A QUANTITATIVE DATA REPRESENTATION FRAMEWORK FOR STRUCTURAL AND FUNCTIONAL MR IMAGING WITH APPLICATION TO PROSTATE CANCER DETECTION

BY SATISH EASWAR VISWANATH

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

Graduate Program in Biomedical Engineering

Written under the direction of

Anant Madabhushi

and approved by

____

____

____

____

____

____

____

New Brunswick, New Jersey

May, 2012

ABSTRACT OF THE DISSERTATION

A Quantitative Data Representation Framework for Structural and Functional MR Imaging with Application to Prostate Cancer Detection

by Satish Easwar Viswanath

Dissertation Director: Anant Madabhushi

Prostate cancer (CaP) is currently the second leading cause of cancer-related deaths in the United States among men, but there is a paucity of non-invasive image-based information for CaP detection and staging *in vivo*. Studies have shown the utility of multi-protocol magnetic resonance imaging (MRI) to improve CaP detection accuracy by using T2-weighted (T2w), dynamic contrast enhanced (DCE), and diffusion weighted (DWI) MRI. In this thesis, we present methods for quantitative representation of structural (T2w) and functional (DCE, DWI) imaging data with the objective of building automated classifiers to improve CaP detection accuracy *in vivo*.

In vivo disease presence was quantified via extraction of textural signatures from T2w MRI. Evaluation of these signatures showed that CaP appearance within each of the two dominant prostate regions (central gland, peripheral zone) is significantly different. A classifier trained on zone-specific features also yielded a higher detection accuracy compared to a simpler, monolithic combination of all the texture features.

While a number of automated classifiers are available, classifier choice must account for limitations in dataset size and annotation (such as with *in vivo* prostate MRI). A comprehensive evaluation of different classifier schemes was undertaken for the specific problem of automated CaP detection via T2w MRI on a zonewise basis. It was found that simple classifiers yielded significantly improved CaP detection accuracies compared to complex classifiers.

Fundamental differences must be overcome when constructing a unified quantitative representation of structural and functional MRI. We present a novel technique, referred to as *consensus embedding*, which constructs a lower dimensional representation (embedding) from a high dimensional feature space such that information (class-based or otherwise) is optimally preserved. Consensus embedding is shown to result in an improved representation of the data compared to alternative DR-based strategies in a variety of experimental domains.

A unified quantitative representation of T2w, DCE, and DWI prostate MRI was constructed via the consensus embedding framework. This yielded an integrated classifier which was more accurate for CaP detection *in vivo* as compared to using structural and functional information individually, or using a naive combination of such differing types of information.

Preface

This dissertation represents the collective published and unpublished works of the author. It is primarily composed from the content of peer-reviewed journal [1–4] and conference [5–7] articles (both published or under review), that were written by the author of this dissertation over the course of his thesis work. Additional papers published or co-published by the author that could not be included herein are available in their entirety on the lab website.

Acknowledgements

First and in most earnest, I would like to thank my advisor, Dr. Anant Madabhushi, for his help and guidance throughout the duration of this degree. It has been an honor to learn so many different things from him during this time. Sincere thanks also to each of my committee members for their valuable input and guidance in this research.

The members and staff of the Laboratory for Computational Imaging and Bioinformatics (LCIB) in the Biomedical Engineering Department at Rutgers University deserve special mention. The collaborative spirit, camaraderie, and bonhomie that exists in our lab is responsible not only for science presented herein, but also for making it fun to work on it. The experimental data and insight from our collaborators has been crucial as well as invaluable to me, the primary sources of which have been Dr. B. Nicolas Bloch, Dr Michael Feldman, Dr. Mark Rosen, Dr. John E. Tomaszewski, Dr. John Kurhanewicz, Dr. Neil Rofsky, Dr. Robert Lenkinski, and Dr. Elizabeth Genega.

This work was made possible via grants from the Department of Defense (W81XWH-08-1-0072), Wallace H. Coulter Foundation, National Cancer Institute (R01CA136535-01, R01CA140772-01, and R03CA143991-01), The Cancer Institute of New Jersey, and the Society for Imaging Informatics in Medicine.

My parents, C.V. and Shanti Easwaran, and my sister, Warsha, have been unstinting in their support of my endeavors for as long as I can remember; they deserve deep gratitude for providing love and encouragement every step of the way. Over the years my friends through undergraduate and graduate degrees have provided outlets and perspectives in ways innumerable: I consider myself lucky to have had them around.

Last and most importantly, my wife Ramya has been in turn a companion, support, guardian, disciplinarian, guide, angel, and so much more. Words cannot express how thankful I am you decided you can put up with me, and how grateful I am for you.

Dedication

To Ramya, for being everything you are, and yet so much more.

And to those who were and will be with me, I wish you were here.

If it worked right the first time, it would be called 'search'.

– Anon

It is a capital mistake to theorize before one has data.
Insensibly one begins to twist facts to suit theories,
instead of theories to suit facts.

– Sherlock Holmes
(written by Sir Arthur Conan Doyle)

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Magnetic Resonance Imaging (MRI) is widely considered a medical imaging modality of high quality, and has recently shown great promise for non-invasively determining cancer presence and extent *in vivo*. The overarching objective of the work presented in this thesis is to develop quantitative methods for detection of prostate cancer, as well as to provide a quantitative understanding of the disease characteristics on MRI.

1.1 Magnetic Resonance Imaging in prostate cancer

Prostatic adenocarcinoma (CaP) will affect one in 6 men in the United States during his lifetime, and one in 36 will die as a result of it[1]. Early detection and staging of this disease offers a 100% 5 year survival rate based on improved treatment options. Over the last decade, there has been significant interest in the use of endorectal MRI to identify prostate cancer in vivo [12, 13], allowing for improved detection and staging accuracy compared to ultrasound imaging [14]. Recent surveys of the literature [15] have reported a joint sensitivity and specificity of 71%-74% when using a 1.5 Tesla (T) endorectal T2-weighted MRI protocol. While initial enthusiasm for MRI was on account of its utility as a staging modality [16], recent developments in image-guided treatments (such as intensity-modulated radiotherapy, high-focused ultrasound therapy) as well as risk assessment, suggest that there is a growing interest in the use of MRI for accurate identification of prostate cancer stage, presence, and extent in vivo [17].

Further, the detection accuracy and qualitative characterization of prostate cancer

[1] Source: American Cancer Society, *Cancer Facts & Figures 2011*

(CaP) *in vivo* has been shown to significantly improve when multiple magnetic resonance imaging (MRI) protocols are considered in combination, as compared to using individual imaging protocols [18]. These protocols include: (1) T2-weighted (T2w), capturing high resolution anatomical information, (2) Dynamic Contrast Enhanced (DCE), characterizing micro-vascular function via uptake and washout of a paramagnetic contrast agent, and (3) Diffusion Weighted (DWI), capturing water diffusion restriction via an Apparent Diffusion Coefficient (ADC) map. DCE and DWI MRI represent functional information, which complements structural information from T2w MRI [18].

1.2 Examining prostate cancer on T2-weighted MRI

T2-weighted (T2w) MRI makes use of long echo times (TE) and repetition times (TR) during acquisition, thereby providing excellent contrast for different tissue structures based on fluid content [19]. McNeal [20] described four distinct histological regions within the prostate, which are visually differentiable on an endorectal T2w MR image

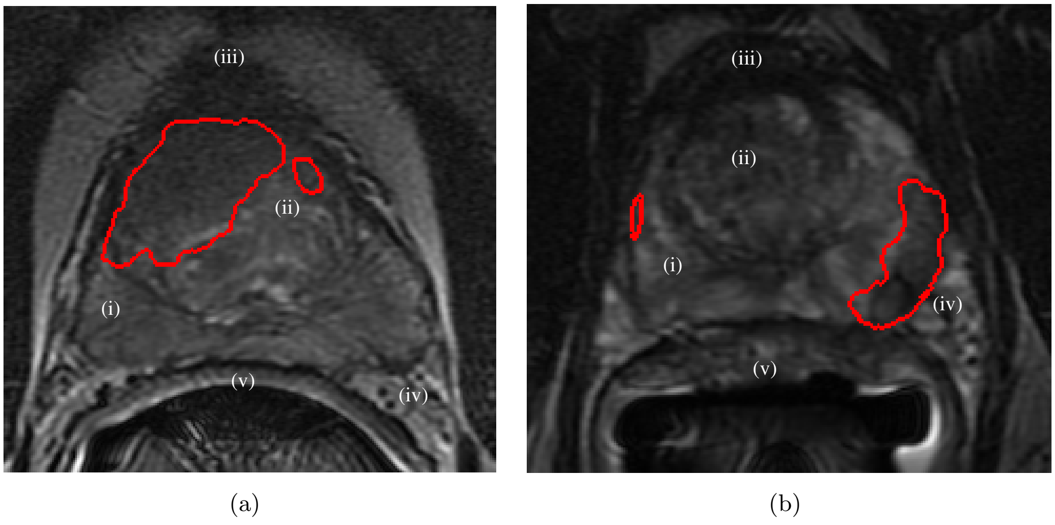


Figure 1.1: Depicting visible structures of the prostate on T2w MRI for representative slices with (a) CG CaP, and (b) PZ CaP. On (a) and (b), the red outline shows the delineation of the mapped extent of CaP presence (obtained via registration with corresponding whole-mount histology sections). Structures have been numbered on the images as follows: (i) peripheral zone (PZ), (ii) central and transitional zones jointly termed "central gland" (CG), (iii) anterior fibro-muscular stroma, (iv) neurovascular bundle, (v) Denonvilliers fascia (rectal wall).

based on their signal intensity (SI) [21] characteristics (see Figure 1.1): the outer peripheral zone, the inner central and transitional zones, and the anterior fibro-muscular stroma. Other structures such as the prostatic capsule, the neurovascular bundle, the seminal vesicles, and the Denonvilliers fascia (rectal wall) can also be seen in significant detail [21]. On account of age, the central and transitional zone regions may not always be easily differentiable on T2w MRI; typically these regions are jointly termed the "central gland" (CG) region [22].

Due to the lack of epithelial structures in the fibro-muscular region, CaP has typically been found to occur in either the peripheral zone (PZ) or the CG, and its appearance has been found to vary as a function of its spatial location in the gland [23]. Within the hyper-intense appearance of PZ on T2w MRI, CaP nodules appear as a region of low SI with incomplete stromal septations within its focus [24]. By comparison, CaP nodules in the CG are discerned as a region of purely homogeneous low SI with ill-defined margins and a lenticular (lens-like) shape. This is in contrast to the generally heterogeneous appearance of the CG region on T2w MRI due to nodular areas of varying SI [22,23]. While a large majority ($\sim 75\%$) of prostate cancers occur in the PZ [24], accurate localization of CaP within the CG or the PZ is extremely important as CG cancers tend to have lower Gleason scores, suggesting that patients with CG tumors might be candidates for less aggressive therapy and/or active surveillance [25]. Additionally, accurate knowledge of the location of CaP within the gland can help in surgical planning by identifying accurate surgical margins [26], targeting biopsies in a more directed fashion [17], as well as in radiation and focal ablation therapy planning to minimize treatment to benign areas [16].

Visual characterization of CaP on T2w MRI has been found to be a function of the resolution and contrast of the imaging technique [15]. Bloch et al [27] demonstrated improved signal to noise ratio (SNR) and high spatial resolution using 3 T endorectal MRI, allowing visualization of microscopic and pathologic details that were invisible at 1.5 T. However, the introduction of a bias field in regions close to the endorectal probe [28] significantly affects visual quality of endorectal T2w MRI; an acquisition artifact that is exacerbated by increasing magnetic field strength [22]. Other acquisition artifacts which can affect expert assessment of endorectal T2w MRI include “ghosting” (due to patient motion) as well as ambient noise [28].

1.3 Identifying quantitative features specific to CaP appearance on T2w MRI

Local staging of CaP in either the CG or PZ of the prostate via T2w MRI has shown significant variability in terms of accuracy (54%-93%) as well as inter-observer agreement [15]. For example, the general appearance of an area of low SI in the PZ may be confounded by the presence of non-malignant disease such as prostatitis, fibro-muscular benign prostatic hyperplasia (BPH), or post-biopsy hemorrhagic change [19, 21]. A significant CaP confounder in the CG particularly is BPH, which has significant overlap with CaP in terms of SI characteristics on T2w MRI. Some researchers have shown that CG and PZ carcinomas may exhibit differences in their apparent diffusion coefficient [29] and *ex vivo* metabolic profiles [30]. However, to our knowledge, no-one has been able to demonstrate significant differences in T2w SI between CG and PZ CaP.

Based on the differences in T2w SI and appearance between CG and PZ CaP, it is reasonable to assume that there may exist quantitative imaging features (and hence a quantitative imaging signature (QIS)) which are specific to CaP appearance within each of the prostatic zones (see Table 1.1). The first problem considered in this thesis is thus to identify unique textural QISes for CaP within each of the PZ and CG. We further posit that these zone-specific QISes will be sufficiently different from each other to allow

	Qualitative appearance of CaP	Quantitative features extracted
T2w	low T2w signal intensity in peripheral zone	1st order statistics, Kirsch/Sobel (gradients), 2nd order co-occurrence (Haralick)
DCE	distinctly quicker contrast enhancement for CaP compared to benign	Multi-time point intensity information
DWI	significantly low ADC compared to benign	ADC values, gradients 1st and 2nd order statistics

Table 1.1: Qualitative CaP appearance on multi-parametric MRI and corresponding quantitative features used.

for building of computerized image-based classifiers for detecting tumor occurrence (on a per voxel basis) in the CG and PZ respectively.

1.4 Developing an automated classifier for CaP detection on T2w MRI

Pattern recognition approaches to distinguishing between object classes (diseased/ normal or cancerous/benign) on medical imagery [31–34] typically involve training classifiers with features extracted from the image. Depending on the type of medical imaging data and the specific classes to be discriminated, a variety of image-derived features have been proposed and evaluated [35, 36]. A problem that has perhaps not received as much attention is the choice of classifier scheme for a particular computer aided diagnosis (CAD) problem [37]. The advent of ensemble schemes (bagging [38], boosting [39]) to overcome known shortcomings of classifier algorithms with respect to bias and variance [40] have further expanded the choices available when choosing an optimal classifier for the pattern recognition task at hand.

The challenges to developing classifiers in the context of medical imaging based CAD problems are two-fold. First, there is the issue of accurately determining class labels (ground truth) for the target region, and secondly, small dataset sizes for classifier training and evaluation. With medical imaging data, target class label determination (e.g. presence of micro-calcifications on mammograms) can usually only be performed by an expert with the appropriate domain knowledge. Expert delineation of the target class for most medical imaging datasets is typically laborious and time-consuming, as well as being subject to inter- and intra-observer variability [41] (e.g. Gleason grading of prostate cancer histopathology specimens [42–45]). Most medical imaging datasets also suffer from the small training set size problem (or Hughes phenomenon [46]), which means that classifiers may not be (a) reproducible, and (b) generalizable to new unseen, test data.

The second major problem considered in this study is hence to determine the optimal classifier in the context of the CAD problem of identifying presence of CaP from pre-operative, endorectal, *in vivo* MRI. The particular dataset and problem considered in this work also serve to illustrate the difficulty in accurately determining ground truth for the target class for classifier training and evaluation in most medical image analysis problems. This study considers a cohort of patients previously identified with CaP on prostate biopsy and scheduled for a radical prostatectomy. Prior to surgery, these patients are imaged via high resolution MRI. The excised prostatectomy sections are then annotated for CaP presence and extent by a pathologist (considered to be the gold standard for CaP ground truth). Even though the number of planar sections on MRI and histology will usually differ, the presence of anatomic fiducials on both modalities allows for a reasonable estimate of sectional correspondences [47]. Corresponding histology and radiology sections are then rigorously aligned via an automated non-linear registration step [11]. This allows for mapping of pathologist annotations of CaP extent from histology onto corresponding MRI; the disease mapping so established on MRI can thus be regarded as the surrogate "ground truth" for CaP extent on MRI. It is clear that this surrogate ground truth for disease extent is laden with different sources of errors, namely, (1) errors in the expert annotation of CaP extent on histology data, (2) errors in estimating correspondences between histological and MR images (true correspondences may not even exist), and (3) registration errors in mapping CaP extent onto MRI from histology. It goes without saying that the fidelity of the surrogate ground truth will have a bearing on classifier performance and hence classifier selection. In this work, we will examine this question in the context of the clinically relevant problem of detecting CaP extent on *in vivo* MRI.

1.5 Challenges in developing a representation and classification scheme for CaP detection using multi-parametric MRI

Quantitative integration of multi-channel (modalities, protocols) information allows for construction of sophisticated meta-classifiers for identification of disease presence [48, 49]. Such multi-channel meta-classifiers have been shown to perform significantly better compared to any individual data channel [48]. From an intuitive perspective, this is because the different channels of information each capture complementary sets of information. By extension, it is expected that an automated classifier for disease characterization which uses multi-channel information (termed a *meta-classifier*) will perform significantly better compared to using individual data channels.

We now consider some of the most significant challenges [49] involved in quantitatively integrating multi-parametric (T2w, DCE, DWI) MRI to construct a meta-classifier to detect CaP.

1. *Data Alignment*: This is done in order to bring the multiple channels of information (T2w, DCE, and DWI MRI) into the same spatial frame of reference. Typically, image registration techniques [11,50] can be utilized, which account for differences in resolution amongst the different protocols.

2. *Knowledge Representation*: This involves quantitative characterization of disease-pertinent information from each protocol. Towards this end, textural and functional image feature extraction schemes previously developed in the context of multi-parametric MRI may be employed [6,51].

3. *Data Fusion*: This involves combination of extracted quantitative descriptors to construct the integrated meta-classifier. This may be done via a *combination of data* (COD) strategy, where the information from each channel is combined prior to classification.

A number of COD strategies with the express purpose of building integrated quantitative meta-classifiers have recently been presented, including DR-based [52], kernel-based [53], and feature-based [54] approaches (See Figure 1.2). The alternative to COD is known as *combination of interpretations* (COI), where independent classifications based on the individual channels are combined. A COI approach has typically been shown to be sub-optimal for quantitative multi-channel data integration as inter-protocol dependencies are not accounted for [52].

The most common approach for quantitative multi-channel image data integration has involved concatenation of multi-channel features, followed by classification in the concatenated feature space [54]. However, a feature-based concatenation approach may be suboptimal from two perspectives. A naïve combination of multi-channel image features may not necessarily yield the most discriminatory combined representation of the information captured by each channel, which would significantly affect the performance of the corresponding meta-classifier. Further, a concatenation approach may not be appropriate when considering disparate channels of information (such as imaging and spectroscopy or imaging and -omics data).

A second approach involves multi-kernel learning (MKL) schemes [53] to represent and fuse multi-modal data based on choice of kernel. This approach may better account for different data channels representing significantly differing types of information. In Lanckriet et al [53] information from amino acid sequences, protein complex data, gene expression data, and protein interactions was transformed to a common kernel space. Thus despite the original data being disparate, the information could be combined and used to train a SVM classifier for classifying functions of yeast proteins. One of the challenges with MKL schemes is to identify an appropriate kernel for a particular problem, followed by learning associated weights. Further, when a large amount of information is present in each input channel, most COD methods, including MKL, suffer from the curse of dimensionality problem.

The final approach, dimensionality reduction (DR) [9], has been shown to be useful

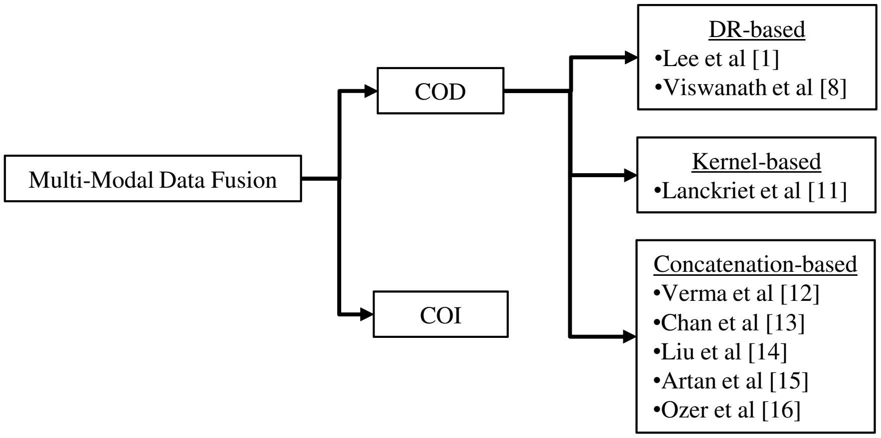


Figure 1.2: Summary of multi-modal data fusion approaches.

for quantitative data representation [52, 55] as it allows for the construction of a lower-dimensional *embedding* space which accounts for differences in scale between different channels, while also ensuring the curse of dimensionality is accounted for. While the individual channel descriptors are divorced from their physical meaning in embedding space (embedding features are not readily interpretable), relevant class-discriminatory information is largely preserved [56]. This implies that a DR-based approach may be ideal for multi-parametric representation, fusion, and subsequent classification. There is hence great need to develop a unified DR-based representation and fusion scheme for multi-parametric MRI, which may subsequently be used to accurately determine CaP presence and extent *in vivo*.

1.6 Use of dimensionality reduction for biomedical data representation and fusion

The analysis and classification of high-dimensional biomedical data has been significantly facilitated via the use of dimensionality reduction techniques, which allow classifier schemes to overcome issues such as the *curse of dimensionality*. This is an issue where the number of variables (features) is disproportionately large compared to the number of training instances (objects) [57]. Dimensionality reduction (DR) involves the projection of data originally represented in a N-dimensional (N-D) space into a lower n-dimensional (n-D) space (known as an *embedding*) such that $n << N$. DR techniques are broadly categorized as linear or non-linear, based on the type of projection method used.

