms, Master of Science Dissertation, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 2006.

[49] Jain, A. & Duin, P, “Statistical pattern recognition: a review,” *IEEE Transactions on pattern analysis and machine intellegence* 22, no. 1, pp. 4–37, 2000.

[50] Amelia Zafra, Mykola Pechenizkiy, Sebastian Ventura, “Feature selection is the relief for multiple instance learning,” in 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 525-532, 2010.

[51] Alppaydin, Ethem, *Introduction to machine learning*, Massachusetts London, England: The MIT Press Cambridge, 2010.

[52] Ripley, Brian D., *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press, 1996.

[53] Weiss, Sholom M., Kulikowski, Casimar A., *Computer systems that learn*, San Meteo, CA: Morgan Kauffmann, 1991.

[54] A.P. Bradley, “The use of area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition* 30, no. 7, pp. 45-1159, 1997.

[55] Jin Huang, Charles X. Ling, “Using AUC and accuracy in evaluating learning algorithms”, *IEEE Transactions on Knowledge and data Engineering*, VOL. 17, No. 3, pp. 299–310, March 2005.

[56] David M. J. Tax, Robert P. W. Duin, “Learning curves for the analysis of multiple instance classifiers,” *Structural, Syntactic, and Statistical Pattern Recognition*, Volume 5342, pp. 724-733, 2008.

[57] Greiner M., Pfeiffer D., Smith R.D., “Principles and practical application of the receiver operating characteristics analysis for diagnostic test,” *Prev Vet Med* 45, pp. 23–41, 2000.

[58] Swets J.A., “ROC analysis applied to the evaluation of medical imaging techniques”, *Invest Radiol 14*, no. 2, pp. 109–121, 1979.

[59] Hanley J.A., McNeil B.J., “The meaning and use of the area under a receiver operating characteristics (ROC) curve,” *Radiology* 143, pp. 29–36, 1982.

[60] C.S. Klifa, A. Shimakawa, Z. Siraj, J.E. Gibbs, L.J. Wilmes, S.C. Partridge, E. Proctor, N.M. Hylton, “Characterization of breast lesions using the 3D FIESTA sequence and contrast-enhanced magnetic resonance imaging,” *J Magn Reson Imaging* 25, pp. 82–88, 2007.

[61] M.G. Hochman, S.G. Orel, C.M. Powell, M.D. Schnall, C.A. Reynolds, and L.N. White, “Fibroadenomas: MR imaging appearances with radiologic-histopathologic correlation,” *Radiology*, 204 (1), pp. 123–129, Jul 1997.

[62] M. Kuwabara, “MRI of breast tumors with emphasis on histopathologic correlation [in Japanese],” *Nippon Igaku Hoshasen Gakkai Zasshi*, 51 (11), pp. 1366–1374, Nov 1991.

[63] L. Liberman, E.A. Morris, M.J.Y. Lee, J.B. Kaplan, L.R. La Trenta, J.H. Menell, A.F. Abramson, S.M. Dashnaw, D.J. Ballon, and D.D. Dershaw, “Breast lesions detected on MR imaging: features and positive predictive value,” *AJR Am J Roentgenol*, 179 (1), pp. 171–178, Jul 2002.

[64] E.A. Morris, L. Liberman, D.J. Ballon, M. Robson, A.F. Abramson, A. Heerdt, and D.D. Dershaw., “MRI of occult breast carcinoma in a high-risk population,” *AJR Am J Roentgenol*, 181 (3), pp. 619–626, Sep 2003.

[65] L.W. Nunes, M.D. Schnall, and S.G. Orel, “Update of breast MR imaging architectural interpretation model,” *Radiology*, 219 (2), pp. 484–494, May 2001.

[66] S. Wurdinger, A.B. Herzog, D.R. Fischer, C. Marx, G. Raabe, A. Schneider, and W.A. Kaiser, “Differentiation of phyllodes breast tumors from fibroadenomas on MRI,” *AJR Am J Roentgenol*, 185 (5), pp. 1317–1321, Nov 2005.

[67] M. Kawashima, Y. Tamaki, T. Nonaka, K. Higuchi, M. Kimura, T. Koida, Y. Yanagita, and S. Sugihara, “MR imaging of mucinous carcinoma of the breast,” *AJR Am J Roentgenol*, 179 (1), pp. 179–183, Jul 2002.

[68] T. Okafuji, H. Yabuuchi, S. Sakai, H. Soeda, Y. Matsuo, T. Inoue, M. Hatakenaka, N. Takahashi, S. Kuroki, E. Tokunaga, and H. Honda, “MR imaging features of pure mucinous carcinoma of the breast,” *Eur J Radiol*, 60 (3), pp. 405–413, Dec 2006.

[69] D.M. Renz, P.A.T. Baltzer, M. Facius, S.O.R. Pfleiderer, M. Gajda, O. Camara, and W.A. Kaiser, “New signs of tumors in magnetic resonance mammography,” *Eur Radiol*, 16 (Suppl 5), pp. E80–E82, 2006.

[70] M.D. Schnall, J. Blume, D.A. Bluemke, G.A. DeAngelis, N. DeBruhl, S. Harms, S.H. Heywang-Kbrunner, N. Hylton, C. K. Kuhl, E.D. Pisano, P. Causer, S.J. Schnitt, D. Thickman, C.B. Stelling, P.T. Weatherall, C. Lehman, and C.A. Gatsonis, “Diagnostic architectural and dynamic features at breast MR imaging: multicenter study,” *Radiology*, 238 (1), pp. 42–53, Jan 2006.

[71] H. Yabuuchi, H. Soeda, Y. Matsuo, T. Okafuji, T. Eguchi, S. Sakai, S. Kuroki, E. Tokunaga, S. Ohno, K. Nishiyama, M. Hatakenaka, and H. Honda., “Phyllodestumor of the breast: Correlation between MR findings and histologic grade,” *Radiology*, 241 (3), pp.702–709, Dec 2006.

[72] F. Yang and W.A. Kaiser, “Different types of edema in benign and malignant lesions of the breast in MR mammography,” *Eur Radiol*, 16 (Suppl 5), pp. E99–E99, 2006.

[73] A. Malich, D.R. Fischer, S. Wurdinger, J. Boettcher, C. Marx, M. Facius, and W.A. Kaiser, “Potential MRI interpretation model: differentiation of benign from malignant breast masses,” *AJR Am J Roentgenol*, 185 (4), pp. 964–970, Oct 2005.

[74] D.R. Fischer, S. Wurdinger, J. Boettcher, A. Malich, and W.A. Kaiser, "Further signs in the evaluation of magnetic resonance mammography: a retrospective study," *Invest Radiol*, 40 (7), pp. 430–435, Jul 2005.

[75] M. Dietzel, P. Baltzer, M. Gajd, and W. Kaiser, "Differential diagnosis of up to 20mm sized breast lesions using morphologic features in T2-w of MR mammography," In European conference of radiology 2007, B-625, Vienna, March 2007, European Society of Radiology.

[76] S. Yuen, T. Uematsu, M. Kasami, K. Tanaka, K. Kimura, J. Sanuki, Y. Uchida, and H. Furukawa, "Breast carcinomas with strong high-signal intensity on T2-weighted MR images: pathological characteristics and differential diagnosis," *J Magn Reson Imaging*, 25 (3), pp. 502–510, Mar 2007.

[77] E.A. Sadowski and F. Kelcz, "Frequency of malignancy in lesions classfied as probably benign after dynamic contrast-enhanced breast MRI examination," *J Magn Reson Imaging*, 21 (5), pp. 556 – 564, May 2005.

[78] M.A. Jacobs, P.B. Barker, D.A. Bluemke, C. Maranto, C.Arnold, E.H. Herskovits, and Z. Bhujwalla, "Benign and malignant breast lesions: diagnosis with multi-parametric MR imaging," *Radiology*, 229 (1), pp. 225–232, Oct 2003.

[79] E.E. Deurloo, S.H. Muller, J.L. Peterse, A.P.E. Besnard, and K.G.A. Gilhuijs, "Clinically and mammographically occult breast lesions on MR images: potential effect of computerized assessment on clinical reading," *Radiology*, 234 (3), pp. 693–701, Mar 2005.