Linear DR techniques make use of simple linear projections and consequently linear cost functions. An example of a linear DR scheme is Principal Component Analysis [58] (PCA) which projects data objects onto the axes of maximum variance. However, maximizing the variance within the data best preserves class discrimination only when distinct separable clusters are present within the data [59]. In contrast, non-linear DR involves a non-linear mapping of the data into a reduced dimensional space. Typically these methods attempt to project data so that relative local adjacencies between high dimensional data objects, rather than some global measure such as variance, are best preserved during data reduction from N- to n-D space [56]. This tends to better retain class-discriminatory information and may also account for any non-linear structures that exist in the data (such as manifolds), as illustrated in [60]. Examples of these techniques include locally linear embedding [60] (LLE), graph embedding [9] (GE), and isometric mapping [61] (ISOMAP). Recent work has shown that in several scenarios, classification accuracy may be improved via the use of non-linear DR schemes (rather than linear DR) for gene-expression data [56, 62] as well as medical imagery [63, 64].

However, typical DR techniques such as PCA, GE, or LLE may not guarantee an optimum result due to one or both of the following reasons:

- Noise in the original N-D space tends to adversely affect class discrimination, even if robust features are used (as shown in [65]). A single DR projection may also fail to account for such artifacts (demonstrated in [66, 67]).

- Sensitivity to choice of parameters being specified during projection; e.g. in [68] it was shown that varying the neighborhood parameter in ISOMAP can lead to significantly different embeddings.

1.7 Developing a novel DR-based representation scheme: Consensus Embedding

In this work, we present a novel DR scheme known as *consensus embedding* which aims to overcome the problems of sensitivity to noise and choice of parameters that plague several popular DR schemes [66–68]. The spirit behind consensus embedding is to construct a single stable embedding by generating and combining multiple uncorrelated, independent embeddings; the hypothesis being that this single stable embedding will better preserve specific types of information in the data (such as class-based separation) as compared to any of the individual embeddings. Consensus embedding may be used in conjunction with either linear or non-linear DR methods and, as we will show, is intended to be easily generalizable to a large number of applications and problem domains.

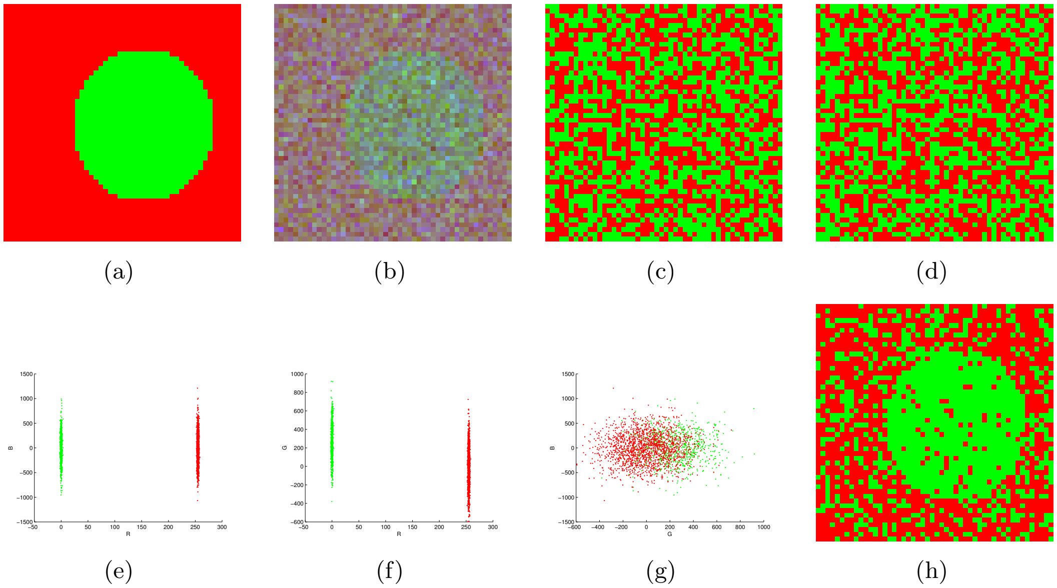


Figure 1.3: (a) Original RGB image to which Gaussian noise was added to create (b) noisy RGB image. Image visualization of classes obtained by replicated k-means clustering [8] of all the pixels via (c) original noisy RGB space, and (d) graph embedding [9] of noisy RGB data. 2D plots of (e) **R**-**G**, (f) **R**-**B**, and (g) **G**-**B** planes are also shown where colors of objects plotted correspond to the region in (b) that they are derived from. The discriminatory 2D spaces ((e) and (f)) are combined via consensus embedding, and the visualized classification result is shown in (h). Note the significantly better image partitioning into foreground and background of (h) compared to (c) and (d).

Figure 1.3 illustrates an application of consensus embedding in separating foreground (green) and background (red) regions via pixel-level classification. Figure 1.3(a) shows a simple RGB image to which Gaussian noise was added to the **G** and **B** color channels (see Figure 1.3(b)). We now consider each of the 3 color channels as features (i.e. $N = 3$) for all of the image objects (pixels). Classification via replicated k-means clustering [8] of all the objects (without considering class information) was first performed using the noisy RGB feature information (Figure 1.3(b)), in order to distinguish the foreground from background. The labels so obtained for each object (pixel) are then visualized in the image shown in Figure 1.3(c), where the color of the pixel corresponds to its cluster label. The 2 colors in Figure 1.3(c) hence correspond to the 2 classes (clusters) obtained. No discernible regions are observable in this figure. Application of DR (via GE) reduces the data to a n = 2-D space, where the graph embedding algorithm [9] non-linearly projects the data such that the object classes are maximally discriminable in the reduced dimensional space. However, as seen in Figure 1.3(d), clustering this reduced embedding space does not yield any obviously discernible image partitions either.

By plotting all the objects onto 2D plots using only the **R**-**G** (Figure 1.3(e)) and **R**-**B** (Figure 1.3(f)) color channels respectively, we can see that separation between the two classes exists only along the **R** axis. In contrast, the 2D **G**-**B** plot (Figure 1.3(g)) shows no apparent separation between the classes. Combining 1D embeddings obtained via applying graph embedding to Figures 1.3(e) and (f), followed by unsupervised clustering, yields the consensus embedding result shown in Figure 1.3(h). Consensus embedding clearly results in superior background/foreground partitioning compared to the results shown in Figures 1.3(c), (d).

1.8 Application of consensus embedding for multi-parametric data representation

In our final goal, we shall examine the application of our consensus embedding framework for multi-parametric data representation and fusion in the context of integrating prostate T2w, DCE and DWI MRI for CaP detection. The information available from each protocol is characterized via a number of quantitative descriptors [51], via application of different feature extraction schemes. Rather than make use of a direct concatenation of all the multi-parametric image features, we utilize an ensemble of embedding representations of the multi-parametric feature data. The final resulting representation is then used to train an automated classifier in order to detect CaP presence on a per-voxel basis from multi-parametric MRI.

1.9 Summary of the major goals of this thesis

We now summarize the major goals from the preceding sections. The remainder of this thesis is presented based on examining each of these topics in turn.

1. Identification of unique quantitative signatures for CaP presence on T2w MRI, stratified by the presence of cancer within each of the PZ and CG.

2. Determination of the optimal automated classifier for identifying presence of CaP from pre-operative, endorectal, *in vivo* MRI, as well as generalizable trends for how best to determine choice of classifier in the context of medical imaging data cohorts.

3. Development of theoretic and algorithmic intuition for a novel DR-based representation scheme known as *consensus embedding*.

4. Demonstrating the application of the consensus embedding methodology within a unified representation, fusion, and classification framework for CaP detection from multi-parametric MRI.

1.10 Organization of this thesis

The organization of the rest of this thesis is as follows. In Chapter 2, existing literature concerning each of the different goals in this thesis is reviewed, and the specific novel contributions of this thesis are presented. In Chapter 3, the consensus embedding methodology is described with associated definitions, theory, and algorithms. Empirical evaluation of consensus embedding for representation and classification in the context of different applications and datasets is also described. Specifics about experimental design, including prostate MRI data acquisition and pre-processing, is described in Chapter 4. Chapters 5 and 6 then present the experimental results for determining quantitative CaP-specific signatures as well as an optimal automated classification scheme for CaP detection from *in vivo* MRI. In Chapter 7, we present the application of consensus embedding for unified representation and fusion of structural and functional MRI resulting in improved CaP detection. Finally, in Chapters 8 and 9, we present our concluding remarks, and suggest directions for future work.

Chapter 2

Previous work and novel contributions

We now discuss relevant work under each of the primary goals set out in the previous chapter. We conclude this chapter by summarizing the major novel contributions of the current work.

2.1 Theoretic and empirical comparisons of generalized automated classification algorithms

The properties of ensemble classifier algorithms and the bounds on their performance (in conjunction with different classifiers) have previously been discussed by several researchers [69, 70]. In a seminal study, Breiman [70] decomposed different classifiers in terms of their bias and variance characteristics to determine bounds on performance. Discriminant analysis (DA) methods were identified as being simpler and demonstrating low variance, leading to possibly worse performance in conjunction with bagging or boosting. By comparison, the Bayes classifier was identified as being more complex with a high bias. However, these theoretical comparisons (borne out by simulated synthetic data results) offer little insight into the generalizability of classifiers on new, unseen testing data. This has therefore resulted in a large number of empirical studies comparing the performance of different classifiers [69, 71–75] on standardized datasets [76] for which unambiguous ground truth class labels are available. However it is not clear whether classifier trends reported in these studies hold in the presence of limited training data and where the validity of class labels is questionable. In fact, even with large synthetic datasets, there has been lack of concordance in the conclusions derived from classifier comparison studies with regard to the optimal classifier. For instance, Hamza et al [74] concluded that bagging and boosting performed comparably, while Opitz et al [75] determined that bagging performed worse than boosting, even if the latter tended to overfit to noisy data. However, Dietterich [73] showed that the performance of either of bagging or boosting was dependent on the level of noise present; as noise increased significantly, bagging could outperform boosting. These seemingly contradictory conclusions from different classifier comparison studies on standardized datasets with unambiguous class labels suggests that the determination of the optimal classifier may be even more difficult for real-world medical imaging studies.

2.2 Developing automated classifiers and examining quantitative features for different diseases via medical imagery

Previous related work comparing classifier strategies on biomedical data has been primarily in the context of high-dimensional molecular (gene- and protein-expression) data [77–80]. In most of these studies, SVMs were typically identified as the most accurate classifier. This conclusion has also been reported in large-scale classifier comparison studies [72, 81], implying that SVMs may work optimally when considering large, well-annotated cohorts (i.e. non SSS-QCL problems). Previous empirical studies that have examined the performance of different classifier strategies specifically for medical imaging cohorts [31–34, 82] are also not of the SSS-QCL variety, as they employ large datasets with largely unambiguous class labels and annotations. Hence, both the choice of classifier as well as classifier trends in these studies [31–34, 82] tend to mirror conclusions arrived at from other large-scale comparison studies involving natural and synthetic data [69, 71, 72, 81]. Many CAD algorithms [51, 52, 83–88], however, do not have the advantage of the size or the quality of the training sets employed in these studies [31–34, 82].

For example, Wei et al [34] found that SVMs yielded the best classifier performance in detecting histologically-proven micro-calcifications on digital mammograms from a cohort of 697 patients. Similarly, Juntu et al [32], who considered 135 cases of histologically-proven soft-tissue tumors, also found that SVMs offered the best performance in distinguishing between benign and malignant tumors on MR images. Kadah et al [33] found that a multi-layer neural network offered the best performance for identifying liver disease on a cohort of 120 pathology-investigated ultrasound patient images, but also found that simple classifiers (such as kNN [89]) could provide comparable performance. Frame et al [31] considered a QCL problem for manually annotated retinal fluorescein angiographic images for detecting diabetic retinopathy on a moderately sized dataset (88 studies). They found that a simple (rule-based) classifier offered the best diagnostic performance compared to linear DA (LDA) and neural networks. Interestingly, Schmah et al [82], in attempting to classify stroke characteristics in 9 fMRI patient studies found that both quadratic DA (QDA) and SVMs offered the best classifier performance for their SSS/non-QCL problem. These findings [31–34, 82] serve to illustrate that (a) classifier trends are widely divergent for different domains and datasets, and (b) the optimal classifier for a specific problem appears to be sensitive to the size of the dataset and the quality of the class labels.

2.3 Recent work on quantitative approaches to CaP detection on MRI

Previously, Seltzer et al [90] showed that the combination of a radiologist and a computerized decision support system significantly improves the accuracy of CaP staging on 1.5 T endorectal T2w MRI. A visual study of fractal texture on prostate MRI was conducted by Lv et al [91], who determined that fractal features showed statistically significantly different values for CaP compared to benign regions on 1.5 T T2w MRI. However, to our knowledge, there has been no previous work quantitatively examining the textural appearance of CaP in a manner specific to the zonal location of disease within the prostate.

Chan et al [83] considered a combination of co-occurrence based [92], discrete cosine transform, and co-ordinate features within an automated classification scheme to obtain an accurate statistical map for CaP presence on 11 *in vivo* 1.5 T endorectal multi-protocol (line-scan diffusion, T2-mapping, and T2w) MRI. They compared the use of maximum likelihood, LDA, and SVM classifiers to detect CaP in the PZ alone; SVMs were identified as offering the best performance; however they did not converge for all the experiments conducted. Madabhushi et al [51] presented a machine learning scheme which intelligently combined 3D texture features (1st order statistical [93], Haralick co-occurence [92], Gabor wavelet [94]) to analyze 4 T *ex vivo* T2w prostate MRI for CaP presence. The system was found to compare favorably to expert radiologists in terms of CaP detection accuracy, with improved reproducibility. Vos et al [84,95] demonstrated the efficacy of SVMs for classifying MP MRI (T2w, dynamic contrast enhanced) to detect CaP in the PZ using 34 patient studies. Vos et al [96] more recently presented a fully automated computer-aided detection method which utilized a concatenation of percentile-based descriptors of T2w, ADC, and pharmacokinetic parameters to yield a region-based classification for CaP presence (for screening purposes). Yetik et al [85,97] compared the use of an unsupervised Markov Random Field algorithm with supervised algorithms (SVMs, relevance vector machines) to detect CaP in the PZ via 20 MP MRI (T2w, dynamic contrast enhanced, diffusion-weighted) patient studies; the supervised methods outperforming the unsupervised method. These classifiers also only employed T2w MRI signal intensity, as opposed to using any textural representations of the original image data. Lopes et al [86] showed that a machine learning classifier employing fractal and multi-fractal features from 1.5 T T2w prostate MRI was more accurate in detecting CaP compared to more traditional texture features (Haralick co-occurence [92], Gabor [94], and Daubechies [98] wavelet features). They also found that SVMs and boosted decision stumps yielded comparable classification accuracies when utilizing fractal features to detect regions of CaP within the PZ in 17 T2w MRI patient studies. Tiwari et al [52] examined SVMs, Probabilistic Boosting Trees (PBTs), and Random Forests for detecting CaP on MP MRI (T2w, MR spectroscopy), and found SVMs offered the best classification performance over 36 patient studies, albeit by a very small margin.

In summary, previous related work in computerized decision support and computer aided detection of CaP from MRI [83–86] suffer from the following limitations,

- Differences in spatial location of disease within the gland have not been considered when attempting to identify CaP presence (i.e. the classifiers are monolithic since they do not attempt to distinguish between the spatially distinct types of CaP). In fact, most CaP detection schemes have been restricted to within the PZ alone, possibly due to the difficulty in obtaining ground truth annotations of the target (CaP) class and associated class labels.

- Complex classifiers (such as SVMs or DTs) have been utilized for this purpose, despite being demonstrated primarily on cohorts of limited size (5-36 patient studies), where class labels (on a per-voxel basis) were obtained by expert radiologist annotations of disease presence (on MRI) to train and evaluate the classifier.
- Most importantly, those approaches concerned with automated CaP detection via multi-parametric MRI have largely adopted a feature concatenation approach to combining and representing the multi-channel image information [83, 85, 96, 97].

There are hence a number of open problems related to quantification, representation, and classification of multi-parametric MRI data for CaP detection. As explained in the previous Chapter, DR-based representation may be considered most optimal for representation and fusion of multi-parametric data, however, there are significant issues which must be overcome with DR methods before they can be applied in this context.

2.4 Limitations with dimensionality reduction and proposed solutions

The problems of sensitivity to noise and choice of parameters which plague DR-based representation methods have previously been addressed in classifier theory via the development of classifier ensemble schemes, such as Boosting [99] and Bagging [38]. These classifier ensembles guarantee a lower error rate as compared to any of the individual members (known as "weak" classifiers), assuming that the individual weak classifiers are all uncorrelated [40]. Similarly a consensus-based algorithm has been presented [8] to find a stable unsupervised clustering of data using unstable methods such as k-means [100]. Multiple "uncorrelated" clusterings of the data were generated and used to construct a co-association matrix based on cluster membership of all the points in each clustering. Naturally occurring partitions in the data were then identified. This idea was further extended in [101] where a combination of clusterings based on simple linear transformations of high-dimensional data was considered. Note that ensemble techniques thus (1) make use of uncorrelated, or relatively independent, analyses (such as classifications or projections) of the data, and (2) combine multiple analyses (such as classifications or projections) to enable a more stable result.

2.4.1 Improved DR schemes to overcome parameter sensitivity

As shown by [61], linear DR methods such as classical multi-dimensional scaling [102] are unable to account for non-linear proximities and structures when calculating an embedding that best preserves pairwise distances between data objects. This led to the development of non-linear DR methods such as LLE [60] and ISOMAP [61] which make use of local neighborhoods to better calculate such proximities. As previously mentioned, DR methods are known to suffer from certain shortcomings (sensitivity to noise and/or change in parameters). A number of techniques have recently been proposed to overcome these shortcomings. In [103, 104] methods were proposed to choose the optimal neighborhood parameter for ISOMAP and LLE respectively. This was done by first constructing multiple embeddings based on an intelligently selected subset of parameter values, and then choosing the embedding with the minimum residual variance. Attempts have been made to overcome problems due to noisy data by selecting data objects known to be most representative of their local neighborhood (landmarks) in ISOMAP [105], or estimating neighborhoods in LLE via selection of data objects that are unlikely to be outliers (noise) [67]. Similarly, graph embedding has also been explored with respect to issues such as the scale of analysis and determining accurate groups in the data [106]. However, all of these methods require an exhaustive search of the parameter space in order to best solve the specific problem being addressed. Alternatively, one may utilize class information within the supervised variants [107, 108] of ISOMAP and LLE which attempt to construct weighted neighborhood graphs that explicitly preserve class information while embedding the data.

2.4.2 Learning in the context of dimensionality reduction

The application of classification theory to DR has begun to be explored recently. Athitsos et al presented a nearest neighbor retrieval method known as BoostMap [109], in which distances from different reference objects are combined via boosting. The problem of selecting and weighting the most relevant distances to reference objects was posed in terms of classification in order to utilize the Adaboost algorithm [99], and BoostMap was shown to improve the accuracy and speed of overall nearest neighbor discovery compared to traditional methods. DR has also previously been formulated in terms of maximizing the entropy [110] or via a simultaneous dimensionality reduction and regression methodology involving Bayesian mixture modeling [111]. The goal in such methods is to probabilistically estimate the relationships between points based on objective functions that are dependent on the data labels [110]. These methods have been demonstrated in the context of application of PCA to non-linear datasets [111]. More recently, multi-view learning algorithms [112] have attempted to address the problem of improving the learning ability of a system by considering several disjoint subsets of features (views) of the data. The work most closely related to our own is that of [113] in the context of web data mining via multi-view learning. Given that a hidden pattern exists in a dataset, different views of this data are each embedded and transformed such that known domain information (encoded via pairwise link constraints) is preserved within a common frame of reference. The authors then solve for a consensus pattern which is considered the best approximation of the underlying hidden pattern being solved for. A similar idea was examined in [114, 115] where 1D projections of image data were co-registered in order to better perform operations such as image-based breathing gating as well as multi-modal registration. Unlike consensus embedding, these algorithms involve explicit transformations of embedding data to a target frame of reference, as well as being semi-supervised in encoding specific link constraints in the data.

2.5 Novel contributions of the current work

2.5.1 Developing quantitative imaging signatures for CaP on T2w MRI

The first goal of this work comprises identifying distinct zone-specific quantitative textural models of CaP using 3 T endorectal T2w MRI, in order to construct QISes which allow for accurate discrimination between CG and PZ CaP. The overarching application of our work is to build an automated decision support system for improved detection and localization of CaP presence at high-resolution (per-voxel) using T2w MRI; the use of zone-specific models is expected to significantly improve the sensitivity and specificity of CaP detection of such a system. We will show that building zone-specific classifiers result in an improved detection of CaP presence on MRI compared to monolithic classifiers. In order to define zone-specific QISes, we will utilize texture-based characterization of T2w MRI data. The underlying hypothesis for our approach is that the CaP appearance within each of the CG and PZ has distinct textural signatures, i.e. we are attempting to quantify visual characteristics of regions with incomplete stromal septations (corresponding to PZ CaP) or those with homogeneous appearance (corresponding to CG CaP). The result of applying texture analysis algorithms [92–94, 98] to prostate T2w MRI is that every voxel in a dataset will now be associated with a set of numerical values (reflecting corresponding textural attributes) which describe the local properties and relationships between the SI at the voxel under consideration and its surrounding local neighborhood voxels. Further, given annotations of CaP presence (labels) for each voxel within a dataset, we can train an automated classifier to utilize these textural signatures (and labels) to give us a per-voxel classification for CaP presence in a new, unseen test dataset. In this work, we will construct separate QISes for each of CG and PZ CaP. These QISes will then be used to train independent classifiers to detect CaP in the CG and PZ, respectively. Our hypothesis is that these zone-specific texture-based classifiers will be more accurate in identifying CaP compared to either (a) an intensity-based approach (which would use only the SI at every voxel to classify for CaP presence), or (b) a zone-agnostic monolithic approach (which would not consider zonal differences when constructing the classifier).