[80] American College of Radiology (ACR), "ACR BIRADS – magnetic resonance imaging," 2008.

[81] Fahira A. Maken, Yaniv Gal, Darryl Mcclymont, Andrew P. Bradley, "Multiple instance learning for breast cancer magnetic resonance imaging," Digital Imaging Computing: Techniques and Applications (DICTA). Wollongong, pp. 1–8, 2014.

[82] Turgay Ayer, Mehmet US Ayvaci, Ze Xiu Liu, Oguzhan Alagoz, Elizabeth S. Burnside, "Computer-aided diagnostic models in breast cancer screening," *Imaging in medicine* 2, no. 3, pp. 313–323, 2010.

[83] Fukunaga, K. *Introduction to statistical pattern recognition.* 2nd eddition, Boston: Academic Press, 1990.

[84] Steven D. Brossi, Andrew P. Bradley, "A Comparison of multiple instance and group based learning," Digital Imaging Computing: Techniques and Applications (DICTA), pp. 1-8, Fremantle, Western Australia Perth, 2012.

[85] Rosset A., Spadola L., Ratib O., "OsiriX: an open-source software for navigation in multidimensional Dicom images," *J Digital Imiging* 17, no. 3, pp. 205–216, 2004.

[86] Paul F. Evangelista, Mark J. Embrechts, Boleslaw K. Szymanski, "Taming the curse of dimensionality in kernels and novelty detection," *Applied soft computing technologies: the challenge of complexity* In Abraham A., Bacts B.d., Koppen M., Nickolay B. (Eds.), pp. 431–444, 2006.

[87] L. Breiman, "Random forests," *Machine learning* 45, pp. 5–32, 2001.

[88] W.G. Touw, J.R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. van Hijum. "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?," *Briefings in bioinformatics* 14, pp. 315–326, 2013.

[89] D.M.J. Tax., MIL: A Matlab toolbox for multiple instance learning, 2013.

[90] R Core Team, R: a language and environment for statistical computing, 2013.

[91] A. Liaw, M.Wiener, "Classification and regression by random Forest," *R News* 2, no. 3, pp. 18–22, 2000.

[92] Darryl G. McClymont, Computer assisted detection and characterization of breast cancer in MRI, PhD Thesis, The University of Queensland, Australia, 2015.

[93] W.V. Aalst, T. Twellmann, H. Buurman, F.A. Gerritsen, M. Bart, H. Romeny, "Computer-aided diagnosis in breast MRI : do adjunct features derived from T2-weighted images improve classification of breast masses?," *Bildverarbeitung für die Medizin*, pp. 11-15, 2008.

[94] Cao, M. M., Hoyt, A.C. and Bassett, L.W., "Mammographic signs of systemic disease*," Radiographics: a review publication of the Radiological Society of North America, Inc,* 31(4), 1085–100, 2011.

[95] Kawashima, H., Kobayashi-Yoshida, M., Matsui, O., Zen, Y., Suzuki, M. and Inokuchi, M., "Peripheral hyperintense pattern on T2-weighted magnetic resonance imaging (MRI) in breast carcinoma:

correlation with early peripheral enhancement on dynamic MRI and histopathologic findings," *Journal of magnetic resonance imaging : JMRI*, 32(5), 1117–23, 2010.

[96] G.Izmirlian, "Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial," *Annals of the New York Academy of Sciences* 1020, pp. 154–174, 2004.

[97] W. Kaiser, *Signs in MR-mammography*, Springer Publishing , 2009.

[98] A. Thomas Stavros, D. Thickman, C. Rapp, M. Dennis, S. Parker, G. Sisney, "Solid breast nodules: use of sonography to distinguish between benign and malignant lesions," *Radiological Society North America* 196, pp. 123–134, 1995.

[99] Fahira Afzal Maken, Andrew P. Bradley, "Multiple instance learning for breast MRI based on generic spatio-temporal features," 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, pp. 902-906, 2015.

[100] Warren R., Coulthard A., *Breast MRI in practice.* London, UK: Martin Dunitz, 2002.

[101] Kuhl C.K., P. Mielcareck, S. Klaschik, C. Leutner, E.Wardelmann, J.Gieseke, H.H. Schild, "Dynaimc breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions?," *Radiology* 211, pp. 101–110, 1999.

[102] Kuhl C.K., Schild H.H., "Dynamic image interpretation of MRI of the breast," *J. Magn. Reson. Imaging* 12, pp. 965–974, 2000.

[103] Fusco, R., Lesion detection and classification in breast cancer: evaluation of approaches based on morphological features, tracer kinetic modelling and semi-quantitative parameters in MR functional imaging (DCE-MRI), PhD thesis, University of Bologna, 2013.

[104] Chan, A. A. & Nelson, S. J., Simplified gamma-variate fitting of perfusion curves, in 'IEEE International Symposium on Biomedical Imaging: Nano to Macro' 2004.

[105] Ricker, W. E., *Handbook of computations for biological statistics of fish populations*, 1958.

[106] Agliozzo, S., De Luca, M., Bracco, C., Vignati, A., Giannini, V., Martincich, L., Carbonaro, L., Bert, A., Sardanelli, F. & Regge, D., "Computer-aided diagnosis for dynamic contrast enhanced breast

MRI of mass-like lesions using a multiparametric model combining a selection of morphological, kinetic, and spatiotemporal features," *Medical physics,* 39(4), 1704–15, 2012.

[107] Tofts, P. S., Berkowitz, B. & Schnall, M. D., "Quantitative analysis of dynamic Gd-DTPA enhancement in breast tumors using a permeability model,", *Magnetic Resonance in Medicine* 33(4), 564–8, 1995.

[108] Brix, G., Kiessling, F., Lucht, R., Darai, S., Wasser, K., Delorme, S. & Griebel, J., "Microcirculation and microvasculature in breast tumors: pharmacokinetic analysis of dynamic MR image series," Magnetic *Resonance in Medicine* 52(2), 420–9, 2004.

[109] P.M. Hayton, "Analysis of contrast-enhanced breast MRI," in Department of Engineering Science, vol. PHD Oxford: The University of Oxford, Oxford, UK, 1998.

[110] K. Rao, P. Yip, *Discrete Cosine Transform : Algorithms, Advantages, Applications,* San Diego, CA, USA : Academic Press Professional, Inc., 1990.

[111] Darryl McClymont, Andrew Mehnert, Adnan Trakic, Dominic Kennedy, Stuart Crozier, "Multimodal features for improved breast MRI CAD," *Journal of Medical Imaging*, 2015 (under review).

[112] O.Schabenberger, *Contemporary statistical models for the plant and soil science,* CRC Press, 2002.

[113] Yaniv Gal, Andrew Mehnert, Andrew Bradley, Kerry McMahon, and Stuart Crozier, "An evaluation of four parametric models of contrast enhancement for dynamic magnetic resonance imaging of the breast," 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 2007.

[114] Andrew Mehnert, Kerry McMahon and Dominic Kennedy, Ewert Bengtsson, Stephen Wilson and Stuart Crozier, "Visualisation of the pattern of contrast enhancement in dynamic breast MRI," APRS Workshop on digital Image Computing (WDIC2005), Brisbane, Australia, 2005.

[115] Yaniv Gal, Computer aided analysis of dynamic contrast enhanced MRI of breast cancer, PhD Dissertation, The University of Queensland Australia, 2010.

[116] Boon-Lock Yeo, Bede Liu, "Volume rendering of DCT-based compressed 3D scalar data," *IEEE Trans. Visualization and Computer Graphics* 1, no. 1, pp. 29–43, March 1995.

[117] Malavika Bhaskaranand, Jerry D. Gibson, “Distributions of 3D DCT coefficients for video,” International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, pp. 793–796, 2009.

[118] Kittler J., “Feature set search algorithms,” *Pattern Recognition and Signal Processing* , pp. 41-60, The Netherlands: Sijthoff and Noordhoff, 1978.

[119] P. Hart, and T. Cover, “Nearest neighbor pattern classification,” *IEEE Trans. Information Theory*, Jan. vol. 13, no. 1, pp. 21–27, 1967.

Appendix A: Analysis of synthetic datasets using multiple instance learning

To study the working of multiple instance learning on synthetic datasets in section 3.1, we search the parameter space for 72 combinations of ‘k reference neighbours’ and ‘c citer’s rank’. Table A shows the [k, c] values utilised in the experiment.

Table A: 72 combination values of ‘k, c’ (column wise)

1, 0	3, 0	7, 0	15, 0	31, 0	63, 0	127, 0	255, 0
1, 1	3, 1	7, 1	15, 1	31, 1	63, 1	127, 1	255, 1
1, 4	3, 4	7, 4	15, 4	31, 4	63, 4	127, 4	255, 4
1, 8	3, 8	7, 8	15, 8	31, 8	63, 8	127, 8	255, 8
1, 16	3, 16	7, 16	15, 16	31, 16	63, 16	127, 16	255, 16
1, 32	3, 32	7, 32	15, 32	31, 32	63, 32	127, 32	255, 32
1, 64	3, 64	7, 64	15, 64	31, 64	63, 64	127, 64	255, 64
1, 128	3, 128	7, 128	15, 128	31, 128	63, 128	127, 128	255, 128
1, 256	3, 256	7, 256	15, 256	31, 256	63, 256	127, 256	255, 256

Figure A.1 and Figure A.2 illustrate the performance of CkNN using dataset 1, over 72 values of [k, c] (as mentioned in Table A) with Hausdorff distance and Minimum distance respectively. Colour bar maps the performance of CkNN in mean AUC. We can see that the performance of CkNN has an increasing trend with increasing percent positives (PP) using Hausdorff distance (Figure A.1), while there is no clear trend with Minimum distance (Figure A.2). Overall, CkNN gives good performance with Hausdorff distance for $PP > 60\%$. Figure A.2 suggests the robust performance of CkNN using Minimum distance (irrespective of the bag composition) over a range of [k, c] values.

Figure A.3 and Figure A.4 respectively present the performance of CkNN over 72 values of [k, c] (as mentioned in Table A), with Hausdorff distance and Minimum distance using dataset 2. Colour bar maps the performance of CkNN in mean AUC. It can be seen that, performance is not improved with increasing PP, rather remains unstable using either distance function. Overall, CkNN performs better with Hausdorff distance as compared to Minimum distance.

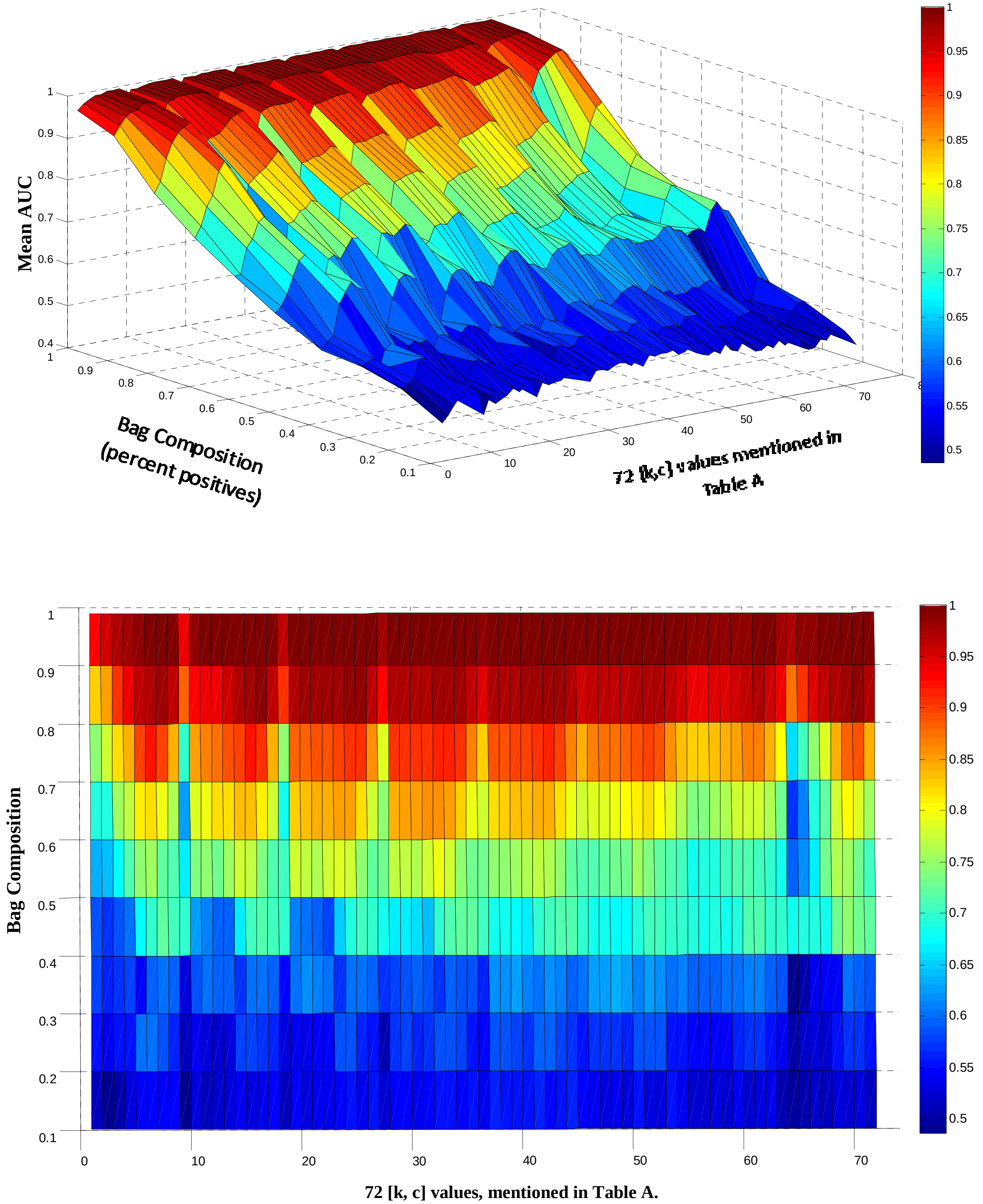


Figure A.1: Performance of CkNN with Hausdorff dist. using dataset 1.