2.5.2 Determining an optimal automated classification scheme for detection of CaP on MRI

In the second goal, we aim to compare classifier performance in the context of a problem where the objective is to distinguish tumors in different prostatic zones via texture-based features derived from T2w MRI. Most supervised classifier schemes such as Bayesian learners [116], Support Vector Machines [117] (SVMs), and Decision Trees [118] (DTs) typically require a large well-labeled training set for model training. This then poses the question whether complex, popular classifiers (such as SVMs [117]), which require extensive parameter tuning and accurately annotated data, represent the appropriate choice of classifier for most CAD and medical imaging problems. Since these data cohorts are typically of the small sample size (SSS), questionable class label (QCL) variety, a simpler, non-parametric classifier requiring less training might be more desirable. In a similar vein one could ask whether bagging [38] and boosting [39], which combine decisions from multiple uncorrelated classifiers (where the individual classifiers may be unstable or marginally better than guessing), are appropriate for SSS-QCL datasets such as with medical imaging studies. The major questions that emerge specifically in the context of SSS-QCL datasets are,

- Do classifier trends on large, well annotated datasets generalize to SSS-QCL problems?
- How will more complex classifiers compare against simpler, non-parametric classifiers for SSS-QCL problems in terms of classifier accuracy, model and computational complexity?

2.5.3 Development of the consensus embedding methodology

Under the third goal of this work, we present a novel DR scheme (consensus embedding) that involves first generating and then combining multiple uncorrelated, independent (or *base*) n-D embeddings. These base embeddings may be obtained via either linear or non-linear DR techniques being applied to a large N-D feature space. Note that we use the terms "uncorrelated, independent" with reference to the method of constructing base embeddings; similar to their usage in ensemble classification literature [40]. Indeed, techniques to generate multiple base embeddings may be seen to be analogous to those for constructing classifier ensembles. In the latter, base classifiers with significant variance can be generated by varying the parameter associated with the classification method (k in kNN classifiers [119]) or by varying the training data (combining decision trees via Bagging [38]). Previously, a consensus method for LLE was examined in [120] with the underlying hypothesis that varying the neighborhood parameter (κ) will effectively generate multiple uncorrelated, independent embeddings for the purposes of constructing a consensus embedding. The combination of such base embeddings for magnetic resonance spectroscopy data was found to result in a low-dimensional data representation which enabled improved discrimination of cancerous and benign spectra compared to using any single application of LLE. In this work we shall consider an approach inspired by random forests [121] (which in turn is a modification of the Bagging algorithm [38]), where variations within the feature data are used to generate multiple embeddings which are then combined via our consensus embedding scheme. Additionally, unlike most current DR approaches which require tuning of associated parameters for optimal performance in different datasets, consensus embedding offers a methodology that is not significantly sensitive to parameter choice or dataset type.

The major contributions of the consensus embedding methodology are hence:

- A novel DR approach which generates and combines embeddings.
- A largely parameter invariant scheme for dimensionality reduction.
- A DR scheme easily applicable to a wide variety of pattern recognition problems including image partitioning, data mining, and high dimensional data classification.

2.5.4 Application of consensus embedding for CaP detection from multi-parametric MRI

Finally under the fourth goal of this work, we present a novel multi-channel data representation and fusion framework referred to as *Enhanced Multi-Protocol Analysis via Intelligent Supervised Embedding* (EMPrAvISE), which is based off consensus embedding. An overview of the different steps within the EMPrAvISE methodology is shown in Figure 2.1. First, the information available via each of the channels is quantified via multiple features. A number of embeddings are then calculated from the multi-channel

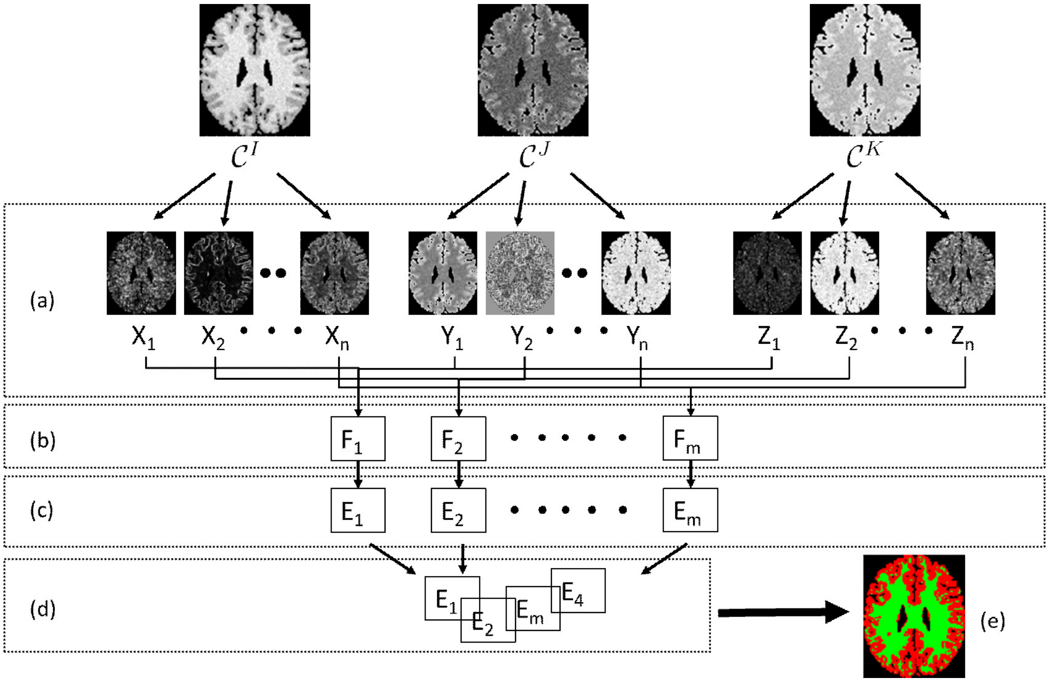


Figure 2.1: Overview of the Enhanced Multi-Protocol Analysis via Intelligent Supervised Embedding (EMPrAvISE) methodology: 3 different information channels ($\mathcal{C}^I$,$\mathcal{C}^J$,$\mathcal{C}^K$) corresponding to 3 distinct multi-parametric MR protocols acquired for a synthetic brain MRI slice [10] are shown. Note that this data comprises only grey and white matter (to ensure a 2-class problem) in which noise and inhomogeneity have been introduced, making it difficult to differentiate between the classes present. (a) Multi-parametric features $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, and $Z_1, \ldots, Z_n$ are extracted from each channel $\mathcal{C}^I$,$\mathcal{C}^J$,$\mathcal{C}^K$. (b) Multi-parametric feature subsets are constructed as $F_1 = \{X_1, Y_1, Z_1\}, F_2 = \{X_2, Y_2, Z_2\}, \cdots, F_n = \{X_n, Y_n, Z_n\}$. (c) Post DR on each of $F_1, F_2, \ldots, F_n$, embeddings $E_1, E_2, \ldots, E_n$ are calculated. (d) These embeddings are intelligently aggregated (using *a priori* class information) within EMPrAvISE. (e) Visualizing the partitioning of the EMPrAvISE representation is seen to result in an optimal white and gray matter segmentation, despite significant levels of noise in the original brain image.

feature space, which are then intelligently aggregated via a synergy of ensemble theory within dimensionality reduction. EMPrAvISE is intended to inherently account for (1) maximally quantifying information available within a multi-parametric feature space (via feature extraction), (2) differences in dimensionalities between individual T2w, DCE, DWI protocols (via DR), (3) noise and parameter sensitivity issues with DR-based representation (via the use of an ensemble theory), and (4) dependencies between the T2w, DCE, DWI protocols (via an intelligent combination scheme). We present the application of EMPrAvISE for representing multi-parametric (T2w, DCE, DWI) prostate MRI within a meta-classifier for CaP detection *in vivo.* Specifically, the major contributions and significant advances within EMPrAvISE are as follows,

- A generalizable framework for multi-channel data integration which can account for differences in dimensionality and scale, as well as relative disparity in the input information channels.

- A novel weighted embedding combination methodology which leverages ensemble theory within dimensionality reduction to result in a more discriminatory representation of input data.

- Validation of the EMPrAvISE methodology on 2 distinct multi-parametric MRI data cohorts, demonstrating significant improvements in CaP detection accuracy over current state-of-the-art representation schemes.

Chapter 3

Consensus embedding: theory, algorithms, and empirical evaluation

3.1 Theory of consensus embedding

The spirit of consensus embedding lies in the generation and combination of multiple embeddings in order to construct a more stable, stronger result. Thus we will first define various terms associated with embedding construction. Based on these, we can mathematically formalize the concept of generating and combining multiple base embeddings, which will in turn allow us to derive necessary and sufficient conditions that must be satisfied when constructing a consensus embedding. Based on these conditions we will describe the specific algorithmic steps in more detail. Notation that is used in this chapter is summarized in Table 3.1.

3.1.1 Preliminaries

An object shall be referred to by its label c and is defined as a point in an N-dimensional space $\mathbb{R}^N$. It is represented by an N-tuple $\mathbf{F}(c)$ comprising its unique N-dimensional co-ordinates. In a sub-space $\mathbb{R}^n \subset \mathbb{R}^N$ such that $n << N$, this object c in a set C is represented by an n-tuple of its unique n-dimensional coordinates $\mathbf{X}(c)$. $\mathbb{R}^n$ is also known as the *embedding* of objects $c \in C$ and is always calculated via some projection of $\mathbb{R}^N$. For example in the case of $\mathbb{R}^3$, we can define $\mathbf{F}(c) = \{f_1, f_2, f_3\}$ based on the co-ordinate locations (f_1, f_2, f_3) on each of the 3 axes for object $c \in C$. The corresponding embedding vector of $c \in C$ in $\mathbb{R}^2$ will be $\mathbf{X}(c) = \{e_1, e_2\}$ with co-ordinate axes locations (e_1, e_2). Note that in general, determining the target dimensionality (n) for any $\mathbb{R}^N$ may be done by a number of algorithms such as the one used in this work [122].

$\mathbb{R}^N$	High(N)-dimensional space	$\mathbb{R}^n$	Low(n)-dimensional space
c, d, e	Objects in set C	Z	Number of unique triplets in C
$\mathbf{F}(c)$	High-dimensional feature vector	$\mathbf{X}(c)$	Embedding vector
Λ^{cd}	Pairwise relationship in $\mathbb{R}^N$	δ^{cd}	Pairwise relationship in $\mathbb{R}^n$
$\Delta(c, d, e)$	Triangle relationship (Defn. 1)	$\psi^{ES}(\mathbb{R}^n)$	Embedding strength (Defn. 2)
$\widehat{\mathbb{R}}^n$	True embedding (Defn. 3)	$\hat{\delta}^{cd}$	Pairwise relationship in $\widehat{\mathbb{R}}^n$
$\ddot{\mathbb{R}}^n$	Strong embedding (Defn. 4)	$\dot{\mathbb{R}}^n$	Weak embedding
$\widetilde{\mathbb{R}}^n$	Consensus embedding (Defn. 5)	$\tilde{\delta}^{cd}$	Pairwise relationship in $\widetilde{\mathbb{R}}^n$
M	Number of generated embeddings	K	Number of selected embeddings
R	Number of objects in C	$\widetilde{\mathbf{X}}(c)$	Consensus embedding vector

Table 3.1: Summary of notation and symbols used in this chapter.

The notation Λ^{cd}, henceforth referred to as the *pairwise relationship*, will represent the relationship between two objects $c, d \in C$ with corresponding vectors $\mathbf{F}(c), \mathbf{F}(d) \in \mathbb{R}^N$. Similarly, the notation δ^{cd} will be used to represent the pairwise relationship between two objects $c, d \in C$ with embedding vectors $\mathbf{X}(c), \mathbf{X}(d) \in \mathbb{R}^n$. We assume that this relationship satisfies the three properties of a metric (e.g. Euclidean distance). Finally, a triplet of objects $c, d, e \in C$ is referred to as an *unique triplet* if $c \neq d$, $d \neq e$, and $c \neq e$. Unique triplets will be denoted simply as (c, d, e).

3.1.2 Definitions

Definition 1 *The function Δ defined on a unique triplet (c, d, e) is called a triangle relationship, $\Delta(c, d, e)$, if when $\Lambda^{cd} < \Lambda^{ce}$ and $\Lambda^{cd} < \Lambda^{de}$, then $\delta^{cd} < \delta^{ce}$ and $\delta^{cd} < \delta^{de}$.*

For objects $c, d, e \in C$ whose relative pairwise relationships in $\mathbb{R}^N$ are preserved in $\mathbb{R}^n$, the triangle relationship $\Delta(c, d, e) = 1$. For ease of notation, the triangle relationship $\Delta(c, d, e)$ will be referred to as Δ where appropriate. Note that for a set of R unique objects ($R = |C|$, $|.|$ is cardinality of a set), $Z = \frac{R!}{3!(R-3)!}$ unique triplets may be formed.

Definition 2 *Given Z unique triplets $(c, d, e) \in C$ and an embedding $\mathbb{R}^n$ of all objects $c, d, e \in C$, the associated embedding strength $\psi^{ES}(\mathbb{R}^n) = \frac{\sum_C \Delta(c,d,e)}{Z}$.*

This definition is based on the idea that the strength of any embedding $\mathbb{R}^n$ will depend on how well pairwise relationships are preserved from $\mathbb{R}^N$. This in turn can written in terms of the triplet relationship as,

$$\psi^{ES}(\mathbb{R}^n) = \frac{\text{total number of preserved triplets}}{\text{total number possible triplets}} = \frac{\sum_C \Delta(c,d,e)}{Z}. \tag{3.1}$$

The embedding strength (ES) of an embedding $\mathbb{R}^n$, denoted $\psi^{ES}(\mathbb{R}^n)$, is hence the fraction of unique triplets $(c,d,e) \in C$ for which $\Delta(c,d,e) = 1$.

Definition 3 *A true embedding, $\widehat{\mathbb{R}}^n$, is an embedding for which $\psi^{ES}(\widehat{\mathbb{R}}^n) = 1$.*

A true embedding $\widehat{\mathbb{R}}^n$ is one for which the triangle relationship is satisfied for all unique triplets $(c,d,e) \in C$, hence perfectly preserving all pairwise relationships from $\mathbb{R}^N$ to $\widehat{\mathbb{R}}^n$. Additionally, for all objects $c, d \in C$ in $\widehat{\mathbb{R}}^n$, the pairwise relationship is denoted as $\hat{\delta}^{cd}$.

Note that according to Definition 3, the most optimal true embedding may be considered to be the original $\mathbb{R}^N$ itself, i.e. $\hat{\delta}^{cd} = \Lambda^{cd}$. However, as $\mathbb{R}^N$ may not be optimal for classification (due to the curse of dimensionality), we are attempting to approximate a true embedding as best possible in n-D space. Note that multiple true embeddings in n-D space may be calculated from a single $\mathbb{R}^N$; any one of these may be chosen to calculate $\hat{\delta}^{cd}$.

Practically speaking, any $\mathbb{R}^n$ will be associated with some degree of error compared to the original $\mathbb{R}^N$. This is almost a given since some loss of information and concomitant error can be expected to occur in going from a high- to a low-dimensional space. We can calculate the probability of pairwise relationships being accurately preserved from $\mathbb{R}^N$ to $\mathbb{R}^n$ i.e. the probability that $\Delta(c,d,e) = 1$ for any unique triplet $(c,d,e) \in C$ in any $\mathbb{R}^n$ by utilizing the traditional formulation of a prior probability,

$$p = \frac{\text{total number of observed instances}}{\text{total number of instances}}. \tag{3.2}$$

Here, triplets are considered to be "instances". Therefore Equation 3.2 becomes,

$$p(\Delta) = \frac{\text{total number of preserved triplets (i.e. } \Delta = 1)}{\text{total number possible triplets}} = \frac{\sum_C \Delta(c,d,e)}{Z}. \tag{3.3}$$

Therefore,

$$p(\Delta | c, d, e, \mathbb{R}^n) = \frac{\sum_C \Delta(c,d,e)}{Z}. \tag{3.4}$$

Note that the probability in Equation 3.4 is binomial as the complementary probability to $p(\Delta|c,d,e,\mathbb{R}^n)$ (i.e. the probability that $\Delta(c,d,e) \neq 1$ for any unique triplet $(c,d,e) \in C$ in any $\mathbb{R}^n$) is given by $1 - p(\Delta|c,d,e,\mathbb{R}^n)$ (in the case of binomial probabilities, event outcomes can be broken down into two probabilities which are complementary, i.e. they sum to 1).

Definition 4 *A strong embedding, $\ddot{\mathbb{R}}^n$, is an embedding for which $\psi^{ES}(\ddot{\mathbb{R}}^n) > \theta$.*

In other words, a strong embedding is defined as one which accurately preserves the triangle relationship for more than some significant fraction (θ) of the unique triplets of objects $c, d, e \in C$ that exist. An embedding $\mathbb{R}^n$ which is not a strong embedding is referred to as a *weak embedding*, denoted as $\dot{\mathbb{R}}^n$.

We can calculate multiple uncorrelated (i.e. independent) embeddings from a single $\mathbb{R}^N$ which may be denoted as $\mathbb{R}^n_m, m \in \{1, \ldots, M\}$, where M is total number of possible uncorrelated embeddings. Note that both strong and weak embeddings will be present among all of the M possible embeddings. All objects $c, d \in C$ can then be characterized by corresponding embedding vectors $\mathbf{X}_m(c), \mathbf{X}_m(d) \in \mathbb{R}^n_m$ with corresponding pairwise relationship δ^{cd}_m. Given multiple δ^{cd}_m, we can form a distribution $p(X = \delta^{cd}_m)$, over all M embeddings. Our hypothesis is that the maximum likelihood estimate (MLE) of $p(X = \delta^{cd}_m)$, denoted as $\tilde{\delta}^{cd}$, will approximate the true pairwise relationship $\hat{\delta}^{cd}$ for objects $c, d \in C$.

Definition 5 *An embedding $\mathbb{R}^n$ is called a consensus embedding, $\widetilde{\mathbb{R}}^n$, if for all objects $c, d \in C$, $\delta^{cd} = \tilde{\delta}^{cd}$.*

We denote the consensus embedding vectors for all objects $c \in C$ by $\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n$. Additionally, from Equation 3.4, $p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n)$ represents the probability that $\Delta(c, d, e) = 1$ for any $(c, d, e) \in C$ in $\widetilde{\mathbb{R}}^n$.

3.1.3 Necessary and sufficient conditions for consensus embedding

While $\widetilde{\mathbb{R}}^n$ is expected to approximate $\widehat{\mathbb{R}}^n$ as best possible, it cannot be guaranteed that $\psi^{ES}(\widetilde{\mathbb{R}}^n) = 1$ as this is dependent on how well $\tilde{\delta}^{cd}$ approximates $\hat{\delta}^{cd}$, for all objects $c, d \in C$. $\tilde{\delta}^{cd}$ may be calculated inaccurately as a result of considering pairwise relationships derived from weak embeddings, $\dot{\mathbb{R}}^n$, present amongst the M embeddings that are generated. As Proposition 1 and Lemma 1 below demonstrate, in order to ensure that $\psi^{ES}(\widetilde{\mathbb{R}}^n) \to 1$, $\widetilde{\mathbb{R}}^n$ must be constructed from a combination of multiple strong embeddings $\ddot{\mathbb{R}}^n$ alone, so as to avoid including weak embeddings.

Proposition 1 *If $K \leq M$ independent, strong embeddings $\mathbb{R}^n_k, k \in \{1, \ldots, K\}$, with a constant $p(\Delta|c,\ d, e, \mathbb{R}^n_k)$ that $\Delta(c, d, e) = 1$ for all $(c, d, e) \in C$, are used to calculate $\widetilde{\mathbb{R}}^n$, $\psi^{ES}(\widetilde{\mathbb{R}}^n) \to 1$ as $K \to \infty$.*

Proof. If $K \leq M$ independent, strong embeddings alone are utilized in the construction of $\widetilde{\mathbb{R}}^n$, then the number of weak embeddings is $(M - K)$. As Equation 3.4 represents a binomial probability, $p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n)$ can be approximated via the binomial formulation of Equation 3.4 as,

$$p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n) = \sum_{K=1}^{M} \binom{M}{K} \alpha^K (1 - \alpha)^{M-K}, \tag{3.5}$$

where $\alpha = p(\Delta|c, d, e, \mathbb{R}^n_k)$ (Equation 3.4) is considered to be constant. Based on Equation 3.5, as $K \to \infty$, $p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n) \to 1$, which in turn implies that $\psi^{ES}(\widetilde{\mathbb{R}}^n) \to 1$; therefore $\widetilde{\mathbb{R}}^n$ approaches $\widehat{\mathbb{R}}^n$. □

Proposition 1 demonstrates that for a consensus embedding to be strong, it is sufficient that strong embeddings be used to construct it. Note that as $K \to M$, $p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n) >> p(\Delta|c, d, e, \mathbb{R}^n_k)$. In other words, if $p(\Delta|c, d, e, \mathbb{R}^n_k) > \theta$, $p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n) >>$

θ. Based on Equation 3.4 and Definitions 2, 4, this implies that as $K \to M$, $\psi^{ES}(\widetilde{\mathbb{R}}^n) >> \theta$. Lemma 1 below demonstrates the necessary nature of this condition i.e. if weak embeddings are considered when constructing $\widetilde{\mathbb{R}}^n$, $\psi^{ES}(\widetilde{\mathbb{R}}^n) << \theta$ (it will be a weak embedding).