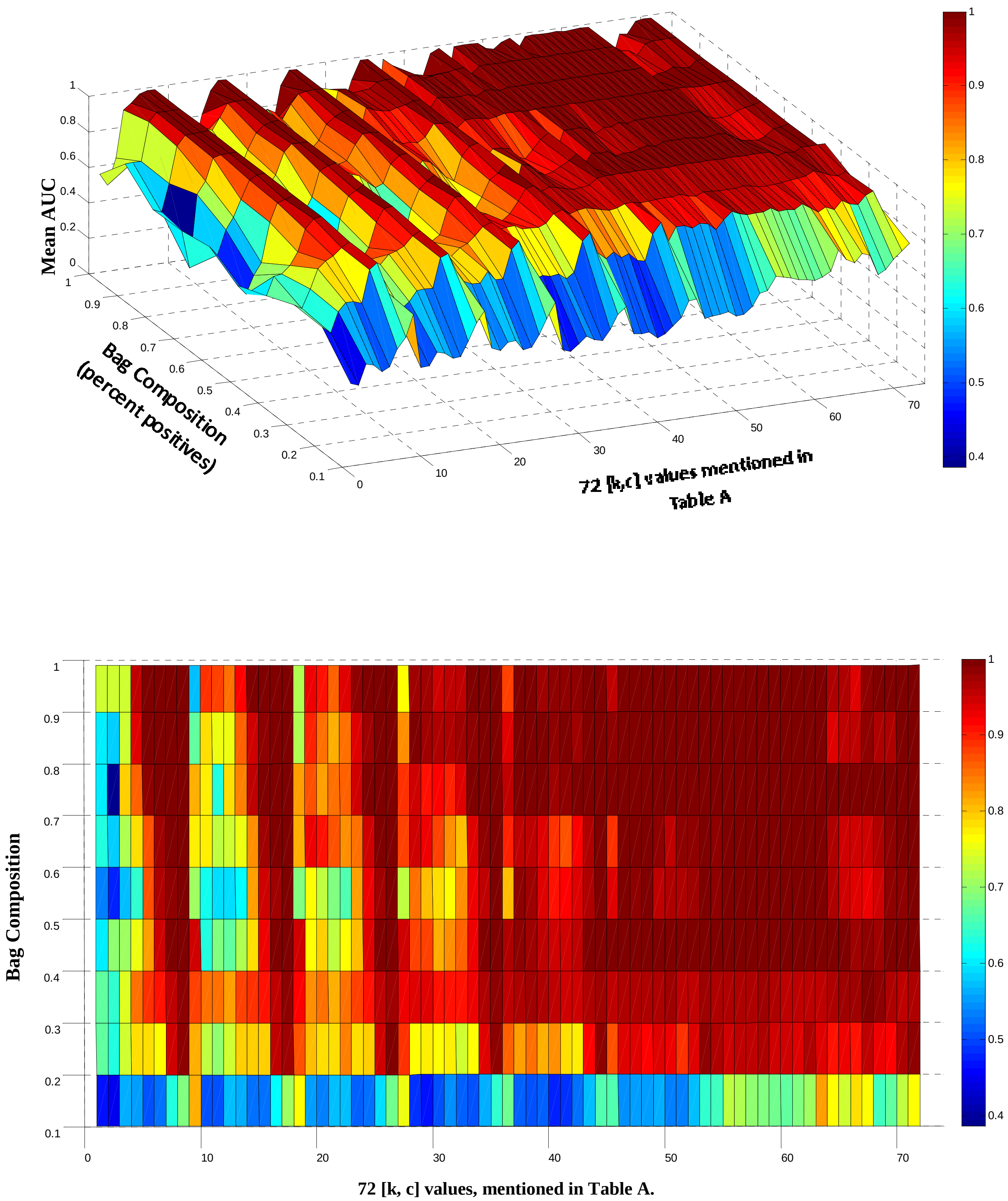
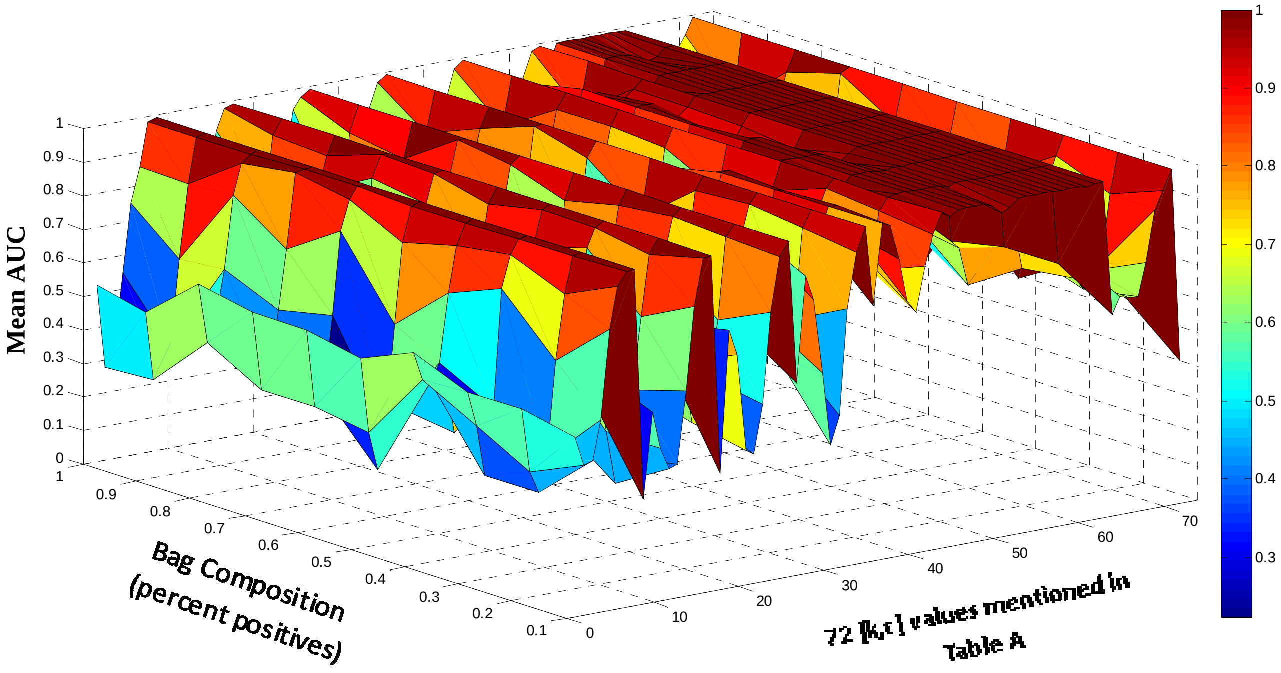


Figure A.2: Performance of CkNN with Minimum dist. using dataset 1.



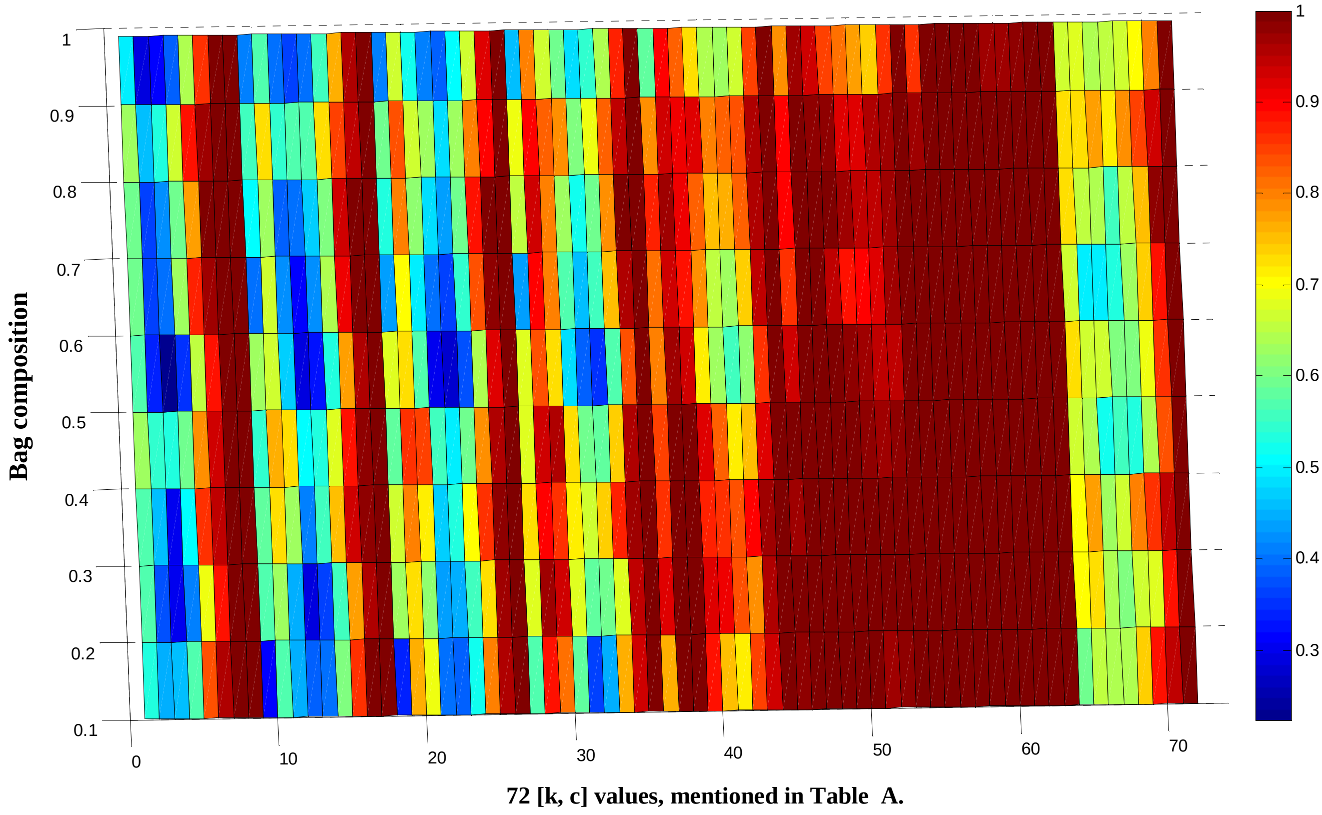


Figure A.3: Performance of CkNN with Hausdorff dist. using dataset 2.