Lemma 1 *If $K \leq M$ independent, weak embeddings $\mathbb{R}^n_k, k \in \{1, \ldots, K\}$, with $\psi^{ES}(\mathbb{R}^n_k) \leq \theta$, are used to calculate $\widetilde{\mathbb{R}}^n$, then $\psi^{ES}(\widetilde{\mathbb{R}}^n) << \theta$.*

Proof. From Equation 3.4 and Definitions 2, 4, if $\psi^{ES}(\mathbb{R}^n_k) \leq \theta$, then $p(\Delta|c, d, e, \mathbb{R}^n_k) \leq \theta$. Substituting $p(\Delta|c, d, e, \mathbb{R}^n_k)$ in Equation 3.5, will result in $p(\Delta|c, d, e, \widetilde{\mathbb{R}}^n) << \theta$. Thus $\psi^{ES}(\widetilde{\mathbb{R}}^n) << \theta$, and $\widetilde{\mathbb{R}}^n$ will be weak. □

Proposition 1 and Lemma 1 together demonstrate the necessary and sufficient nature of the conditions required to construct a consensus embedding: that if a total of M base embeddings are calculated from a single $\mathbb{R}^N$, some minimum number of strong embeddings ($K \leq M$) must be considered to construct a $\widetilde{\mathbb{R}}^n$ that is a strong embedding. Further, a $\widetilde{\mathbb{R}}^n$ so constructed will have an embedding strength $\psi(\widetilde{\mathbb{R}}^n)$ that will increase significantly as we include more strong embeddings in its computation.

3.1.4 Properties of consensus embedding

The following proposition will demonstrate that $\widetilde{\mathbb{R}}^n$ will have a lower inherent error in its pairwise relationships compared to the strong embeddings $\mathbb{R}^n_k, k \in \{1, \ldots, K\}$, used in its construction.

We first define the mean squared error (MSE) in the pairwise relationship between every pair of objects $c, d \in C$ in any embedding $\mathbb{R}^n$ with respect to the true pairwise relationships in $\widehat{\mathbb{R}}^n$ as,

$$\epsilon_X = \mathrm{E}_{cd}(\hat{\delta}^{cd} - \delta^{cd})^2. \tag{3.6}$$

where E_{cd} is the expectation of the squared error in the pairwise relationships in $\mathbb{R}^n$ calculated over all pairs of objects $c, d \in C$. We can hence calculate the expected MSE over all K base embeddings specified above as,

$$\epsilon_{K,X} = \mathrm{E}_K\left[\epsilon_X\right] = \mathrm{E}_K\left[\mathrm{E}_{cd}(\hat{\delta}^{cd} - \delta_k^{cd})^2\right]. \tag{3.7}$$

Given K observations δ_k^{cd}, $k \in \{1, \ldots, K\}$ (derived from selected base embeddings $\mathbb{R}_k^n$), we define the pairwise relationship in the consensus embedding $\widetilde{\mathbb{R}}^n$ as $\tilde{\delta}^{cd} = \mathrm{E}_K(\delta_k^{cd})$, where E_K is the expectation of δ_k^{cd} over K observations. The MSE in $\tilde{\delta}^{cd}$ with respect to the true pairwise relationships in $\widehat{\mathbb{R}}^n$ may be defined as (similar to Equation 3.6),

$$\epsilon_{\widetilde{X}} = \mathrm{E}_{cd}(\hat{\delta}^{cd} - \tilde{\delta}^{cd})^2, \tag{3.8}$$

where E_{cd} is the expectation of the squared error in the pairwise relationships in $\widetilde{\mathbb{R}}^n$ calculated over over all pairs of objects $c, d \in C$. It is clear that if for all $c, d \in C$ that $\tilde{\delta}^{cd} = \hat{\delta}^{cd}$, then $\widetilde{\mathbb{R}}^n$ is also a true embedding.

Proposition 2 *Given K independent, strong embeddings, $\mathbb{R}_k^n, k \in \{1, \ldots, K\}$, which are used to construct $\widetilde{\mathbb{R}}^n$, $\epsilon_{K,X} \geq \epsilon_{\widetilde{X}}$.*

Proof.

Expanding Equation 3.7,

$$\epsilon_{K,X} = \mathrm{E}_{cd}\left(\hat{\delta}^{cd}\right)^2 - 2\,\mathrm{E}_{cd}(\hat{\delta}^{cd})\,\mathrm{E}_K(\delta_k^{cd}) + \mathrm{E}_{cd}\,\mathrm{E}_K\left(\delta_k^{cd}\right)^2$$

$$\text{Now, } \mathrm{E}_K\left(\delta_k^{cd}\right)^2 \geq \left(\mathrm{E}_K\,\delta_k^{cd}\right)^2,$$

$$\geq \mathrm{E}_{cd}\left(\hat{\delta}^{cd}\right)^2 - 2\,\mathrm{E}_{cd}(\hat{\delta}^{cd})(\tilde{\delta}^{cd}) + \mathrm{E}_{cd}\left(\tilde{\delta}^{cd}\right)^2$$

$$\geq \epsilon_{\widetilde{X}}$$

□

Proposition 2 implies that $\widetilde{\mathbb{R}}^n$ will never have a higher error than the maximum error associated with any individual strong embedding $\mathbb{R}_k^n, k \in \{1, \ldots, K\}$, involved in its construction. However if $\epsilon_{K,X}$ is low, $\epsilon_{\widetilde{X}}$ may not significantly improve on it. Similar to Bagging [38] where correlated errors across weak classifiers are preserved in the ensemble result, if the pairwise relationship δ_k^{cd} is incorrect across all K embeddings, $\tilde{\delta}^{cd}$ will be incorrect as well. However Proposition 2 guarantees that $\epsilon_{\widetilde{X}}$ will never be worse than $\epsilon_{K,X}$.

3.2 Algorithms and implementation

Based on Proposition 1, 3 distinct steps are typically required for calculating a consensus embedding. First, we must generate a number of base embeddings (M), the steps for which are described in *CreateEmbed*. We then select for strong embeddings from amongst M base embeddings generated, described in *SelEmbed*. We will also discuss criteria for selecting strong embeddings. Finally, selected embeddings are combined to result in the final consensus embedding representation as explained in *CalcConsEmbed*. We also discuss some of the computational considerations of our implementation.

3.2.1 Creating n-dimensional data embeddings

One of the requirements for consensus embedding is the calculation of multiple uncorrelated, independent embeddings $\mathbb{R}^n$ from a single $\mathbb{R}^N$. This is also true of ensemble classification systems such as Boosting [99] and Bagging [38] which require multiple uncorrelated, independent classifications of the data to be generated prior to combination. As discussed previously, the terms "uncorrelated, independent" are used by us with reference to the method of constructing embeddings, as borrowed from ensemble classification literature [40]. Similar to random forests [121], we make use of a *feature space perturbation* technique to generate uncorrelated (base) embeddings. This is implemented by first creating M bootstrapped feature subsets of V features each (every subset $\eta_m, m \in \{1, \ldots, M\}$ containing $\binom{N}{V}$ features, no DR involved). Note, that the number of samples in each V-dimensional subset is the same as in the original N-dimensional space. Each V-dimensional η_m is then embedded in n-D space via DR (i.e. projecting from $\mathbb{R}^V$ to $\mathbb{R}^n$). M is chosen such that each of N dimensions appears in at least one η_m.

Algorithm *CreateEmbed*

Input: $\mathbf{F}(c) \in \mathbb{R}^N$ for all objects $c \in C$, n

Output: $\mathbf{X}_m(c) \in \mathbb{R}^n_m, m \in \{1, \ldots, M\}$

Data Structures: Feature subsets η_m, total number of subsets M, number of features in each subset V, DR method Φ

```
begin
    0. for $m = 1$ to $M$ do
    1.     Select $V < N$ features from $\mathbb{R}^N$, forming subset $\eta_m$;
    2.     Calculate $\mathbf{X}_m(c) \in \mathbb{R}^n_m$, for all $c \in C$ using $\eta_m$ and method $\Phi$;
    3. endfor
end
```

As discussed in the introduction, multiple methods exist to generate base embeddings, such as varying a parameter associated with a method (e.g. neighborhood parameter in LLE, as shown in [120]) as well as the method explored in this thesis (feature space perturbation). These methods are analogous to methods in the literature for generating base classifiers in a classifier ensemble [40], such as varying k in kNN classifiers (changing associated parameter) [123], or varying the training set for decision trees (perturbing the feature space) [121].

3.2.2 Selection of strong embeddings

Having generated M base embeddings, we first calculate their embedding strengths $\psi^{ES}(\mathbb{R}^n_m)$ for all $\mathbb{R}^n_m, m \in \{1, \ldots, M\}$. The calculation of ψ^{ES} can be done via performance evaluation measures such as those described below, based on the application and prior domain knowledge. Embeddings for which $\psi^{ES}(\mathbb{R}^n_m) > \theta$ are then selected as strong embeddings, where θ is a pre-specified threshold.

Algorithm *SelEmbed*

Input: $\mathbf{X}_m(c) \in \mathbb{R}^n_m$ for all objects $c \in C$, $m \in \{1, \ldots, M\}$

Output: $\mathbf{X}_k(c) \in \mathbb{R}^n_k, k \in \{1, \ldots, K\}$

Data Structures: A list Q, embedding strength function ψ^{ES},

```
embedding strength threshold $\theta$
begin
    for $m = 1$ to $M$ do
        Calculate $\psi^{ES}(\mathbb{R}^n_m)$;
        if $\psi^{ES}(\mathbb{R}^n_m) > \theta$
            Put $m$ in $Q$;
        endif
    endfor
    For each element $k$ of $Q$, store $\mathbf{X}_k(c) \in \mathbb{R}^n_k$ for all objects $c \in C$;
end
```

Note that while θ may be considered to be a parameter which needs to be specified to construct the consensus embedding, we have found in our experiments that the results are relatively robust to variations in θ. In general, θ may be defined based on the manner of evaluating the embedding strength, as discussed in the next section.

3.2.3 Evaluation of embedding strength

We present two performance measures in order to evaluate embedding strength: one measure being supervised and relying on label information; the other being unsupervised and driven by the separability of distinct clusters in the reduced dimensional embedding space. In Section 3.4.3 we compare the two performance measures against each other to determine their relative effectiveness in constructing a strong consensus embedding.

Supervised evaluation of embedding strength: We have demonstrated that embedding strength increases as a function of classification accuracy (Theorem 1, Appendix B), implying that strong embeddings will have high classification accuracies. Intuitively, this can be explained as strong embeddings showing greater class separation compared to weak embeddings. Given a binary labeled set of samples C, we denote the sets of objects corresponding to the two classes as S^+ and S^-, such that $C = S^+ \cup S^-$ and $S^+ \cap S^- = \emptyset$. When using a classification algorithm that does not consider class labels, we can evaluate classification accuracy as follows:

1. Apply classification algorithm to C (embedded in $\mathbb{R}^n$) to find T clusters (unordered, labeled set of objects), denoted via $\widehat{\Psi}_t, t \in \{1, \ldots, T\}$.
2. For each $\widehat{\Psi}_t$
   (a) Calculate $DTP = |\widehat{\Psi}_t \cap S^+|$.
   (b) Calculate $DTN = |(C - \widehat{\Psi}_t) \cap S^-|$.
   (c) Calculate classification accuracy for $\widehat{\Psi}_t$, as $\phi^{Acc}(\widehat{\Psi}_t) = \frac{DTP+DTN}{|S^+ \cup S^-|}$.
3. Calculate classification accuracy of $\mathbb{R}^n$ as $\phi^{Acc}(\mathbb{R}^n) = \max_T \left[\phi^{Acc}(\widehat{\Psi}_t)\right]$.

As classification has been done without considering label information, we must evaluate which of the clusters so obtained shows the greatest overlap with S^+ (the class of interest). We therefore consider the classification accuracy of the cluster showing the most overlap with S^+ as an approximation of the embedding strength of $\mathbb{R}^n$, i.e. $\psi^{ES}(\mathbb{R}^n) \approx \phi^{Acc}(\mathbb{R}^n)$.

Unsupervised evaluation of embedding strength: We utilize a measure known as the R-squared index (RSI), based off cluster validity measures [124], which can be calculated as follows:

1. Apply classification algorithm to C (embedded in $\mathbb{R}^n$) to find T clusters (unordered, labeled set of objects), denoted via $\widehat{\Psi}_t, t \in \{1, \ldots, T\}$.
2. Calculate $SST = \sum_{j=1}^{n} \left[\sum_{i=1}^{R} \left(\mathbf{X}(c_i) - \overline{\mathbf{X}(c_j)}\right)^2\right]$ (where $\overline{\mathbf{X}(c_j)}$ is the mean of data values in the j^{th} dimension).
3. Calculate $SSB = \sum_{\substack{j=1\cdots n\\ t=1\cdots T}} \left[\sum_{i=1}^{|\widehat{\Psi}_t|} \left(\mathbf{X}(c_i) - \overline{\mathbf{X}(c_j)}\right)^2\right]$.
4. Calculate R-squared index of $\mathbb{R}^n$ as $\phi^{RS}(\mathbb{R}^n) = \frac{SST-SSB}{SST}$.

RSI may be considered both a measure of the degree of difference between clusters found in a dataset as well as measurement of the degree of homogeneity between them. The value of ϕ^{RS} ranges between 0 and 1, where if $\phi^{RS} = 0$, no difference exists among clusters. Conversely, a value close to $\phi^{RS} = 1$ suggests well-defined, separable clusters in the embedding space. Note that when using RSI to evaluate embedding strength, it will be difficult to ensure that all selected embeddings are strong without utilizing *a priori* information. In such a case we can attempt to ensure that a significant majority of the embeddings selected are strong, which will also ensure that the consensus embedding $\widetilde{\mathbb{R}}^n$ is strong (based off Proposition 1).

3.2.4 Constructing the consensus embedding

Given K selected embeddings $\mathbb{R}^n_k, k \in \{1, \ldots, K\}$, we quantify pairwise relationships between all the objects in each $\mathbb{R}^n_k$ via Euclidean pairwise distances. Euclidean distances were chosen for our implementation as they are well understood, satisfy the metric assumption of the pairwise relationship, as well as being directly usable within the other methods used in this work. Ω_k denotes the ML estimator used for calculating $\tilde{\delta}^{cd}$ from K observations δ^{cd}_k for all objects $c, d \in C$.

```
Algorithm $CalcConsEmbed$
**Input**: $\mathbf{X}_k(c) \in \mathbb{R}^n_k$ for all objects $c \in C$, $k \in \{1, \ldots, K\}$
**Output**: $\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n$
**Data Structures**: Confusion matrix $W$, ML estimator $\Omega$, projection method $\gamma$
begin
    0. for $k = 1$ to $K$ do
    1.     Calculate $W_k(i,j) = \|\mathbf{X}_k(c) - \mathbf{X}_k(d)\|_2$ for all objects $c, d \in C$ with indices $i, j$;
    2. endfor
    3. Apply normalization to all $W_k, k \in \{1, \ldots, K\}$;
    4. Obtain $\widetilde{W}(i,j) = \Omega_k \left[ W_k(i,j) \right] \forall c, d \in C$;
    5. Apply projection method $\gamma$ to $\widetilde{W}$ to obtain final consensus embedding $\widetilde{\mathbb{R}}^n$;
end
```

Corresponding entries across all W_k (after any necessary normalization) are used to estimate $\tilde{\delta}^{cd}$ (and stored in $\widetilde{W}$). In our implementation, we have used the median as the ML estimator as (1) the median is less corruptible to outliers, and (2) the median and the expectation are interchangeable if one assumes a normal distribution [125]. In Section 3.4.2 we compare classification results using both the mean and median individually as the ML estimator. We apply a projection method γ, such as multi-dimensional scaling (MDS) [102], to the resulting $\widetilde{W}$ to embed the objects in $\widetilde{\mathbb{R}}^n$ while preserving the pairwise distances between all objects $c \in C$. The underlying intuition for this final step is based on a similar approach adopted in [8] where MDS was applied to the co-association matrix (obtained by accumulating multiple weak clusterings of the data) in order to visualize the clustering results. As $\widetilde{W}$ is analogous to the co-association matrix, the projection method γ will allow us to construct the consensus embedding space $\widetilde{\mathbb{R}}^n$.

One can hypothesize that $\widetilde{W}$ is an approximation of distances calculated in the original feature space. Distances in the original feature space can be denoted as $\widehat{W}(i,j) = \|\mathbf{F}(c) - \mathbf{F}(d)\|_2 \, \forall c, d \in C$ with indices i, j. An alternative approach could therefore be to calculate $\widehat{W}$ in the original feature space and apply γ to it instead. However, noise artifacts in the original feature space may prevent it from being truly optimal for analysis [65]. As we will demonstrate in Section 3.4.1, simple DR, as well as consensus DR, provide superior representations of the data (by accounting for noise artifacts) as compared to using the original feature space directly.

3.2.5 Computational efficiency of consensus embedding

The most computationally expensive operations in consensus embedding are (1) calculation of multiple uncorrelated embeddings (solved as an eigenvalue problem in $O(n^3)$ time for n objects), and (2) computation of pairwise distances between all the objects in each strong embedding space (computed in time $O(n^2)$ for n objects). A slight reduction in both time and memory complexity can be achieved based on the fact that distance matrices will be symmetric (hence only the upper triangular need be calculated). Additionally, multiple embeddings and distance matrices can be computed via code parallelization. However these operations still scale polynomially based on the number of objects n.

To further reduce the computational burden we embed the consensus embedding paradigm within an intelligent sub-sampling framework. We make use of a fast implementation [126] of the popular mean shift algorithm [127] (MS) to iteratively represent data objects via their most representative cluster centers. As a result, the space retains its original dimensionality, but now comprises only some fractional number (n/t) of the original objects. These n/t objects are used in the calculations of consensus embedding as well as for any additional analysis. A mapping (Map) is retained from all n original objects to the final n/t representative objects. We can therefore map back results and analyses from the lowest resolution (n/t objects) to the highest resolution (n objects) easily. The fewer number of objects ($n/t << n$) ensures that consensus embedding is computationally feasible. In our implementation, t was determined automatically based on the number of stable cluster centers detected by MS.

```
Algorithm $ConsEmbedMS$
Input: $\mathbf{F}(c) \in \mathbb{R}^N$ for all objects $c \in C$, $n$
Output: $\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n$
Data Structures: Reduced set of objects $\bar{c} \in \bar{C}$
begin
    0. Apply MS [126] to $\mathbb{R}^N$ resulting in $\bar{\mathbb{R}}^N$ for sub-sampled set of objects $\bar{c} \in \bar{C}$;
    1. Save $Map$ from sub-sampled set of objects $\bar{c} \in \bar{C}$ to original set of objects $c \in C$ ;
    2. $\mathbf{X}_m(\bar{c}) = CreateEmbed(\mathbf{F}(\bar{c})|\eta_m, \Phi, M, V), \forall m \in \{1, \ldots, M\}$;
    3. $\mathbf{X}_k(\bar{c}) = SelEmbed(\mathbf{X}_m(\bar{c})|Q, \psi, \theta), \forall k \in \{1, \ldots, K\}, \forall m \in \{1, \ldots, M\}$;
    4. $\widetilde{\mathbf{X}}(\bar{c}) = CalcConsEmbed(\mathbf{X}_k(\bar{c})|W, \Omega, \gamma), \forall k \in \{1, \ldots, K\}$;
    5. Use MS and $Map$ to calculate $\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n$ from $\widetilde{\mathbf{X}}(\bar{c}) \in \bar{\mathbb{R}}^n$ for all objects $c \in C$;
end
```

For an MRI image comprising 5589 pixels (objects) for analysis, the individual algorithms $CreateEmbed$, $SelEmbed$ and $CalcConsEmbed$ took 121.33, 12.22, and 35.75 seconds respectively to complete (on average). By implementing our mean-shift optimization it took only 119 seconds (on average) for $ConsEmbedMS$ to complete analysis of an MRI image comprising between 15,000 and 40,000 pixels (objects); a calculation that would have been computationally intractable otherwise. All experiments were conducted using MATLAB 7.10 (Mathworks, Inc.) on a 72 GB RAM, 2 quad core 2.33 GHz 64-bit Intel Core 2 processor machine.

3.3 Experimental design for evaluating consensus embedding

3.3.1 Description of datasets used for evaluation

The different datasets used for empirical evaluation of consensus embedding included: (1) synthetic brain image data, and (2) gene-expression data (comprehensively summarized in Table 3.2). The overarching goal in each experiment described was to determine the degree of improvement in class-based separation via the consensus embedding representation as compared to alternative representations (quantified in terms of classification accuracy). Note that in the case of gene-expression data we have tested the robustness of the consensus embedding framework via the use of independent training and testing sets.