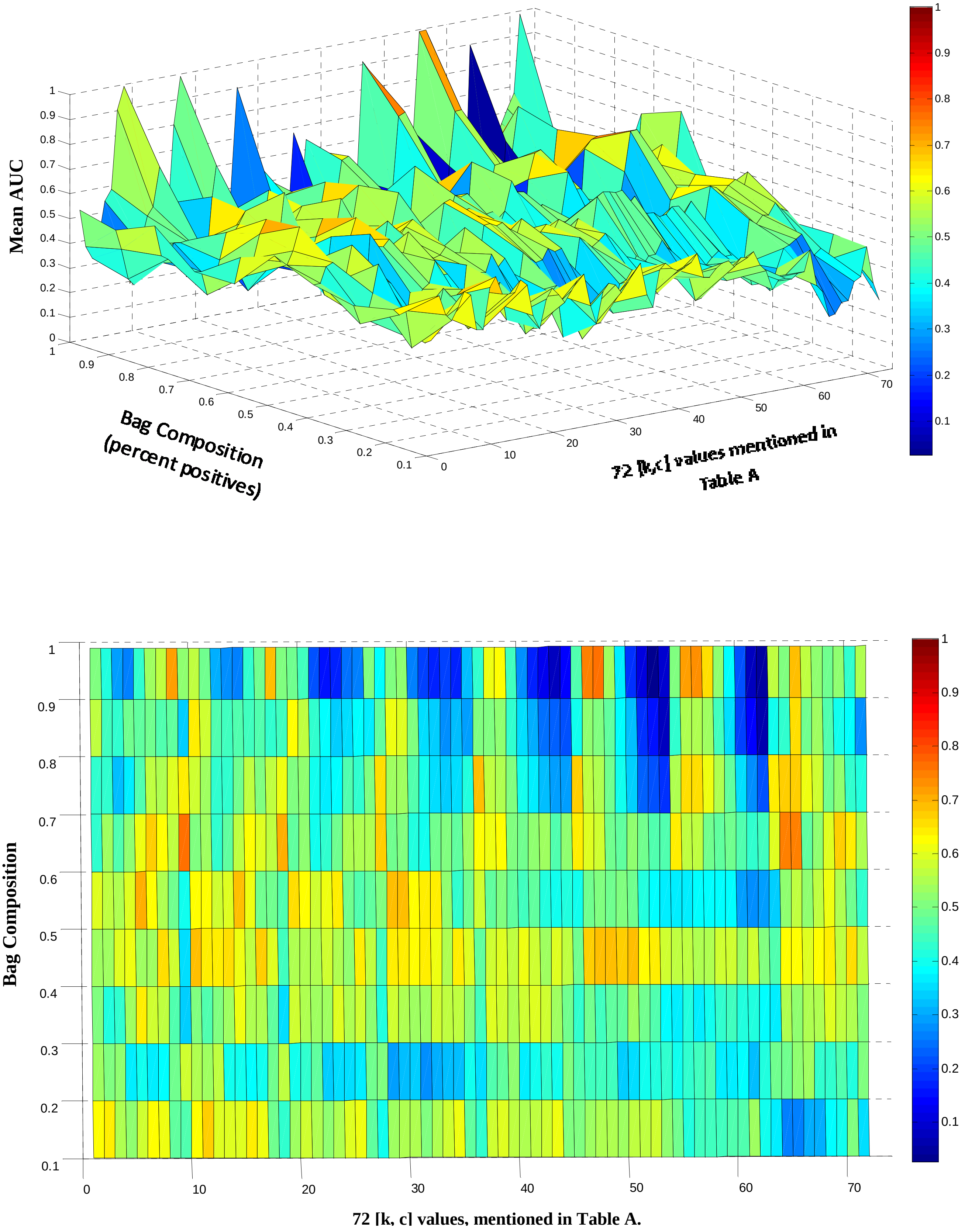


Figure A.4: Performance of CkNN with Minimum dist. using dataset 2.

Mean performance of CkNN over 10 different bag compositions is summarized using box plots in Figure A.5 and Figure A.6 for dataset 1 and dataset 2 respectively. These plots illustrate the performance of CkNN with Hausdorff distance in comparison to Minimum distance on each dataset. On each box, the central mark is the median of the performance, while the edges are 25th and 75th percentiles. Outliers are plotted individually, beyond the extent of whiskers. Figure A.5 suggests over all better performance of CkNN (using dataset 1) with Hausdorff distance as compared to Minimum distance particularly for small *k* and *c* combination values. For dataset 2, CkNN performs better with Hausdorff distance as compared to Minimum distance as shown in Figure A.6.

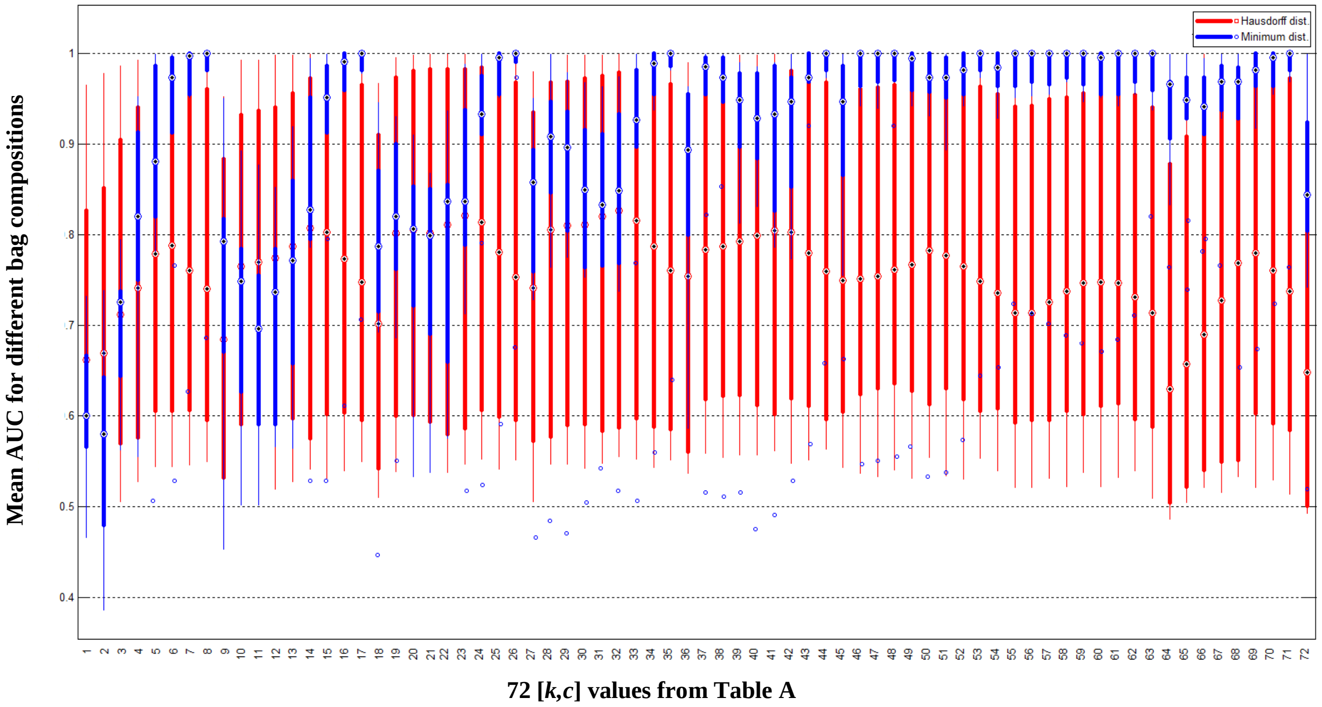


Figure A.5: Performance of CkNN with Hausdorff dist. in comparison to Minimum dist. using dataset 1.

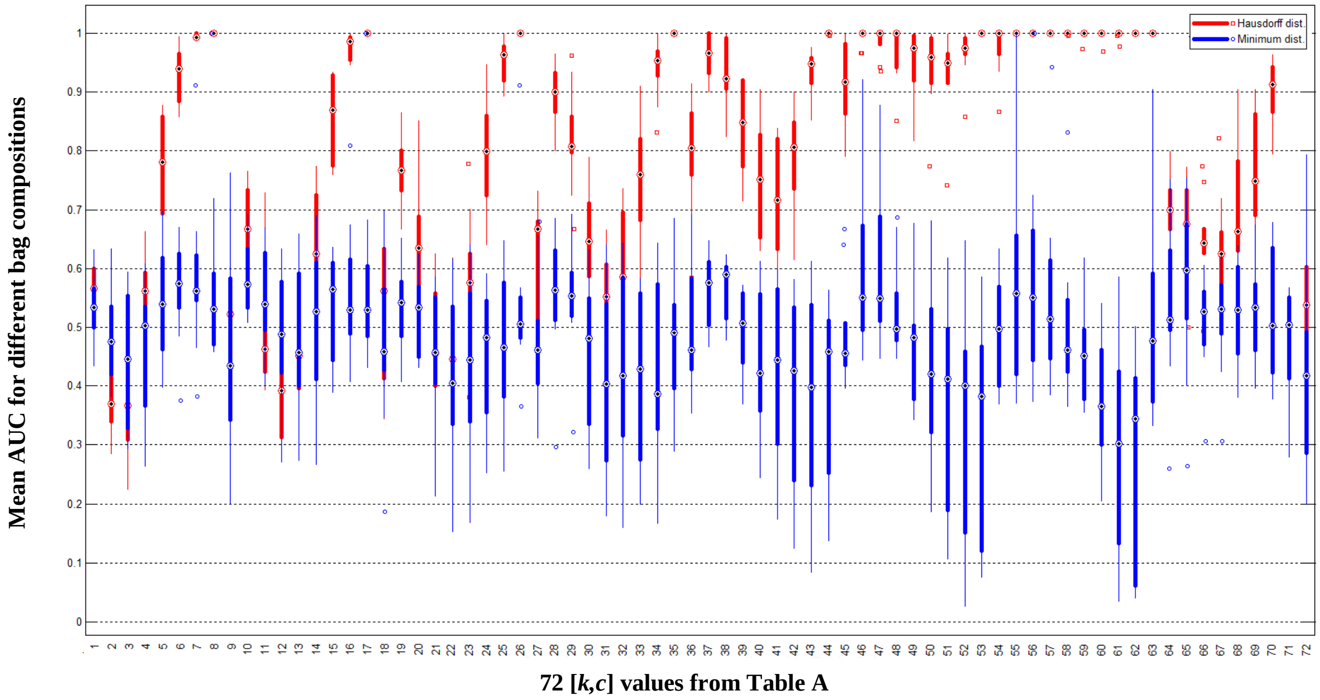


Figure A.6: Performance of CkNN with Hausdorff dist. in comparison to Minimum dist. using dataset 2.

Appendix B: ROC curves from Chapter 3

Average ROC curves from CADx and CADe studies (section 3.4) are presented below. The corresponding AUCs are also provided. The circles indicate the specificity and sensitivity at the current (default) threshold. These curves illustrate either better or identical performance of CkNN to the tile-based and ROI-based SIL for the classification of benign and malignant lesions.

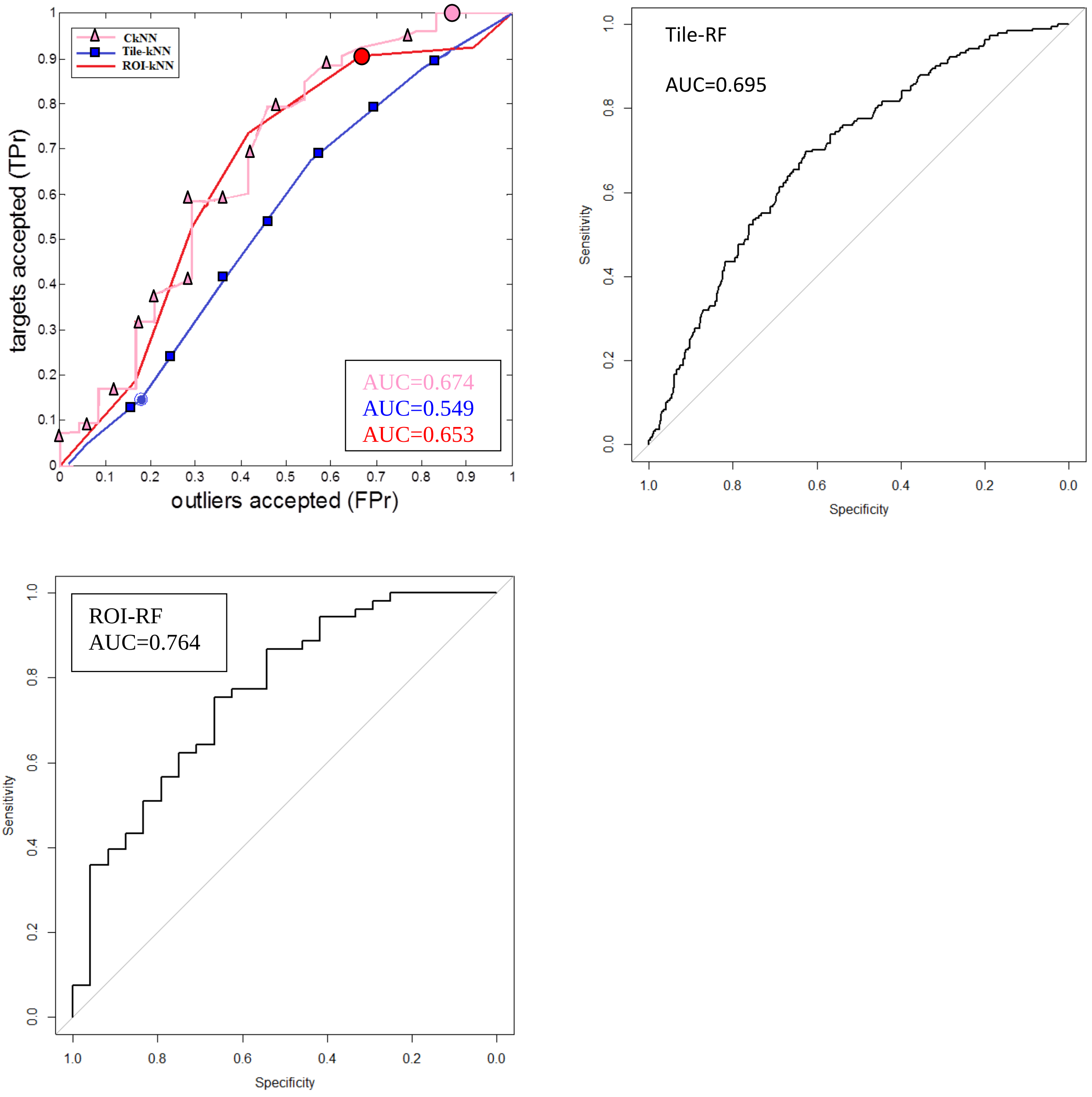


Figure B.1: Average ROC curves from CADx study using dataset A.