In the case of brain MR image data, we have derived texture features [92] on a per-pixel basis from each image. These features are based on calculating statistics

Datasets	Description	Features
Synthetic brain MRI images	10 slices (109×131 comprising 5589 pixels), 6 noise levels (0%, 1%, 3%, 5%, 7%, 9%) 3 RF inhomogeneity levels (0%, 20%, 40%)	Haralick (14)
Gene-expression:		
Prostate Tumor	102 training, 34 testing, 12,600 genes	300 most class-informative genes
Cancer Relapse	78 training, 19 testing, 24,481 genes	
Lymphoma	38 training, 34 testing, 7130 genes	
Lung Cancer	32 training, 149 testing, 12,533 genes	

Table 3.2: Image and gene-expression datasets used in our experiments.

from a gray level intensity co-occurrence matrix constructed from the image, and were chosen due to previously demonstrated discriminability between different types of brain matter [35] for MRI data. Following feature extraction, each pixel c in the MR image is associated with a N dimensional feature vector $\mathbf{F}(c) = [f_u(c)|u \in \{1, \ldots, N\}] \in \mathbb{R}^N$, where $f_u(c)$ is the response to a feature operator for pixel c. In the case of gene-expression data, every sample c is considered to be associated with a high-dimensional gene-expression vector, also denoted $\mathbf{F}(c) \in \mathbb{R}^N$.

DR methods utilized to reduce $\mathbb{R}^N$ to $\mathbb{R}^n$ were graph embedding (GE) [9] and PCA [58]. These methods were chosen in order to demonstrate instantiations of consensus embedding using representative linear and non-linear DR schemes. Additionally, these methods have been leveraged both for segmentation as well as classification of similar biomedical image and bioinformatics datasets in previous work [128, 129]. The dimensionality of the embedding space, n, is calculated as the intrinsic dimensionality of $\mathbb{R}^N$ via the method of [122]. To remain consistent with notation defined previously, the result of DR on $\mathbf{F}(c) \in \mathbb{R}^N$ is denoted $\mathbf{X}_\Phi(c) \in \mathbb{R}^n$, while the result of consensus DR will be denoted $\widetilde{\mathbf{X}}_\Phi(c) \in \widetilde{\mathbb{R}}^n$. The subscript Φ corresponds to the DR method used, $\Phi \in \{GE, PCA\}$. For ease of description, the corresponding classification results are denoted $\Psi(\mathbf{F}), \Psi(\mathbf{X}_\Phi), \Psi(\widetilde{\mathbf{X}}_\Phi)$, respectively.

3.3.2 Experiment 1: Synthetic MNI brain data

Synthetic brain data [10] was acquired from BrainWeb[1], consisting of simulated proton density (PD) MRI brain volumes at various noise and bias field inhomogeneity levels. Gaussian noise artifacts have been added to each pixel in the image, while inhomogeneity artifacts were added via pixel-wise multiplication of the image with an intensity non-uniformity field. Corresponding labels for each of the separate regions within the brain, including white matter (WM) and grey matter (GM), were also available. Images comprising WM and GM alone were obtained from 10 sample slices (ignoring other brain tissue classes). The objective was to successfully partition GM and WM regions

[1] http://www.bic.mni.mcgill.ca/brainweb/

on these images across all 18 combinations of noise and inhomogeneity, via pixel-level classification (an application similar to Figure 1.3). Classification is done for all pixels $c \in C$ based on each of,

(i) the high-dimensional feature space $\mathbf{F}(c) \in \mathbb{R}^N, N = 14$,

(ii) simple GE on $\mathbf{F}(c)$, denoted $\mathbf{X}_{GE}(c) \in \mathbb{R}^n, n = 3$,

(iii) multi-dimensional scaling (MDS) on distances calculated directly in $\mathbb{R}^N$, denoted as $\mathbf{X}_{MDS}(c) \in \mathbb{R}^n, n = 3$ (alternative to consensus embedding, explained in Section 3.2.4),

(iv) consensus embedding, denoted $\widetilde{\mathbf{X}}_{GE}(c) \in \widetilde{\mathbb{R}}^n, n = 3$.

The final slice classification results obtained for each of these spaces are denoted as $\Psi(\mathbf{F})$, $\Psi(\mathbf{X}_{GE})$, $\Psi(\mathbf{X}_{MDS})$, $\Psi(\widetilde{\mathbf{X}}_{GE})$, respectively.

3.3.3 Experiment 2: Comparison of ML estimators in consensus embedding

For the synthetic brain data [10], over all 18 combinations of noise and inhomogeneity and over all 10 images, we compare the use of mean and median as ML estimators in *CalcConsEmbed*. This is done by preserving outputs from *SelEmbed* and only changing the ML estimator in the *CalcConsEmbed*. We then compare classification accuracies for detection of white matter in each of the resulting consensus embedding representations, $\widetilde{\mathbf{X}}_{GE}^{Med}$ and $\widetilde{\mathbf{X}}_{GE}^{Mean}$ (superscript denotes choice of ML estimator).

3.3.4 Experiment 3: Gene-expression data

Four publicly available binary class gene-expression datasets were obtained[2] with corresponding class labels for each sample [56]; the purpose of the experiment being to differentiate the two classes in each dataset. This data comprises the gene-expression vectorial data profiles of normal and cancerous samples for each disease listed in Table

[2] These datasets were downloaded from the Biomedical Kent-Ridge Repositories at http://datam.i2r.a-star.edu.sg/datasets/krbd/

2, where the total number of samples range from 72 to 181 patients and the number of corresponding features range from 7130 to 24,481 genes or peptides. All 4 data sets comprise independent training (S^{tr}) and testing (S^{te}) subsets, and these were utilized within a supervised framework for constructing and evaluating the consensus embedding representation.

Prior to analysis, each dataset was first pruned to the 300 most class-informative features based on t-statistics as described in [130]. The supervised cross-validation methodology for constructing the consensus embedding using independent training and testing sets is as follows,

(a) First, $CreateEmbed$ is run concurrently on data in S^{tr} and S^{te}, such that the same subsets of features are utilized when generating base embeddings for each of S^{tr} and S^{te}.

(b) $SelEmbed$ is then executed on base embeddings generated from S^{tr} alone, thus selecting strong embeddings from amongst those generated. Strong embeddings were defined based on $\theta = 0.15 \times \max_M [\psi(\mathbb{R}^n_m)]$.

(c) Corresponding (selected) embeddings for data in S^{te} are then combined within $CalcConsEmbed$ to obtain the final consensus embedding vectors denoted as $\widetilde{\mathbf{X}}_\Phi(c) \in \widetilde{\mathbb{R}}^n, \Phi \in \{GE, PCA\}, n = 4$.

For this dataset, both supervised (via clustering classification accuracy, superscript S) and unsupervised (via RSI, superscript US) measures of embedding strength were evaluated in terms of the classification accuracy of the corresponding consensus embedding representations.

In lieu of comparative DR strategies, a semi-supervised variant of GE [131] (termed SSAGE) was implemented, which utilizes label information when constructing the embedding. Within this scheme, higher weights are given to within-class points and lower weights to points from different classes. When running SSAGE, both S^{tr} and S^{te} were combined into a single cohort of data, and labels corresponding to S^{tr} alone were revealed to the SSAGE algorithm.

An additional comparison was conducted against a supervised random forest-based

kNN classifier operating in the original feature space to determine whether DR provided any advantages in the context of high-dimensional biomedical data. This was implemented by training a kNN classifier on each of the feature subsets for S^{tr} (that were utilized in *CreateEmbed*), but without performing DR on the data. Each such kNN classifier was then used to classify corresponding data in S^{te}. The final classification result for each sample in S^{te} is based on ensemble averaging to calculate the probability of a sample belonging to the target class. Classifications compared in this experiment were $\Psi(\mathbf{F})$, $\Psi(\mathbf{X}_{SSGE})$, $\Psi(\widetilde{\mathbf{X}}^{S}_{GE})$, $\Psi(\widetilde{\mathbf{X}}^{S}_{PCA})$, $\Psi(\widetilde{\mathbf{X}}^{US}_{GE})$, $\Psi(\widetilde{\mathbf{X}}^{US}_{PCA})$, respectively.

3.3.5 Classification to evaluate consensus embedding

For image data, classification was done via replicated k-means clustering [8], while for gene-expression data, classification was done via hierarchical clustering [132]. The choice of clustering algorithm was made based on the type of data being considered in each of the different experiments, as well as previous work in the field. Note that both these clustering techniques do not consider class label information while classifying the data, and have been demonstrated as being deterministic in nature (hence ensuring reproducible results). The motivation in using such techniques for classification was to ensure that no classifier bias or fitting optimization was introduced during evaluation. As our experimental intent was purely to examine improvements in class separation offered by the different data representations, all improvements in corresponding classification accuracies may be directly attributed to improved class discriminability in the corresponding space being evaluated (without being dependent on optimizing the technique used for classification).

3.3.6 Evaluating and visualizing experimental results for consensus embedding

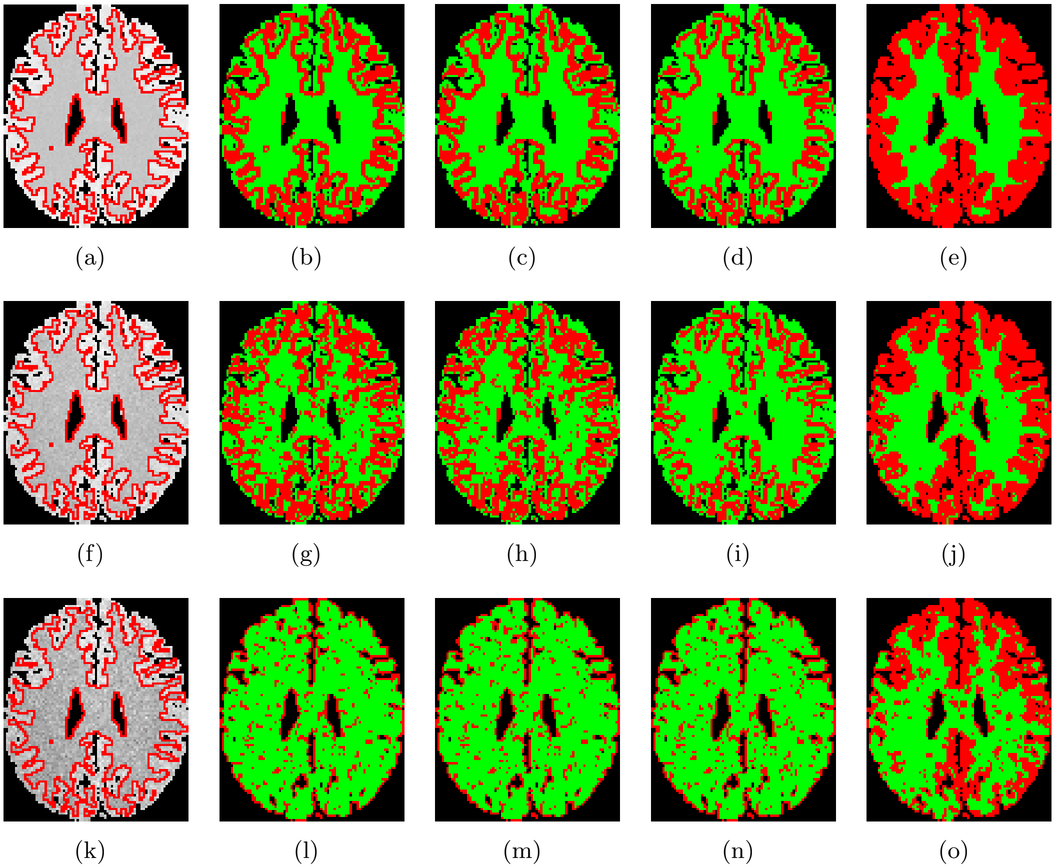
To visualize classification results as region partitions on the images, all the pixels were plotted back onto the image and assigned colors based on their classification label membership. Similar to the partitioning results shown in Figure 1.3, pixels of the same color were considered to form specific regions. For example, in Figure 1.3(h), pixels colored green were considered to form the foreground region, while pixels colored red were considered to form the background.

Classification accuracy of clustering results for images as well as gene-expression data can be quantitatively evaluated as described previously (Section 3.2.3). Image region partitioning results as well as corresponding classification accuracies of the different methods (GE, PCA, consensus embedding) were used to determine what improvements are offered by consensus embedding.

3.4 Results

3.4.1 Experiment 1: Synthetic MNI brain data

Figure 3.1 shows qualitative pixel-level WM detection results on MNI brain data for comparisons to be made across 3 different noise and inhomogeneity combinations (out of 18 possible combinations). The original PD MRI image for selected combinations of noise and inhomogeneity with the ground truth for WM superposed as a red contour is shown in Figures 3.1(a), (f), (k). Note that this is a 2 class problem, and GM (red) and WM (green) region partitions are visualized together in all the result images, as explained previously. Other brain tissue classes were ignored. Comparing the different methods used, when only noise (1%) is added to the data, all three of $\Psi(\mathbf{F})$ (Figure 3.1(b)), $\Psi(\mathbf{X}_{MDS})$ (Figure 3.1(c)), and $\Psi(\mathbf{X}_{GE})$ (Figure 3.1(d)) are only able to identify the outer boundary of the WM region. However, $\Psi(\widetilde{\mathbf{X}}_{GE})$ (Figure 3.1(e)) shows more accurate detail of the WM region in the image (compare with the ground truth WM region outlined in red in Figure 3.1(a)). When RF inhomogeneity (20%) is added to the data for intermediate levels of noise (3%), note the poor WM detection results for $\Psi(\mathbf{F})$ (Figure 3.1(g)), $\Psi(\mathbf{X}_{MDS})$ (Figure 3.1(h)), and $\Psi(\mathbf{X}_{GE})$ (Figure 3.1(i)). $\Psi(\widetilde{\mathbf{X}}_{GE})$ (Figure 3.1(j)), however, yields a more accurate WM detection result (compared to the ground truth WM region in Figure 3.1(f)). Increasing the levels of noise (7%) and inhomogeneity (40%) results in further degradation of WM detection performance for $\Psi(\mathbf{F})$ (Figure 3.1(l)), $\Psi(\mathbf{X}_{MDS})$ (Figure 3.1(m)), and $\Psi(\mathbf{X}_{GE})$ (Figure 3.1(n)). Note from Figure 3.1(o) that $\Psi(\widetilde{\mathbf{X}}_{GE})$ appears to fare far better than $\Psi(\mathbf{F})$, $\Psi(\mathbf{X}_{MDS})$, and


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

Figure 3.1: Pixel-level WM detection results visualized for one image from the MNI brain MRI dataset, each row corresponding to a different combination of noise and inhomogeneity: (a)-(e) 1% noise, 0% inhomogeneity, (f)-(j) 3% noise, 20% inhomogeneity, (k)-(o) 7% noise, 40% inhomogeneity. The first column shows the original PD MRI image with the ground truth for WM outlined in red, while the second, third, fourth, and fifth columns show the pixel-level WM classification results for $\Psi(\mathbf{F}), \Psi(\mathbf{X}_{MDS}), \Psi(\mathbf{X}_{GE})$, and $\Psi(\widetilde{\mathbf{X}}_{GE})$, respectively. The red and green colors in (b)-(e), (g)-(j), (l)-(o) denote the GM and WM regions identified in each result image.

$\Psi(\mathbf{X}_{GE})$.

For each of the 18 combinations of noise and inhomogeneity, we averaged the WM detection accuracies $\phi^{Acc}(\mathbf{F})$, $\phi^{Acc}(\mathbf{X}_{MDS})$, $\phi^{Acc}(\mathbf{X}_{GE})$, $\phi^{Acc}(\widetilde{\mathbf{X}}_{GE})$ (calculated as described in Section 3.2.3) over all 10 images considered (a total of 180 experiments). These results are summarized in Table 3.3 (corresponding trend visualization in Figure 3.2) with accompanying standard deviations in accuracy. Note that $\phi^{Acc}(\widetilde{\mathbf{X}}_{GE})$ shows a consistently better performance than the remaining methods ($\phi^{Acc}(\mathbf{F})$, $\phi^{Acc}(\mathbf{X}_{MDS})$, $\phi^{Acc}(\mathbf{X}_{GE})$) in 17 out of 18 combinations of noise and inhomogeneity. This trend is

Noise	Inhomogeneity	$\phi^{Acc}(\mathbf{F})$	$\phi^{Acc}(\mathbf{X}_{MDS})$	$\phi^{Acc}(\mathbf{X}_{GE})$	$\phi^{Acc}(\widetilde{\mathbf{X}}_{GE})$
0%	0%	65.55±1.84	65.55±1.84	65.55±1.84	**66.86±2.89**
	20%	55.75±1.65	55.75±1.65	55.75±1.65	**61.65±4.58**
	40%	70.03±2.79	**70.08±2.82**	51.84±0.99	64.28±5.93
1%	0%	59.78±1.31	59.74±1.29	74.71±9.06	**80.62±1.03**
	20%	59.36±1.30	59.32±1.33	60.95±8.67	**73.07±8.97**
	40%	59.20±1.12	59.12±1.15	56.38±1.53	**66.46±9.80**
3%	0%	53.35±1.31	53.39±1.27	59.94±7.00	**85.38±0.75**
	20%	55.01±2.92	54.91±3.11	63.88±10.85	**84.61±0.81**
	40%	57.63±1.78	57.71±1.67	57.33±1.38	**79.19±7.56**
5%	0%	62.90±0.72	62.84±0.66	66.67±10.22	**89.68±1.36**
	20%	61.49±1.38	61.49±1.42	82.61±7.39	**86.81±1.38**
	40%	61.02±0.99	61.03±1.09	74.91±9.09	**81.67±1.51**
7%	0%	64.28±0.71	64.26±0.76	66.95±6.25	**87.81±0.73**
	20%	64.07±1.03	64.01±0.96	74.22±10.59	**86.07±1.05**
	40%	64.05±1.19	64.04±1.14	64.44±1.25	**81.53±1.57**
9%	0%	64.96±0.90	64.94±0.88	66.36±1.66	**75.51±14.35**
	20%	64.85±0.97	64.79±0.95	65.68±1.32	**78.18±9.86**
	40%	64.65±0.83	64.63±0.84	65.30±0.74	**77.83±5.00**

Table 3.3: Pixel-level WM detection accuracy and standard error averaged over 10 MNI brain images and across 18 combinations of noise and inhomogeneity for each of: (1) $\Psi(\mathbf{F})$, (2) $\Psi(\mathbf{X}_{MDS})$, (3) $\Psi(\mathbf{X}_{GE})$, (4) $\Psi(\widetilde{\mathbf{X}}_{GE})$ (with median as MLE). Improvements in classification accuracy via $\Psi(\widetilde{\mathbf{X}}_{GE})$ were found to be statistically significant.

also visible in Figure 3.2.

For each combination of noise and inhomogeneity, a paired Students' t-test was conducted between $\phi^{Acc}(\widetilde{\mathbf{X}}_{GE})$ and each of $\phi^{Acc}(\mathbf{F})$, $\phi^{Acc}(\mathbf{X}_{MDS})$, and $\phi^{Acc}(\mathbf{X}_{GE})$, with the null hypothesis being that there was no improvement via $\Psi(\widetilde{\mathbf{X}}_{GE})$ over all 10 brain images considered. $\Psi(\widetilde{\mathbf{X}}_{GE})$ was found to perform significantly better ($p < 0.05$) than all of $\Psi(\mathbf{F})$, $\Psi(\mathbf{X}_{MDS})$, and $\Psi(\mathbf{X}_{GE})$ in 16 out of 18 combinations of noise and inhomogeneity.

Comparing $\phi^{Acc}(\mathbf{F})$, $\phi^{Acc}(\mathbf{X}_{MDS})$, and $\phi^{Acc}(\mathbf{X}_{GE})$, it can be observed that $\Psi(\mathbf{F})$ and $\Psi(\mathbf{X}_{MDS})$ perform similarly for all combinations of noise and inhomogeneity (note that the corresponding red and blue trend-lines completely overlap in Figure 3.2). In contrast, $\Psi(\mathbf{X}_{GE})$ shows improved performance at every combination of noise and inhomogeneity as compared to either of $\Psi(\mathbf{F})$ and $\Psi(\mathbf{X}_{MDS})$. $\Psi(\widetilde{\mathbf{X}}_{GE})$ was seen to significantly improve over all of $\Psi(\mathbf{F})$, $\Psi(\mathbf{X}_{MDS})$, and $\Psi(\mathbf{X}_{GE})$, reflecting the advantages

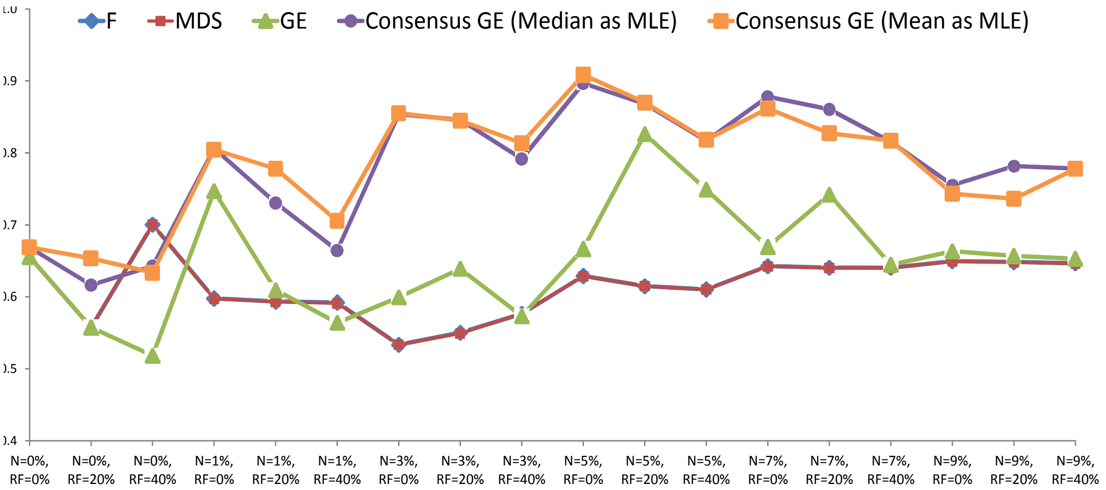


Figure 3.2: Visualization of classification accuracy trends (Tables 3 and 4). $\Psi(\widetilde{\mathbf{X}}_{GE})$ (consensus embedding) performs significantly better than comparative strategies (original feature space, GE, MDS); using median as ML estimator (purple) may be marginally more consistent than using mean as ML estimator (orange). $\Psi(\mathbf{F})$ (blue) and $\Psi(\mathbf{X}_{MDS})$ (red) perform similarly (corresponding trends directly superposed on one another).

of consensus embedding.

3.4.2 Experiment 2: Comparison of ML estimators

WM pixel-level detection accuracy results for consensus embedding using two different ML estimators (median and mean) were averaged over all 10 MNI brain images considered and summarized in Table 3.4, for each of the 18 combinations of noise and inhomogeneity (total of 180 experiments). We see that the accuracy values are generally consistent across all the experiments conducted. No statistically significant difference in classifier performance was observed when using $\Psi(\widetilde{\mathbf{X}}_{GE}^{Med})$ and $\Psi(\widetilde{\mathbf{X}}_{GE}^{Mean})$. It would appear that $\Psi(\widetilde{\mathbf{X}}_{GE}^{Med})$ is less susceptible to higher noise and bias field levels compared to $\Psi(\widetilde{\mathbf{X}}_{GE}^{Mean})$ (trends in Figure 3.2).

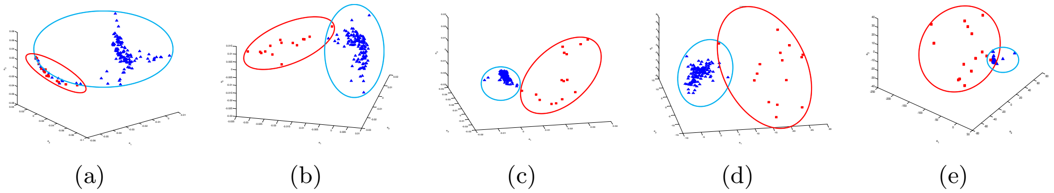
3.4.3 Experiment 3: Gene-expression Data

Table 3.6 summarizes classification accuracies for each of the strategies compared: supervised consensus-PCA and consensus-GE ($\Psi(\widetilde{\mathbf{X}}_{PCA}^{S})$, $\Psi(\widetilde{\mathbf{X}}_{GE}^{S})$, respectively), unsupervised consensus-PCA and consensus-GE ($\Psi(\widetilde{\mathbf{X}}_{PCA}^{US})$, $\Psi(\widetilde{\mathbf{X}}_{GE}^{US})$, respectively), SSAGE ($\Psi(\mathbf{X}_{SSGE})$), as well as supervised classification of the original feature space ($\Psi(\mathbf{F})$).

Noise	Inhomogeneity	$\phi^{Acc}(\widetilde{\mathbf{X}}_{GE}^{Med})$	$\phi^{Acc}(\widetilde{\mathbf{X}}_{GE}^{Mean})$
0%	0%	66.86±2.89	**66.89±2.91**
	20%	61.65±4.58	**65.34±4.12**
	40%	**64.28±5.93**	63.39±6.51
1%	0%	**80.62±1.03**	80.45±1.07
	20%	73.07±8.97	**77.81±0.96**
	40%	66.46±9.80	**70.56±7.15**
3%	0%	85.38±0.75	**85.53±0.84**
	20%	**84.61±0.81**	84.49±0.76
	40%	79.19±7.56	**81.37±1.39**
5%	0%	89.68±1.36	**90.85±1.32**
	20%	86.81±1.38	**87.01±1.83**
	40%	81.67±1.51	**81.82±1.32**
7%	0%	**87.81±0.73**	86.17±6.11
	20%	**86.07±1.05**	82.73±8.23
	40%	81.53±1.57	**81.72±1.47**
9%	0%	**75.51±14.35**	74.32±16.11
	20%	**78.18±9.86**	73.63±12.75
	40%	**78.18±9.86**	73.63±12.75

Table 3.4: Pixel-level WM detection accuracy and standard error averaged over 10 MNI brain images and for 18 combinations of noise and inhomogeneity (180 experiments) with each of the 2 ML estimators considered in *CalcConsEmbed*: (1) median $(\Psi(\widetilde{\mathbf{X}}_{GE}^{Med}))$, (2) mean $(\Psi(\widetilde{\mathbf{X}}_{GE}^{Mean}))$.

These results suggest that consensus embedding yields a superior classification accuracy compared to alternative strategies. We posit that this improved performance is due to the more accurate representation of the data obtained via consensus embedding.



(a) (b) (c) (d) (e)

Figure 3.3: 3D visualization of embedding results for lung cancer gene-expression data: (a) $\mathbf{X}_{SSGE}$, (b) $\widetilde{\mathbf{X}}_{GE}^{S}$, (c) $\widetilde{\mathbf{X}}_{GE}^{US}$, (d) $\widetilde{\mathbf{X}}_{PCA}^{S}$, (e) $\widetilde{\mathbf{X}}_{PCA}^{US}$. The 3 axes correspond to the primary 3 eigenvalues obtained via different DR methods (SSAGE, consensus-GE and consensus-PCA), while the colors of the objects (red and blue) are based on known class information (cancer and non-cancer, respectively). Note the relatively poor performance of (a) semi-supervised DR compared to (b)-(e) consensus DR. Both supervised ((b) and (d)) and unsupervised ((c) and (e)) consensus DR show relatively consistent separation between the classes with distinct, tight clusters. The best clustering accuracy for this dataset was achieved by (c) unsupervised consensus GE $(\widetilde{\mathbf{X}}_{GE}^{US})$.

	$\phi^{Acc}(\mathbf{F})$	$\phi^{Acc}(\mathbf{X}_{SSGE})$	$\phi^{Acc}(\widetilde{\mathbf{X}}^{S}_{PCA})$	$\phi^{Acc}(\widetilde{\mathbf{X}}^{US}_{PCA})$	$\phi^{Acc}(\widetilde{\mathbf{X}}^{S}_{GE})$	$\phi^{Acc}(\widetilde{\mathbf{X}}^{US}_{GE})$
1	73.53	73.53	97.06	**100**	**100**	76.47
2	**68.42**	63.16	63.16	57.89	63.16	57.89
3	89.93	10.07	99.33	96.64	98.66	**100**
4	58.82	61.76	**97.06**	76.47	**97.06**	67.65

Table 3.5: Classification accuracies for testing cohorts of 4 different binary class gene-expression datasets, comparing (1) supervised random forest classification of original feature space ($\mathbf{F}$), (2) unsupervised hierarchical clustering of semi-supervised DR space ($\mathbf{X}_{SSGE}$), and (3) unsupervised hierarchical clustering of consensus embedding space ($\widetilde{\mathbf{X}}_{GE}, \widetilde{\mathbf{X}}_{PCA}$).

The presence of a large noisy, high-dimensional space was seen to adversely affect supervised classification performance of $\mathbf{F}$, which yielded a worse classification accuracy than unsupervised classification (of consensus-GE and consensus-PCA) in 3 out of the 4 datasets. Moreover, semi-supervised DR, which utilized label information to construct $\mathbf{X}_{SSGE}$, was also seen to perform worse than consensus embedding (both supervised and unsupervised variants). We posit that this is because SSAGE does not explicitly account for noise, but only modifies the pairwise relationships between points based on label information (possibly exacerbating the effects of noise). By contrast, any label information used by consensus embedding is used to account for noisy samples when approximating the "true" pairwise relationships between points. The difference in the final embedding representations can be visualized in 3D in Figure 6, obtained by plotting

	$\phi^{Acc}(\widetilde{\mathbf{X}}^{S}_{PCA})$			$\phi^{Acc}(\widetilde{\mathbf{X}}^{US}_{PCA})$		
	$M = 200$	$M = 500$	$M = 1000$	$M = 200$	$M = 500$	$M = 1000$
1	97.06	97.06	97.06	100	100	100
2	57.89	63.16	57.89	57.89	57.89	52.63
3	99.33	99.33	99.33	96.64	95.97	96.64
4	94.12	97.06	97.06	76.47	67.65	61.76

Table 3.6: Classification accuracies for testing cohorts of 4 different binary class gene-expression datasets for $\widetilde{\mathbf{X}}^{S}_{PCA}$ and $\widetilde{\mathbf{X}}^{US}_{PCA}$, while varying the number of subsets M generated within $CreateEmbed$.

	$\phi^{Acc}(\widetilde{\mathbf{X}}^{S}_{GE})$			$\phi^{Acc}(\widetilde{\mathbf{X}}^{US}_{GE})$		
	$M = 200$	$M = 500$	$M = 1000$	$M = 200$	$M = 500$	$M = 1000$
1	100	100	97.06	76.47	76.47	76.47
2	57.89	57.89	57.89	57.89	57.89	57.89
3	98.66	98.66	97.99	100	100	90.60
4	61.76	97.06	55.88	67.65	67.65	67.65

Table 3.7: Classification accuracies for testing cohorts of 4 different binary class gene-expression datasets for $\widetilde{\mathbf{X}}^{S}_{GE}$ and $\widetilde{\mathbf{X}}^{US}_{GE}$, while varying the number of subsets M generated within $CreateEmbed$.

all the samples in the lung cancer gene-expression dataset in 3D Eigen space. Note that consensus DR (Figures 3.3(b)-(e)) shows significantly better separation between the classes with more distinct, tighter clusters as well as fewer false positives compared to SSAGE (Figure 3.3(a)).

Further, comparing the performance of supervised ($\Psi(\widetilde{\mathbf{X}}^{S}_{PCA})$, $\Psi(\widetilde{\mathbf{X}}^{S}_{GE})$) and unsupervised ($\Psi(\widetilde{\mathbf{X}}^{US}_{PCA})$, $\Psi(\widetilde{\mathbf{X}}^{US}_{GE})$) variants of consensus embedding demonstrates comparable performance between them, though a supervised measure of embedding strength shows a trend towards being more consistent. The relatively high performance of $\Psi(\widetilde{\mathbf{X}}^{US}_{PCA})$ and $\Psi(\widetilde{\mathbf{X}}^{US}_{GE})$ demonstrate the feasibility of a completely unsupervised framework for consensus embedding.

For both consensus PCA and consensus GE we tested the parameter sensitivity of our scheme by varying the number of feature subsets generated ($M \in \{200, 500, 1000\}$) in the $CreateEmbed$ algorithm (Tables 3.6 & 3.7). The relatively low variance in classification accuracy as a function of M reflects the invariance to parameters of consensus embedding. No consistent trend was seen in terms of either of consensus-GE or consensus-PCA outperforming the other.

3.5 Discussion

We have presented a novel dimensionality reduction scheme called consensus embedding which can be used in conjunction with a variety of DR methods for a wide range of high-dimensional biomedical data classification and segmentation problems. Consensus embedding exploits the variance within multiple base embeddings and combines them to produce a single stable solution that is superior to any of the individual embeddings, from a classification perspective. Specifically, consensus embedding is able to preserve pairwise object-class relationships from the high- to the low-dimensional space more accurately compared to any single embedding technique. Using an intelligent subsampling approach (via mean-shift) and code parallelization, computational feasibility and practicability of our method is ensured.

Results of quantitative and qualitative evaluation in over 200 experiments on toy, synthetic, and clinical images in terms of detection and classification accuracy demonstrated that consensus embedding shows significant improvements compared to traditional DR methods such as PCA. We also compared consensus embedding to using the feature space directly, as well as to using an embedding based on distance preservation directly from the feature space (via MDS [102]), and found significant performance improvements when using consensus embedding. Even though the features and classifier used in these experiments were not optimized for image segmentation purposes, consensus embedding outperforms state-of-the-art segmentation schemes (graph embedding, also known as normalized cuts [9]), differences being statistically significant in all cases. Incorporating spatial constraints via algorithms such as Markov Random Fields [36] could be used to further bolster the image segmentation results via consensus embedding.

In experiments for high-dimensional biomedical data analysis using gene-expression signatures, consensus embedding also demonstrated improved results compared to semi-supervised DR methods (SSAGE [131]). Evaluating these results further illustrates properties of consensus embedding: (1) the consensus of multiple projections improves upon any single projection (via either linear PCA or non-linear GE), (2) the error rate for consensus embedding is not significantly affected by parameters associated with the method, as compared to traditional DR. Finally, the lower performance of a supervised classifier using the original noisy feature space as compared to using the consensus embedding representation demonstrates the utility of DR to obtain improved representations of the data for classification.

It is however worth noting that in certain scenarios, consensus embedding may not yield optimal results. For instance, if very few embeddings are selected for consensus, the improvement in performance via consensus embedding over simple DR techniques may not be as significant. This translates to having a sparsely populated distribution for the estimation of the consensus pairwise relationship, resulting in a lower confidence being associated with it. Such a scenario may arise due to incorrectly specified selection criteria for embeddings; however, it is relatively simple to implement self-tuning for the selection parameter (θ). Note we have reported results for a fixed value of θ in all our experiments/applications, further demonstrating the robustness of our methodology to choice of parameters.

In this work, consensus embedding has primarily been presented within a supervised framework (using class label information to evaluate embedding strength). Preliminary results in developing an unsupervised evaluation measure using the R-squared cluster validity index [124] are extremely promising. However, additional tuning and testing of the measure is required to ensure robustness.

Another area of future work is developing algorithms for the generating uncorrelated, independent embeddings. This is of great importance as generating truly uncorrelated, independent embeddings will allow us to capture the information from the data better, hence ensuring in an improved consensus embedding result. As mentioned previously, methods to achieve this could include varying the parameter associated with the DR method (e.g. neighborhood parameter in LLE [60]) as well as the feature space perturbation method explored in this thesis. These approaches are analogous to methods of generating weak classifiers within a classifier ensemble [40], such as varying the k parameter in kNN classifiers [123] or varying the training set for decision trees [121]. Note our feature space perturbation method to generate multiple, uncorrelated independent embeddings is closely related to the method used in random forests [121] to generate multiple weak, uncorrelated classifiers. Thus the embeddings we generate, as with the multiple classifiers generated in ensemble classifier schemes, are not intended to be independent in terms of information content, but rather in their method of construction.

The overarching goal of consensus embedding is to optimally preserve pairwise relationships when projecting from high- to low-dimensional space. In this work, pairwise relationships were quantified by us using the popular Euclidean distance metric. This was chosen as it is well understood in the context of these methods used within our algorithm (e.g. the use of MDS). Alternative pairwise relationship measures could include the geodesic distance [61] or the symmetrized Kullback-Leibler divergence [133]. It is important to note that such measures will need to satisfy the properties of a metric to ensure that they correctly quantify both triangle as well as pairwise relationships. We currently use MDS [102] to calculate the final consensus embedding (based on the consensus pairwise distance matrix). We have chosen to use MDS due to ease of computational complexity, but this method could be replaced by a non-linear variant instead. Finally, our intelligent sub-sampling approach to ensure computational feasibility comes with a caveat of the out-of-sample extension problem [134]. We currently handle this using a mapping of results from high to low-resolutions, but are currently identifying more sophisticated solutions. We intend to study these areas in greater detail to further validate the generalizability of consensus embedding.

Chapter 4
Experimental Design

4.1 Data Description

A total of 24 patient datasets were considered in this work. These datasets were obtained from a prospective study approved by the Institutional Review Board at Beth Israel Deaconess Medical Center, Boston, MA. This data was acquired from patients who had first been confirmed to have CaP via positive core needle biopsies and were scheduled for a radical prostatectomy. Prior to surgery, the patients were imaged using a combined torso-phased array and endorectal coil (MedRad, Pittsburgh, PA) using a 3 Tesla whole-body MRI scanner (Genesis Signa LX Excite; GE Medical Systems, Milwaukee, Wisconsin), in the axial plane and included T2w, DCE, and DWI protocols. The parameters for axial T2w MR imaging were TR/TE = 6375/165 msec with a slice thickness of 1.5-2mm (no gap between the slices). The DCE-MR images were acquired during and after a bolus injection of 0.1 mmol/kg of body weight of gadopentetate dimeglumine using a 3-dimensional gradient echo sequence with a temporal resolution of 1 min 35 sec. Two pre-contrast and 5 post-contrast sequential acquisitions were obtained. DWI imaging had B-values of 0 and 1000, with the number of directions imaged being 25, based on which an ADC map was calculated. Matrix size of acquisition was 320×224–192 voxels with a field of view of 12 x 12 cm.

Following radical prostatectomy and prior to sectioning, the excised prostate was embedded in a paraffin block while maintaining the orientation to keep the urethra perpendicular to the plane of slicing. This procedure facilitated the identification of a corresponding *in vivo* 2D axial T2w MRI slice for each 2D histology slice, for the purposes of registration. No *ex vivo* MRI or gross pathology photographs were acquired. Preparation of the digitized WMH sections proceeded as follows: (1) the excised prostate was cut into sections that are 3-4 mm thick by slicing axial sections from the paraffin block using a circular blade, (2) a microtome was used to further cut the sections into thin slices that are about 5 μm thick, and (3) a single thin slice from each 3-4 mm thick section was chosen and stained with Haematoxylin and Eosin (H & E). The stained section was then examined under a light microscope using up to 40x apparent magnification to identify and delineate the regions of CaP. Further details of MRI and histological acquisition have been described previously [11, 27].

The 24 patient studies considered in this work were chosen from a larger cohort of 124 cases, all of which had MR imaging performed prior to a radical prostatectomy procedure. Of the 124 cases, only 65 studies included usable T2w MRI with corresponding digitized whole mount histological sections. Of these, 25 cases were identified with histological sections on which CaP was visible and could be annotated by a pathologist. One additional case had to be discarded due to the poor quality of the MR imaging. A pathologist and radiologist working in unison visually identified corresponding 2D whole-mount histological sections (WMHS) and axial T2w MRI slices from these 24 studies. These correspondences were established via anatomical fiducials such as the urethra, veromontanum, as well as prominent BPH nodules that were visually discernible on both the histology and pre-operative MRI.

4.1.1 Data utilized for identifying features and classifier for automated CaP detection

The first set of objectives of this thesis involve (1) identifying significant features, and (2) an optimal classifier for CaP detection on T2w MRI, with further stratification for CaP detection on zonal (CG, PZ) basis. Based on the recommendations by McNeal [20], a patient study was classified as having CG or PZ CaP if more than 70% of prostate cancer volume was found to be present in a particular zone. Of the 24 datasets, 16 were thus determined as having PZ CaP alone (50 2D sections), while the remaining 6 were identified as having CaP in the CG alone (30 2D sections). The remaining 2 studies (14

2D sections) were found to exhibit nodules of CaP in both the CG and the PZ. In order to ensure that the sets of CG and PZ CaP were as distinct from each other as possible, only those sections were included which showed an explicit focus of CaP in either the CG or the PZ, i.e. for the purposes of the first 2 aims, only 22 patient studies (80 2D sections) were utilized.

4.1.2 Data utilized for CaP detection via multi-parametric MRI

The primary objective in this aim was to identify CaP presence and extent via constructing a unified fused representation of multi-parametric (T2w, DCE, DWI) MRI. Of the 24 datasets, 15 studies comprised the full complement of T2w, DCE, and DWI MR acquisition. The remaining 9 datasets comprised T2w and DCE acquisitions alone. We therefore demonstrate the utility of our consensus embedding framework for multi-parametric MR representation and subsequent classification (termed EMPrAvISE) on the 15 studies comprising the full complement of T2w, DCE, and DWI data. Additional stratification of the studies into CG and PZ CaP was not performed in this objective as this would have cause the results to be significantly underpowered.

4.2 Inter-protocol alignment of T2w, DCE, DWI MRI

T2w and ADC (from DWI) must be brought into spatial alignment with DCE MRI in order to facilitate analysis of all the data within the same frame of reference. This is

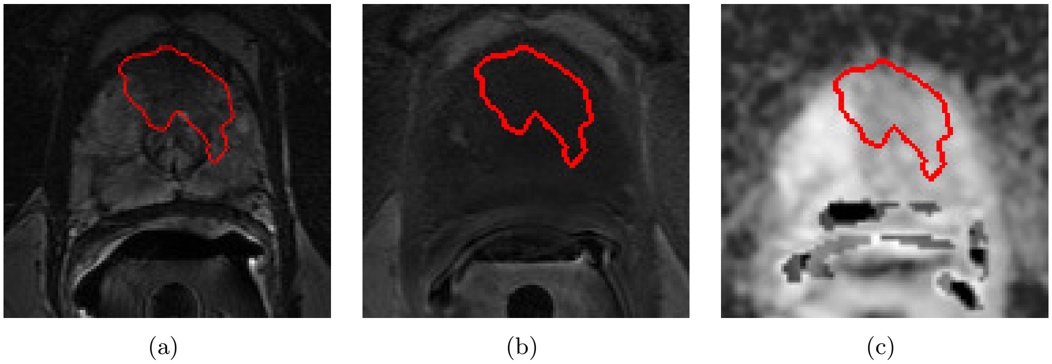


Figure 4.1: Corresponding co-registered multi-parametric MR images shown for (a) $\mathcal{C}^{T2}$, (b) $\mathcal{C}^{T1,5}$, and (c) $\mathcal{C}^{ADC}$, after inter-protocol alignment. The mapped CaP extent from WMH (not shown) is outlined in red on each figure (see Section 4.4 for details).

done via volumetric affine registration [11], hence correcting for inter-acquisition movement and resolution differences between the MRI protocols. Stored DICOM[1] image header information was used to determine relative voxel locations and sizes as well as slice correspondences between T2w, DCE, and ADC imagery. Figure 4.1 shows representative results of inter-protocol registration. Note the similarity in spatial alignment and resolution in Figures 4.1(a)-(c).