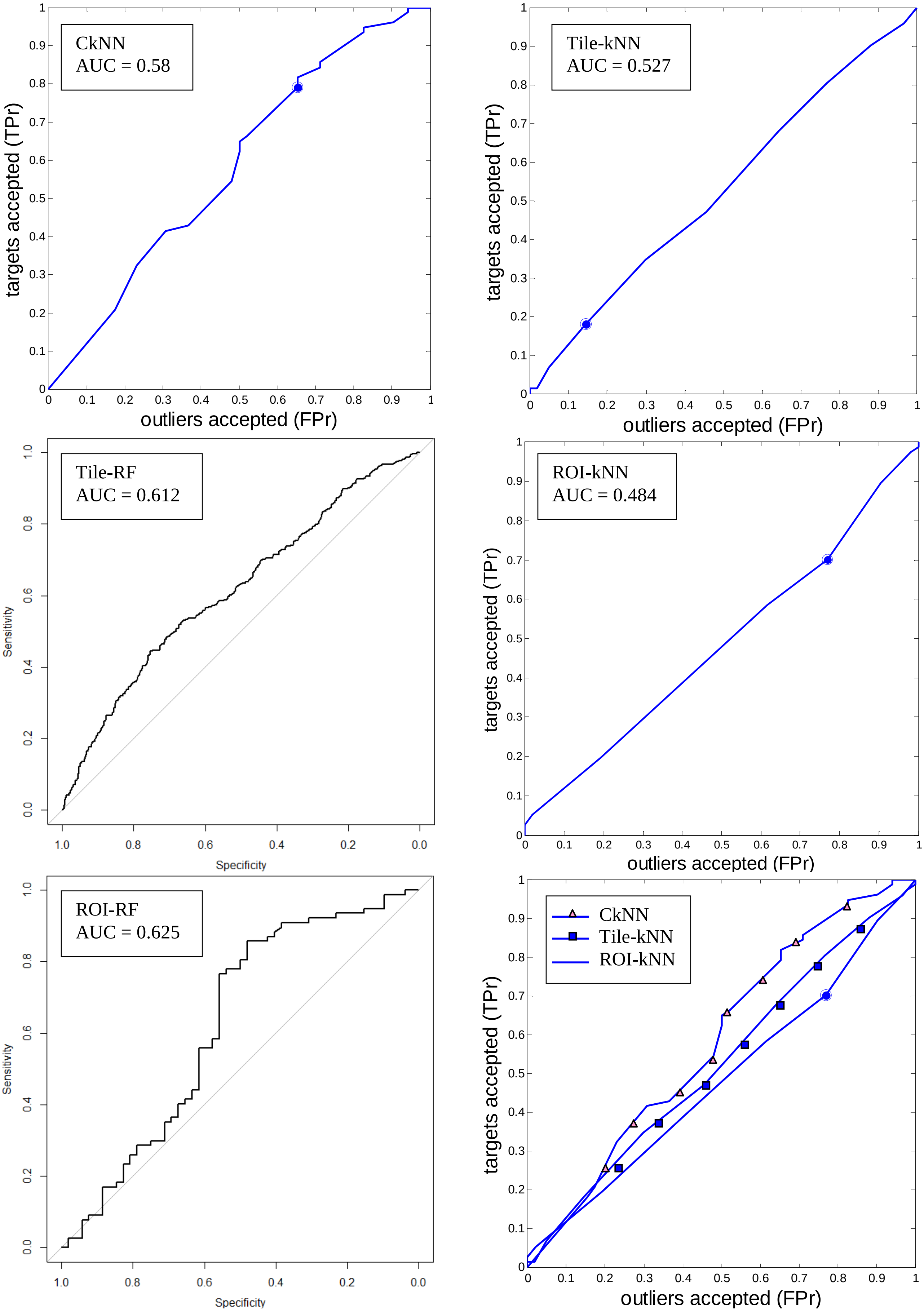


Figure B.2: Average ROC curves from CADx study using dataset B.

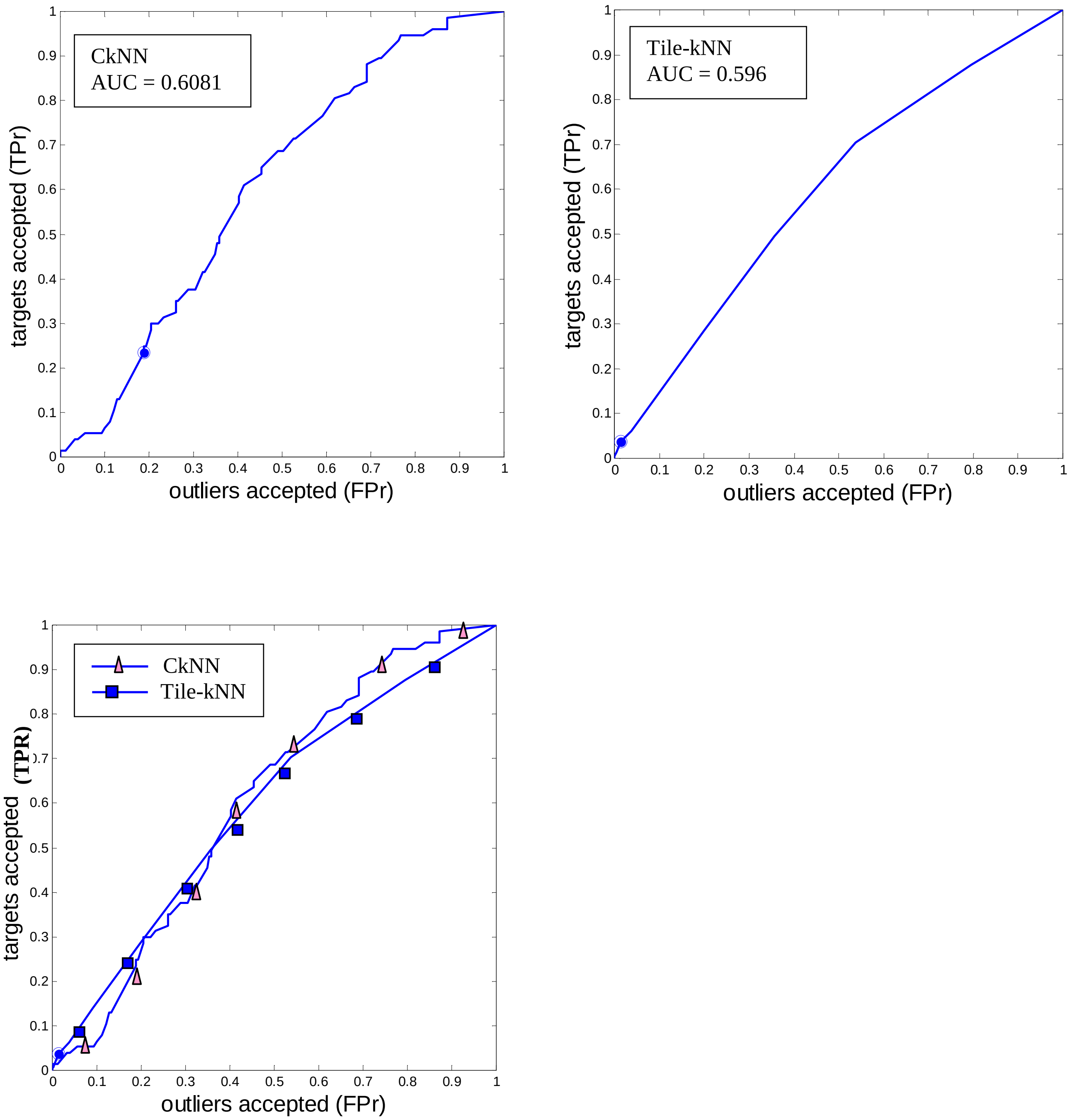


Figure B.3: Average ROC curves from CADe study.

Appendix C: Tuned parameter values

In subsection *E* of section 4.2, due to limited amount of data we optimise the parameters of the learners used in the 3D-DCT by randomly selecting coefficients from the 2D-DCT on T2-w MRI. The tuned parameter values for CkNN and kNN in each fold are illustrated in Table C.

Table C: Tuned parameter values of the learners for the 3D-DCT spatio-temporal features

Validation Fold	Tuned parameters of CkNN [*k*, *c*, dist. function]	Tuned parameters of kNN (*k* values)
1	[1, 1, Hausdorff]	1
2	[1, 7, Hausdorff]	17
3	[1, 0, Hausdorff]	18
4	[7, 2, Hausdorff]	1
5	[1, 5, Minimum]	10
6	[1, 4, Hausdorff]	11
7	[2, 2, Hausdorff]	16
8	[2, 0, Hausdorff]	18
9	[2, 0, Hausdorff]	12
10	[9, 5, Hausdorff]	18

Appendix D: ROC curves from Chapter 4

Average ROC curves using 3D-DCT, and 2D-DCT plus temporal features are presented below. The corresponding AUCs are also provided. The circles indicate the specificity and sensitivity at the current (default) threshold. These curves demonstrate almost identical performance of CkNN to the tile-based and ROI-based SIL for the classification of benign and malignant lesions. Moreover, comparison of mean ROC curves using 3D-DCT features (Figure D.1) against those of 2D-DCT plus temporal features (Figure D.2) suggests better performance of learners using 3D-DCT features.