4.3 Pre-processing of MRI to account for intensity-based artifacts

For all datasets considered, the prostate gland was segmented out from the larger field of view of the axial endorectal T2w MRI image using an automated prostate capsule segmentation scheme [135]. Briefly, this scheme involves application of a statistical shape model to segment the capsule; the shape model being manually and interactively initialized in the vicinity of the capsule. All remaining analysis for CaP presence was thus localized to the prostate region-of-interest (ROI) alone. Figures 4.2(a) and (c) show the result of delineating the prostate ROI (in yellow) on 2 different T2w MRI sections.

The prostate ROI was then corrected for known acquisition-based intensity artifacts; bias field inhomogeneity [16] and intensity non-standardness [136]. The effects of bias field occur due to the usage of an endorectal probe [16], and manifest as a smooth variation of signal intensity across the T2w MR image. Bias field has been shown to significantly affect the automated classification of tissue regions [51], and was corrected for via the popular N3 algorithm [137]. Intensity non-standardness [136] refers to the issue of MR image "intensity drift" across different imaging acquisitions, resulting in MR image intensities lacking tissue-specific numeric meaning within the same MRI protocol, for the same body region, or for images of the same patient obtained on the same scanner [136]. This artifact was corrected for via an interactive implementation of the generalized scale algorithm [136], whereby the image intensity histograms across different patient MRI studies were non-linearly aligned.

[1] `http://medical.nema.org/`

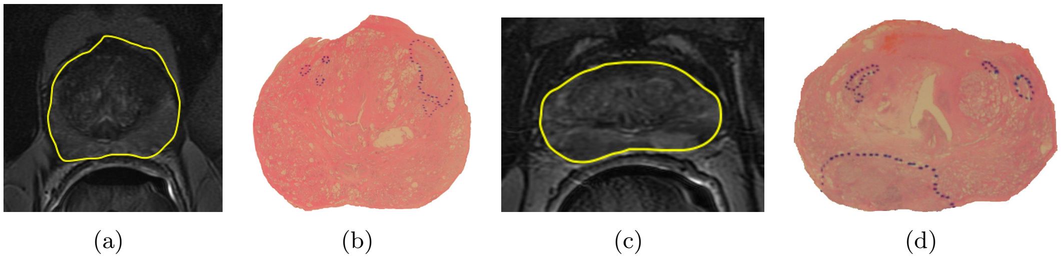


Figure 4.2: Representative MRI and histology images corresponding to (a)-(b) CG CaP and (c)-(d) PZ CaP. (a), (c) show the original T2w MRI images, with the prostate ROI outlined in yellow. (b), (d) show corresponding WMH sections, with CaP extent outlined in blue (by a pathologist).

4.4 Multi-modal registration of WMHS and MRI to obtain "ground truth" CaP extent

Registration of images from different modalities such as WMHS and MRI is complicated on account of the vastly different image characteristics of the individual modalities [11]. For example, the appearance of tissue and anatomical structures (e.g. hyperplasia, urethra, ducts) on MRI and histology are significantly different [138]. These differences are further exacerbated due to histological processing on WMHS (uneven tissue fixation, gland slicing and sectioning result in duct dilation and tissue loss) and the use of an endo-rectal coil on MRI (causing gland deformation). This may cause registration based on traditional intensity-based similarity measures, such as MI, to fail [11]. We have previously complemented intensity information with features derived by transformations of these intensities to drive multi-modal registration [50]. Additionally, achieving correct alignment of such imagery requires elastic transformations to overcome the non-linear shape differences.

In [11], Chappelow et al leveraged the availability of multiple imaging protocols (T2w, DCE, DWI) to introduce complementary sources of information for registration via a novel image similarity measure, Multi-Attribute Combined MI (MACMI) [11]. MACMI was found to be capable of simultaneously encoding the information from multiple protocols within a multivariate MI formulation. It therefore has the ability to handle images that significantly vary in terms of intensities and deformation characteristics, such as for *in vivo* MRI and *ex vivo* WMHS. Additionally, it involves a simple optimization procedure whereby a sequence of individual image transformations is determined.

The registration procedure [11] comprises the following 2 steps:

Step 1: WMHS and corresponding T2w MRI images are affinely aligned to enable correction of large translations, rotations, and differences in image scale,

Step 2: A non-linear alignment is performed via fully automated, non-linear hierarchical (multi-scale) B-spline registration driven by a higher-order variant of mutual information [11].

Figures 4.2(b) and (d) show WMH sections prior to registration, with CG and PZ CaP respectively (CaP extent outlined in blue on Figures 4.2(b) and (d)), which correspond to the T2w MRI sections shown in Figures 4.2(a), (c).

Note that while image registration was done in 2D, all subsequent feature extraction and classification operations were done in 3D. Consequently,

4.5 General overview of notation used in this thesis

The following is a generalized set of notation for the remainder of this thesis. All MRI data analyzed in this thesis is considered at the DCE-MRI resolution (256 × 256 voxels). The DCE MR image is denoted $\mathcal{C}^{T1,t} = (C, f^{T1,t})$, where $f^{T1,t}(c)$ assigns an intensity value to every voxel $c \in C$ at time point $t, t \in \{1, \ldots, 6\}$), Post inter-protocol registration, we obtain the T2w MR image $\mathcal{C}^{T2} = (C, f^{T2})$ and the corresponding ADC map $\mathcal{C}^{ADC} = (C, f^{ADC})$ in alignment with images in $\mathcal{C}^{T1,t}$. Therefore for every voxel $c \in C$, $f^{T2}(c)$ is the T2w MR image intensity value and $f^{ADC}(c)$ is the corresponding ADC value. We denote the transformed WMHS $\mathcal{C}^{H} = (C, f^{H})$, in alignment with $\mathcal{C}^{T1,t}$,$\mathcal{C}^{T2}$,$\mathcal{C}^{ADC}$. CaP extent on $\mathcal{C}^{H}$ is then mapped onto the DCE coordinate frame C, yielding the set of CaP voxels $G(C)$ (surrogate ground truth CaP extent). We thus assign a label to each voxel $c \in G(C), l(c) = 1$, with $l(c) = 0$ otherwise.

Feature vectors associated with every $c \in C$ are denoted $\mathbf{F}(c)$, comprising the feature responses at voxel $c \in C$ to N different feature operators (types of features extracted in this work are detailed in Chapter 5). The corresponding per-voxel classifier associated with every $c \in C$ (constructed based on $\mathbf{F}(c)$) is typically denoted $\mathbf{h}(c)$.

Individual chapters may additionally define notation to clarify details specific to that set of experiments. For example, in experiments concerning CaP quantification on MRI (Chapter 5) and determining an optimal classifier for CaP detection (Chapter 6), the T2w MR image intensity value is denoted $f(c)$ instead of $f^{T2}(c)$. Additional super- or subscripts to the symbols above may be used to differentiate between different feature sets or different types of classifiers considered in that chapter. Appendix A summarizes commonly used notation, symbols, and abbreviations appearing in the remainder of this thesis.

Chapter 5

Determining quantitative imaging signatures for central gland and peripheral zone prostate tumors on T2-weighted MRI

5.1 Specific notation for this chapter

In this chapter, the set of all voxels on T2w MRI is denoted as $\mathbf{C} = \{c_1, c_2, \ldots, c_n\}$; note that all analysis in this study has been performed at a per-voxel level as far as possible. After feature extraction (see below), a feature vector $F_i = \{f_i(c_1), \ldots, f_i(c_n)\}$, $i \in 1, \ldots, N$ is obtained as the collection of feature responses $f_i(c)$ for all $c \in \mathbf{C}$; hence F_i is a $1 \times n$ vector, where i represents the feature operator. The set of feature vectors corresponding to all operators i is given as $\mathbf{F} = \{F_1, \ldots, F_N\}$. Note that n is the total number of voxels considered, while $N = 110$ is the total number of features extracted. Additionally, every voxel $c \in \mathbf{C}$ is associated with a label $l(c) \in \{0, 1\}$, corresponding to cancer/benign annotations on T2w MRI (obtained via registration with histology, see below); the corresponding label vector is given by L (of size $1 \times n$). Additional notation and symbols are summarized in Appendix A.

5.2 Extracting CG and PZ CaP specific texture features from T2w MRI

It has previously been demonstrated that CaP appearance on T2w MRI may be better modeled by image texture features [51,52]; many of which have been frequently used in different image processing and computer vision tasks [93,139]. A total of 110 image features corresponding to 4 different types of texture were extracted, including Gabor [94] and Haar [98] wavelet features, as well as first and second order texture [92,93] features.

The goal of our study is to identify unique textural signatures for CaP presence in each of the CG and PZ regions, respectively. It may be expected that some combination of different types of textural features will thus play a role in accurately characterizing CaP appearance in these zones. We have therefore attempted to model CaP appearance via image texture features which have been frequently used in the image processing and computer vision fields [93,139], as well as shown to be useful for characterizing the appearance of different pathologies on MRI [140–142]. These textural descriptors include co-occurrence [92] (which capture spatial greylevel characteristics) and gradient-based features [93] (which capture edge and directional characteristics); these features have previously been shown to be useful in characterizing appearance of CaP on T2w MRI [51,86]. A number of different wavelet filter decomposition approaches [143] have also been employed by us as they allow for extraction of fine structural image detail at different orientations and scales, and could prove useful in quantitatively characterizing the micro- and macroscopic visual cues used by radiologists when identifying CaP regions on MRI. Most popular amongst these are the Gabor (continuous) [94] and the Haar (discrete) [98] wavelet transforms. Table 5.1 summarizes the texture features used in this study as well as the visual significance of such features for identifying CaP on prostate T2w MRI.

In all, 110 texture features corresponding to 4 different texture feature classes were extracted on a per-voxel basis from each MRI dataset. All feature extraction methods were implemented using MATLAB®(The Mathworks Inc, MA).

After feature extraction, every voxel $c \in C$ is associated with a 110-dimensional feature vector denoted $\widehat{\mathbf{F}}(c) = \{f_1(c), f_2(c), \ldots, f_{110}(c)\}$, for every $c \in C$.

5.3 Feature selection to construct zone-specific texture feature sets

After extracting texture features, we utilized the minimum Redundancy Maximum Relevance (mRMR) feature selection scheme [144] in order to identify an ensemble of features that will allow for optimal identification of CaP presence on MRI. During feature selection, 2 separate sets of data were considered; one comprising feature and label data

Feature	Implementation	Purpose	Significance for quantifying CaP appearance
Gabor wavelet transform (48)	Modulation of a complex sinusoid by a Gaussian function	Attempt to match localized frequency characteristics at multiple scales and orientations [94]	Quantify visual processing features used by radiologists when examining appearance of the carcinoma
Haar wavelet transform (12)	Decomposition coefficients via wavelet decomposition at multiple scales	Attempt decomposition of a signal in the discrete space while offering localization in the time and frequency domains [98]	Differentiate the amorphous nature of the non-CaP regions within foci of low SI
Haralick texture feature (36)	Construct joint probability distribution of the occurrence of greylevel intensities in an image (spatial relationship between pixels used to restrict counting of greylevel co-occurrences). Statistical features are then calculated from this distribution	Differentiate between different types of texture excellently due to calculation of 2nd order statistics (which quantify perceptual appearance of image) [92]	Useful in differentiating homogeneous low SI regions (CaP) from more hyper-intense appearance of normal prostate
Greylevel statistical features (14)	Mean, standard deviation as well as derivative features such as via convolution with the Sobel and Kirsch operators are calculated	Provide 1st order information, quantifying macroscopic appearance of image e.g. variation of intensities within image [93] etc.	May help localize regions of significant differences on T2w MR image, accurately detect region boundaries

Table 5.1: Summary of texture features used in this study as well as their significance for localization of CaP on T2w MRI (numbers in brackets signify how many features of each texture category were computed)

corresponding to voxels from the 16 datasets with PZ CaP, and the other comprising feature and label data corresponding to voxels from the 6 datasets with CG CaP. Thus the result of mRMR feature selection was 2 distinct QISes (one corresponding to CG CaP denoted $\mathbf{Q}^{CG}$, the other to PZ CaP denoted $\mathbf{Q}^{PZ}$); each set comprising texture features which can be considered highly discriminatory for differentiating between CaP and benign regions in the corresponding prostatic zone.

In the following description, the selected subset of features $\mathbf{Q}$ is comprised of feature vectors $F_i, i \in \{1, \ldots, |\mathbf{Q}|\}$ (note that $\mathbf{Q} \subset \mathbf{F}$ and $|Q| < N$). The mRMR scheme attempts to simultaneously optimize two distinct criteria. The first is "maximum relevance" which selects features F_i that have the maximal mutual information (MI) with respect to the corresponding label vector L. This is expressed as

$$U = \frac{1}{|\mathbf{Q}|} \sum_{F_i \in \mathbf{Q}} MI(F_i, L). \tag{5.1}$$

The second is "minimum redundancy" which ensures that selected features $F_i, F_j \in \mathbf{Q}, i, j \in \{1, \ldots, |\mathbf{Q}|\}$, are those which have the minimum MI with respect to each other, given as

$$V = \frac{1}{|\mathbf{Q}|^2} \sum_{F_i, F_j \in \mathbf{Q}} MI(F_i, F_j). \tag{5.2}$$

Under the second constraint, the selected features will be maximally dissimilar with respect to each other, while under the first, the feature selection will be directed by the similarity with respect to the class labels. There are two major variants of the mRMR scheme: the MI difference (MID, given by U-V) and the MI quotient (MIQ, given by U/V). These represent different techniques to optimize the conditions associated with mRMR feature selection. We evaluated the use of both MID and MIQ for feature selection in this study, as well as determined an optimal number of features comprising each QIS by varying $|\mathbf{Q}|$ within the mRMR algorithm.

5.4 Experiments conducted to evaluate textural CaP signatures on a zone-wise basis

The specific experiments performed to evaluate $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ utilized a voxel-level QDA classifier [145] (see Appendix C) within the following,

- To determine the ability of $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ to specifically discriminate between CaP and benign regions on T2w MRI data in each of the CG and the PZ.
- To compare against alternatives to constructing zone-specific QISes, including:
  - utilizing all 110 extracted texture features with no feature selection ($\mathbf{F}$),
  - a randomly selected subset of texture features (denoted by $\mathbf{R}$).

Note that 2 separate voxel-level classification tasks were performed using each of $\mathbf{Q}, \mathbf{F}, \mathbf{R}$: first to identify regions of CG CaP, and then to identify PZ CaP. The CaP detection accuracy of the QDA classifier in each case was then considered as being reflective of the validity of $\mathbf{Q}, \mathbf{F}, \mathbf{R}$ in characterizing CG and PZ CaP respectively.

5.5 Classifier evaluation for examining zone-specific CaP signatures

Classification of the T2w MRI data was done on a per-voxel basis, with evaluation of the results against corresponding per-voxel annotations of CaP presence (via registration with histology). Thus, for all samples $c \in \mathbf{C}$, we directly compared the probabilistic classifier result $\mathbf{h}(c)$ with the label $l(c)$, at different thresholds of $\mathbf{h}(c)$. Plotting the true- and false-positive rates as function of varying the threshold of $\mathbf{h}(c)$ allowed us to perform Receiver-Operating Characteristic (ROC) curve analysis, with the Area under the ROC curve (AUC) being used as a measure of classifier performance [51, 83, 86].

In order to ensure robustness of the classifier to training and testing data, a randomized 3 fold cross-validation procedure was implemented. In a single cross-validation run, the datasets being considered (16 in the case of PZ CaP, 6 in the case of CG CaP) were divided into 3 randomized subsets (comprising 6, 5, and 5 studies for the PZ CaP detection problem and 2, 2, and 2 in the case of CG CaP). 2 subsets were considered as training data and the remaining as testing data, following which classification is performed. This was repeated until all 3 subsets were classified, and the entire cross-validation procedure was iterated 25 times. Feature selection and classifier construction were done separately for each set of training data (for all 3 folds over all 25 runs), with corresponding testing data only used for evaluation of classifier performance. All classifications were performed on a per-voxel basis.

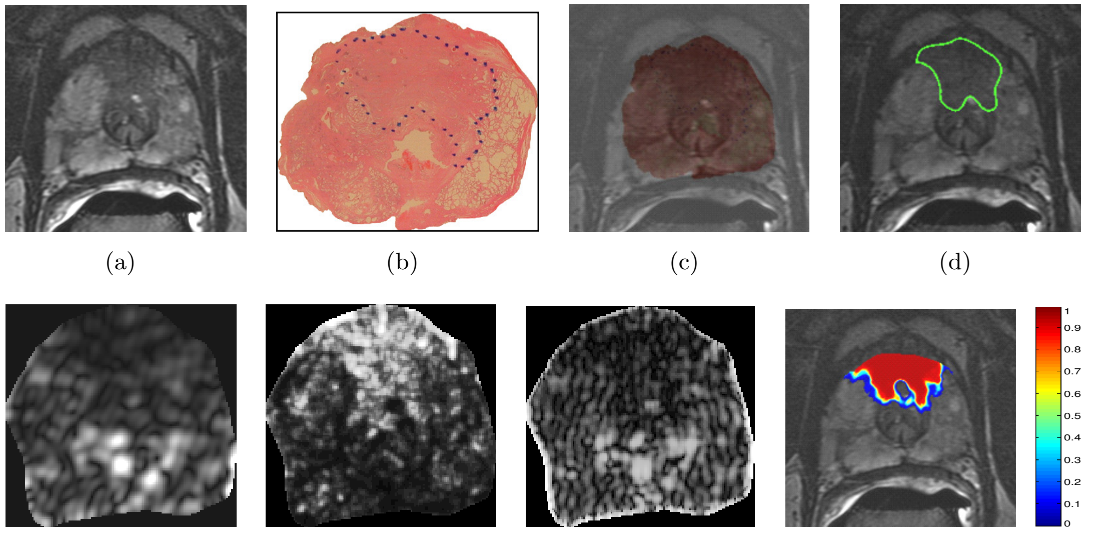
Each run of cross-validation yielded an AUC value (cumulatively calculated over all 3 folds); 25 AUC values were therefore calculated for each classification task (identifying CaP presence using different feature sets $\mathbf{Q}^{CG}, \mathbf{Q}^{PZ}, \mathbf{F}, \mathbf{R}$). The non-parametric pair-wise Wilcoxon rank-sum test was used to test for statistical significance when comparing the CaP detection AUCs of different feature sets ($\mathbf{Q}^{CG}, \mathbf{Q}^{PZ}, \mathbf{F}, \mathbf{R}$). The Bonferroni correction was employed to address the issue of multiple comparisons, thus correcting the p value used in testing for statistical significance within each individual Wilcoxon test from 0.05 to 1.67e-02. The null hypothesis for the Wilcoxon test in every case is that no statistically significant differences were observed when using $\mathbf{Q}^{CG}$ (or $\mathbf{Q}^{PZ}$ for QDA classification as compared to alternative feature sets ($\mathbf{F}, \mathbf{R}$).

5.6 Results

5.6.1 Feature selection to construct $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$

Table 5.2 summarizes the top 20 features ranked by the mRMR algorithm in terms of their discriminative ability for differentiating between CaP and benign regions within each zone (CG, PZ); these features constitute $\mathbf{Q}^{CG}$, $\mathbf{Q}^{PZ}$ respectively. Note that the combination of features 1 and 2 are considered to be more relevant (while minimizing redundancy) as compared to feature 1 alone, and so on. We examined this property of the QISes in more detail in Section 5.6.2 below.

$\mathbf{Q}^{CG}$ was seen to largely comprise of Gabor features, while $\mathbf{Q}^{PZ}$ mainly comprised Haralick texture features. 3 features were found to be in common between $\mathbf{Q}^{CG}$, $\mathbf{Q}^{PZ}$ (highlighted in italics in Table 5.2). Figure 5.1 shows representative $\mathbf{Q}^{CG}$ features derived from a T2w MRI section with CG CaP, while Figure 5.2 shows representative



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.1: Representative CG CaP dataset: (a) original 2D T2w MR image, (b) original 2D whole-mount histological section, with CG CaP extent outlined in blue (by a pathologist), (c) overlay of WMHS and T2w MRI after non-linear multi-modal registration, (d) mapped CG CaP extent on T2w MRI (outlined in green, from blue-outlined region in (b)). Representative texture features for this section: (e) Gabor (Orient = 157.5, Wavelength = 5.6569), (f) Haralick (energy, $w = 5$), and (g) Gabor (Orient = 0, Wavelength = 2.8284). Note the improved visual differentiability between CG CaP and benign regions on (e)-(g) texture feature images compared to the (a) original intensity image. (h) shows the probabilistic heatmap of CG CaP presence mapped back onto the image (via automated QDA classification). In (h), red indicates a high probability of CG CaP presence, blue indicates low probabilities of CG CaP presence, and no CaP was detected in the uncolored regions. This study was not used in training the classifier.