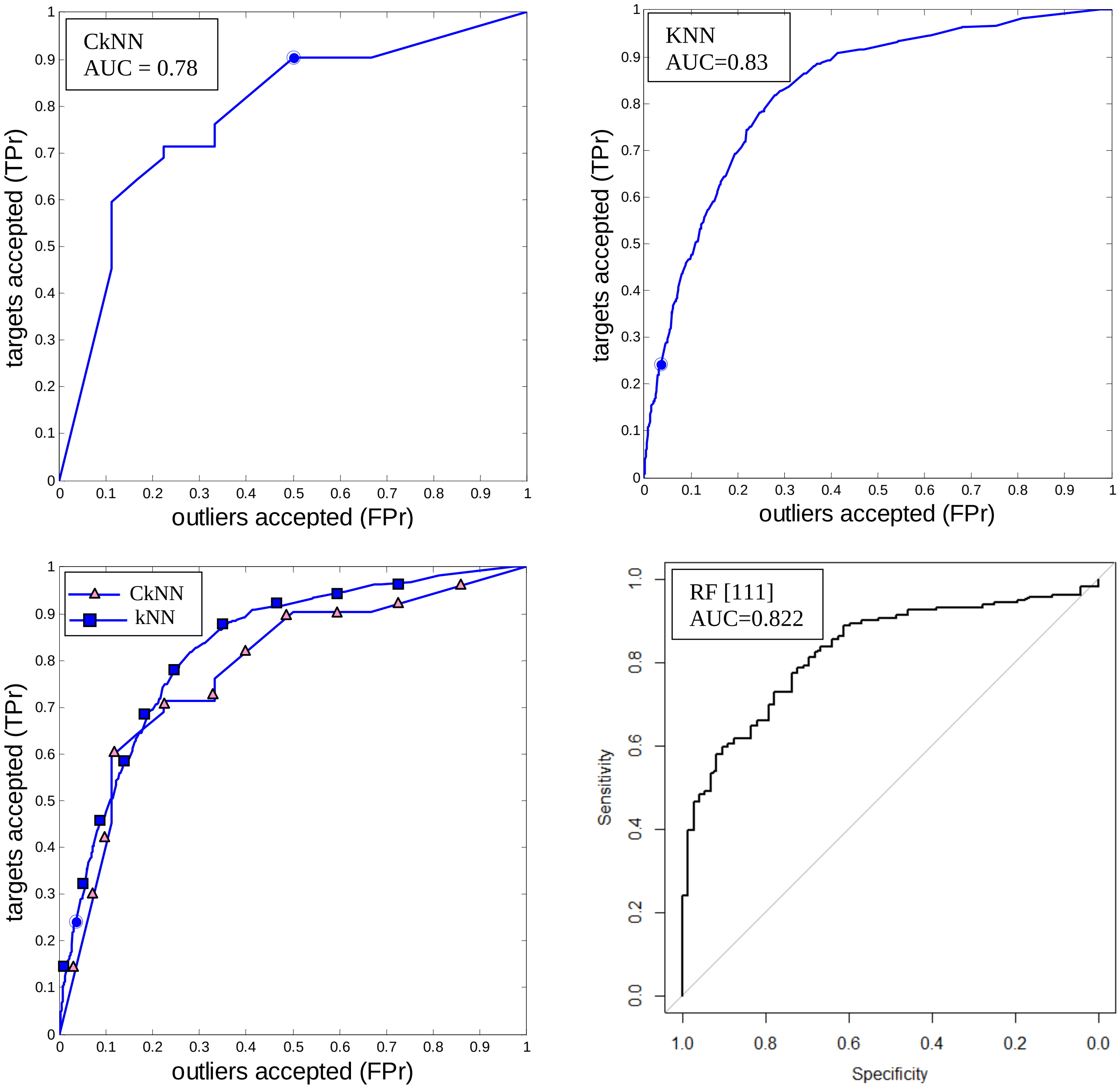


Figure D.1: Average ROC curves using the 3D-DCT features.

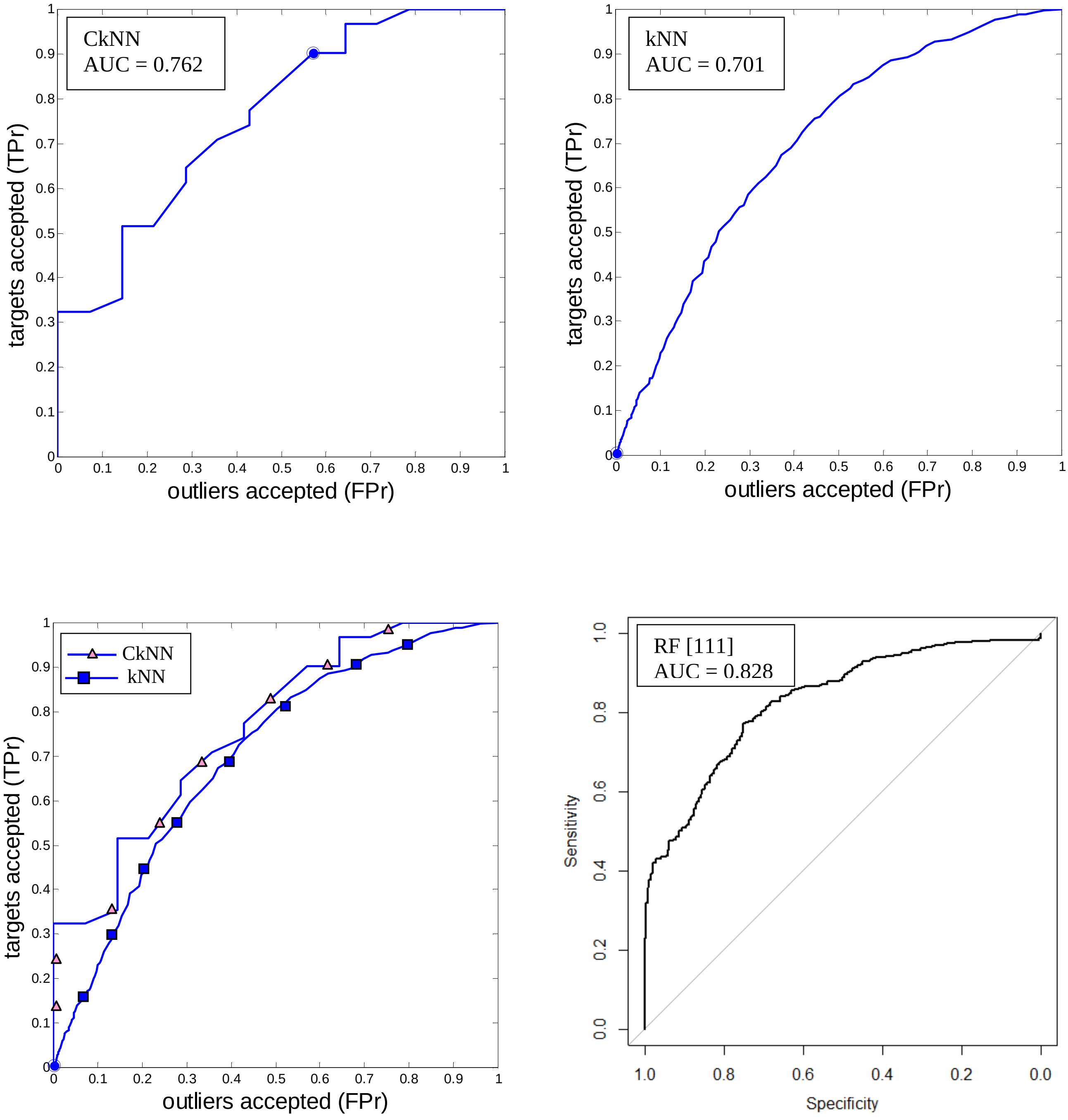


Figure D.2: Average ROC curves using 2D-DCT plus temporal features.