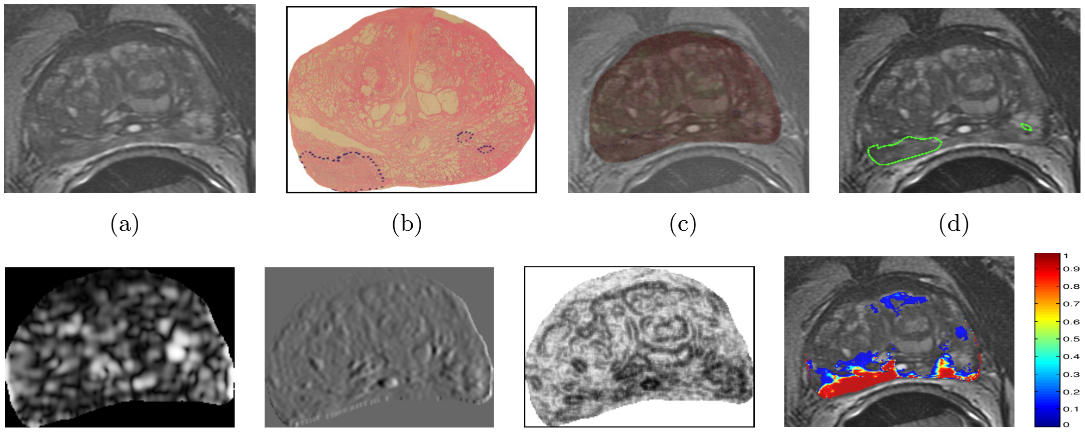
$\mathbf{Q}^{PZ}$ feature images corresponding to PZ CaP.

5.6.2 Selecting an optimal number of features to comprise each of $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{CG}$

The main free parameter associated with the MRMR scheme is the number of features to be selected ($|\mathbf{Q}|$). We empirically varied the number of features that were selected to comprise $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ (from 5 to 50), and evaluated the classification performance of each resulting $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ using a QDA classifier. The aim behind this experiment was to identify the minimal number of features required to yield a classifier AUC that most closely approached that obtainable by considering all 110 features. Figures 5.3(a)

	Top 20 discriminatory features for CG CaP		**Top 20 discriminatory features for PZ CaP**	
1	Gabor	Orient=157.5, Wavelength=5.6569	*Haralick*	*Difference Average* $(w = 7)$
2	Gabor	Orient=45, Wavelength=2.8284	Haralick	Sum Entropy $(w = 7)$
3	Gabor	Orient=0, Wavelength=2.8284	Haralick	Information Measure $(w = 5)$
4	Gabor	Orient=135, Wavelength=45.2548	Haralick	Difference Variance $(w = 7)$
5	Haralick	Energy $(w = 5)$	Gabor	Orient=0, Wavelength=5.6569
6	Haralick	Sum Average $(w = 5)$	Haar	Level 4 Horizontal Coefficient
7	Gabor	Orient=67.5, Wavelength=5.6569	Haralick	Entropy $(w = 7)$
8	Greylevel	Mean $(w = 5)$	Gabor	Orient=157.5, Wavelength=11.3137
9	*Haralick*	*Difference Average* $(w = 7)$	Gabor	Orient=157.5, Wavelength=8.2
10	*Haralick*	*Difference Entropy* $(w = 7)$	Gabor	Orient=135, Wavelength=11.3137
11	Gabor	Orient=0, Wavelength=45.2548	Haar	Level 2 Vertical Coefficient
12	Gabor	Orient=112.5, Wavelength=2.8284	*Haralick*	*Difference Entropy* $(w = 7)$
13	Haralick	Inverse Difference Moment $(w = 5)$	Gabor	Orient=112.5, Wavelength=45.2548
14	Gabor	Orient=90, Wavelength=45.2548	Gabor	Orient=0, Wavelength=8.2
15	Gabor	Orient=157.5, Wavelength=22.6274	*Gabor*	*Orient=67.5, Wavelength=2.8284*
16	Gabor	Orient=90, Wavelength=11.3137	Haralick	Information Measure $(w = 7)$
17	*Gabor*	*Orient=67.5, Wavelength=2.8284*	Gabor	Orient=0, Wavelength=11.3137
18	Greylevel	Standard Deviation $(w = 5)$	Gabor	Orient=112.5, Wavelength=5.6569
19	Gabor	Orient=112.5, Wavelength=11.3137	Greylevel	Kirsch
20	Gabor	Orient=157.5, Wavelength=45.2548	Gabor	Orient=67.5, Wavelength=45.2548

Table 5.2: Summary of top 20 features selected to accurately identify CG and PZ CaP respectively, obtained by voting of selected features across 25 cross-validation runs. Note that the 2 sets of features are relatively unique. 3 features that were found to be in common have been highlighted in italics (w stands for window size, an associated parameter setting for the feature).


(a) (b) (c) (d)

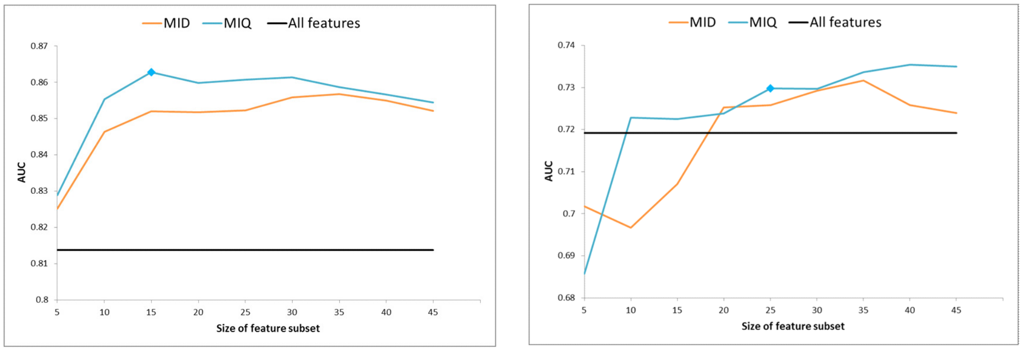
(e) (f) (g) (h)

Figure 5.2: Representative PZ CaP dataset: (a) original 2D T2w MR image, (b) original 2D whole-mount histological section, with PZ CaP extent outlined in blue (by a pathologist), (c) overlay of WMHS and T2w MRI after non-linear multi-modal registration, (d) mapped PZ CaP extent on T2w MRI (outlined in green, from blue-outlined region in (b)). Representative texture features for this section: (e) Gabor (Orient = 0, Wavelength = 5.6569), (f) Haar (Level 4 vertical coefficient), and (g) Haralick (Information measure, $w = 5$). Note the improved visual differentiability between PZ CaP and benign regions on (e)-(g) texture feature images compared to the (a) original intensity image. (h) shows the probabilistic heatmap of PZ CaP presence mapped back onto the image (via automated QDA classification). In (h), red indicates a high probability of PZ CaP presence, blue indicates low probabilities of PZ CaP presence, and no CaP was detected in the uncolored regions. This study was not used in training the classifier.

and 5.3(b) summarize the results of this evaluation for the CG and PZ, respectively.

For the CG, the performance of both variants of the mRMR scheme (MIQ and MID) was consistently found to improve upon using all 110 features (black). Overall, MIQ (blue) was seen to outperform MID (orange). The best performing subset (the most optimal $\mathbf{Q}^{CG}$) contained 15 features (selected via MIQ, highlighted with a blue polygon), with an AUC of 0.863 (standard error of 0.002 across 25 cross validation runs). In comparison, using all 110 features (black) yielded a significantly lower AUC of 0.814 (standard error of 0.002, p value = 1.37e-09).

For the PZ, the performance of both variants of the mRMR scheme (MIQ and MID) generally improved upon using all 110 features (black). Overall, MIQ (blue) was seen to outperform MID (orange). The best performing subset (the most optimal $\mathbf{Q}^{PZ}$) contained 25 features (selected via MIQ, highlighted with a blue polygon), with an AUC of 0.730 (standard error of 0.004). In comparison, using all 110 features (black)

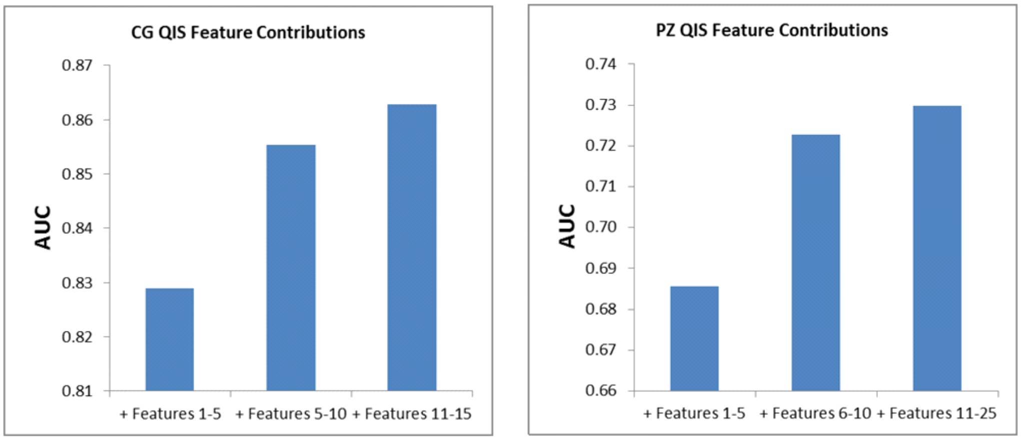


(a) (b)

Figure 5.3: Experiment to determine an optimal number of features to comprise (a) QIS for CG CaP and (b) QIS for PZ CaP. Trend in AUC value (Y-axis) of a QDA classifier is plotted as a function of different numbers of features comprising the QIS (X-axis). Note significant improvement when using mRMR feature selection (blue, orange) over using all features (black). The QIS comprising 15 features (selected via mRMR-MIQ, highlighted with a blue polygon) was found to provide the best overall performance for CG CaP detection. In comparison the QIS comprising 25 features (selected via mRMR-MIQ, highlighted with a blue polygon) was found to provide the best overall performance for PZ CaP detection. These performances were significantly better in comparison to to any other QIS size, as well as compared to using all 110 features (black).

yielded a lower AUC of 0.720 (standard error of 0.003).

We additionally evaluated the significance as well as the effectiveness (for CaP localization) of each feature $F_i \in \mathbf{Q}^{CG}$, $i \in \{1, \ldots, 15\}$, and $F_j \in \mathbf{Q}^{PZ}$, $j \in \{1, \ldots, 25\}$. Figure 5.4 summarizes the classification performance (in terms of AUC) achieved by incrementally considering features comprising each of $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$. For both the CG and PZ, it was seen that the top 5 features contributed most significantly to CaP classification performance. The inclusion of additional features only marginally improved the CaP classification performance over what was obtained by using the first 5 features to be included within $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ respectively. However, $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ in their entireties yielded the highest AUC values for localizing CaP in the CG (using the top 15 features) and PZ (using the top 25 features).



(a) (b)

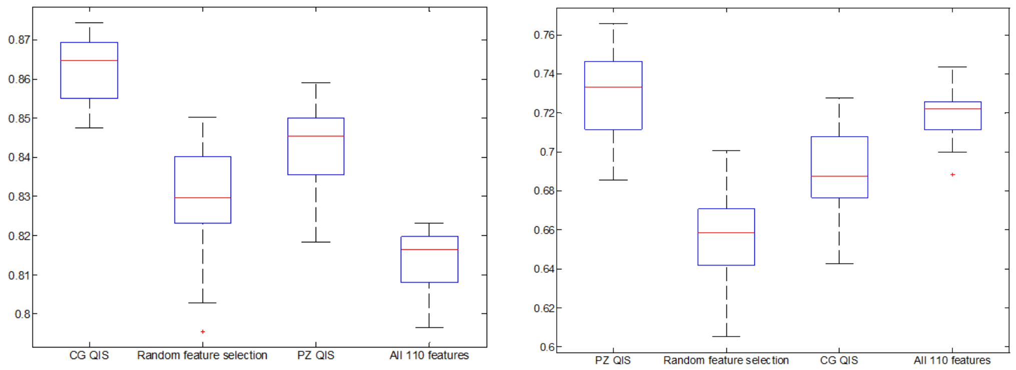
Figure 5.4: Performance of incrementally considering features which comprise CG and PZ QISes. Note the improvement in classification performance with addition of each additional set of features.

5.6.3 Comparing the use of $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$ for CaP detection against alternative feature sets

Figures 5.5(a) and 5.5(b) show box-and-whisker plots reflecting the CaP classification performance (AUC) of $\mathbf{Q}^{CG}$ and $\mathbf{Q}^{PZ}$, as compared to alternate strategies: (1) considering all 110 features ($\mathbf{F}$), (2) a random subset of features $\mathbf{R}$ (with the same number of features as the QIS), (3) using $\mathbf{Q}^{CG}$ to classify for PZ CaP and $\mathbf{Q}^{PZ}$ to classify for CG CaP. This last experiment is intended to evaluate the specificity of the QIS to the appearance of CaP in a particular zone, as well as highlight the importance of doing a zone-based classification. For the CG, a statistically superior performance ($p < 1.67e-02$) was obtained when using $\mathbf{Q}^{CG}$ as compared to using $\mathbf{F}$ ($p = 1.37e-09$), $\mathbf{R}$ ($p = 5.77e-09$), or $\mathbf{Q}^{PZ}$ ($p = 1.29e-07$). Similarly for the PZ, a statistically superior performance ($p < 1.67e-02$) was obtained when using $\mathbf{Q}^{PZ}$ as compared to using $\mathbf{R}$ ($p = 2.29e-09$) and $\mathbf{Q}^{CG}$ ($p = 1.12e-06$), but not $\mathbf{F}$ ($p = 7.42e-02$).

5.7 Discussion

Current clinical intuition suggests that the appearance of CG and PZ tumors on endorectal prostate T2w MRI is significantly different [23, 24]. Given the differing prognoses and outcomes of prostate cancer (CaP) based on its zonal location, there is a



(a) (b)

Figure 5.5: Box-and-whisker plot comparing performance of using (a) CG QIS for CG CaP detection and (b) PZ QIS for PZ CaP detection against alternative strategies: (1) using 15 (or 25) randomly selected features, and (2) using all 110 features (without feature selection/ranking). Additionally, we evaluate the performance of using the PZ QIS (25 features) for CG CaP detection and using the CG QIS (15 features) for PZ CaP detection. Using the zone-specific QIS was found to provide a statistically significantly superior performance for CaP detection in each zone, ratifying the effectiveness of the zone-specific textural signature compared to alternative strategies.

significant need for examining quantitative imaging signatures (QISes) specifically targeted to identifying CaP *in vivo*. The 2 major goals of our study were thus to, (1) define distinct textural signatures specific to CaP appearance in the CG and the PZ, and (2) quantitatively demonstrate the differences in these zone-specific QISes as well as evaluate their effectiveness for detecting CaP on T2w MRI. In this work, we have presented the first attempt at quantitatively defining QISes for CG and PZ tumors on T2w MRI. We believe that this work will allow for the building of targeted classifiers with the ability to incorporate spatial location of the disease into the model. This is a significantly different approach compared to the current trend of building monolithic computer-aided diagnostic models [83–86], which do not consider the zonal location of CaP in the prostate. Defining and utilizing zone-specific QISes to identify CG and PZ tumors will allow us to build computerized classifiers with improved sensitivity and specificity for CaP detection.

Our results showed that the QISes for each of CG and PZ CaP comprised largely non-overlapping textural attributes. These results appear to confirm current clinical intuition [23, 24], which suggests that CG and PZ CaP may have inherently differing appearances on MRI [29]. We found that the CG QIS was primarily comprised of the Gabor class of texture features (representative CG QIS images in Figure 5.1). The multi-scale, multi-orientated Gabor filters appears to be able to accurately model localized frequency characteristics [94] of the hypo-intense, homogeneous appearing CG CaP [23], allowing for discrimination from the heterogeneously appearing normal CG region. By comparison, the PZ CaP QIS was comprised largely of Haralick texture features (representative images in Figure 5.2). These features, which involve calculating image statistics derived from the co-occurrence of greylevel image intensities, appear to allow for accurate characterization and separation of low SI regions of CaP from hyper-intense normal prostatic regions in the PZ. Features that were found to overlap between the two QISes were high-level macro-resolution features (such as the Haralick energy and difference average features at large window sizes), implying a similarity between the two types of tumors at a gross scale. However, the largely non-overlapping nature of the two QISes at finer scales and resolutions appears to suggest fundamental micro-level textural differences between CG and PZ CaP on T2w MRI.

Significantly improved classification accuracy (via QDA) was achieved when using the zone-specific QISes to detect for CaP presence, as compared to (1) using all 110 features that were extracted, as well as (2) using a random subset of features. More importantly, interchanging the QISes (i.e. using $\mathbf{Q}^{CG}$ to classify for PZ CaP, and vice versa) also performed significantly worse compared to using the zone-specific QISes (i.e. using $\mathbf{Q}^{CG}$ to classify for CG CaP). Our findings suggest that CaP presence in different regions of the prostate is characterized by different structural and textural attributes, as captured by the QISes. The relatively high accuracy associated with the zone-specific QISes in detecting CaP (AUC of 0.86 in CG, 0.73 in the PZ) imply that we have largely optimized the classifier for CaP detection on T2w MRI. We expect that further combining the T2w information with DCE or DWI information will allow us to improve on CaP classification accuracy even more.

Our automated classification results are comparable, and in many cases superior, to other computerized decision support schemes in the literature; most of which have only been applied to CaP localization in the PZ alone. When using T2w intensities alone, Chan et al [83] reported an AUC of 0.59±0.14 (11 studies), Vos et al [84] reported an AUC of 0.84 (34 studies), and Ozer et al [85] reported an AUC of 0.7-0.8 (depending on the classifier used). Lopes et al [86] achieved a mean AUC of 0.88 for CaP localization by employing a subset of texture features considered in this work. When examining these results, it is worth noting that the current study differs from previous work in the following ways.

- *Annotation of CaP on MRI*: We have performed automated non-linear registration of T2w MRI and histology [11] in order to map CaP extent onto MRI. This is distinctly more rigorous compared to using either manual annotations [83,85,86] or an approximate affine registration between WMHS and MRI establish CaP extent on MRI [84]. The use of more rigorous techniques for mapping of CaP extents onto MRI leads us to have higher confidence in our results.
- *Resolution of analysis*: All of our analysis and evaluation was performed on a per-voxel basis, within an automatically determined ROI (via capsule segmentation [146]), and in 3D. In comparison, classification and evaluation in previous approaches were either done (1) using manually extracted regions-of-interest [83,84] from within the PZ alone, or (2) a per-pixel analysis of representative 2D sections [85,86].
- *Strength of magnet*: To our knowledge, this is the first work to present automated quantitative analysis of 3 Tesla endorectal prostate T2w MRI, compared to the use of 1.5 T endorectal or whole-body MRI in previous work [83–86].

This study did however have its limitations. First, our entire study comprised 22 pre-operative patient imaging datasets, 16 with PZ CaP and 6 with CG CaP. However this cohort size is comparable to other CAD studies for prostate cancer detection on MRI: Chan et al (11 datasets) [83], Ozer et al (20 datasets) [85], Lopes et al (27 datasets) [86], Madabhushi et al (5 datasets) [51]. To address the issue of the study being under-powered, a randomized cross validation procedure was employed for learning and evaluating the QISes. Additionally, we did not consider textural differences which may exist between different Gleason grades of CaP. We are currently collecting data which will allow us to explore such differences in future work. Registration of histology and MRI was performed in 2D. Performing the spatial alignment in 3D would have required 3D reconstruction of the histology volume, which was not possible due to (1) an insufficient number of histology slices being retained as part of the clinical protocol, leading to coarse and uneven inter-slice spacing of WMHS, and (2) fixation artifacts such as tissue loss or distortion (caused by prostate sectioning) and inconsistent H & E staining. It is also possible that the candidate zone-specific QISes determined by us in this study were selected based on textural differences between normal CG and PZ tissue, as opposed to the textural differences between CG and PZ CaP. However, an experiment to conclusively demonstrate that there are fundamental textural differences between CG and PZ CaP alone (rather than CG and PZ normal tissue) will require an additional segmentation of the CG and PZ within the prostate ROI. This additional zone-wise segmentation was not performed by us in this study as it is a significantly difficult problem, both for human experts as well as for image segmentation algorithms.

We believe developing quantitative zone-specific models of prostate cancer represent an important step in developing computational image analysis models for improved staging and detection of disease *in vivo*. Future work will involve further prospective validation of the candidate textural signatures determined in this study on a larger cohort of data, as well as incorporating additional protocols and features into our work, in order to develop a comprehensive multi-functional signature for CaP presence in vivo for use within an automated decision support system. We also intend to study the differences in textural appearance between CaP regions and common confounders for CaP presence (prostatitis, BPH).

Chapter 6

Comparing classifier ensembles for discriminating central gland and peripheral zone prostate tumors on T2-weighted prostate MRI

6.1 Overview of methodology to determine best automated classifier scheme to detect CaP on MRI

Below we describe the steps comprising the experimental design of this aim (see Figure 6.1). A more detailed explanation of the individual pre-processing steps was provided in Chapter 4. Appendix C contains a detailed description of each of the classifier techniques compared in this section.

Step Ia: Correcting for bias field and intensity non-standardness: The prostate region of interest is first segmented on the T2w MR image [135] (via a statistical shape model algorithm), following which it is corrected for intensity-based acquisition artifacts [16, 136].

Step Ib: Registration of *ex vivo* histology and *in vivo* T2w MRI data: In order to obtain CaP extents that are as accurate as possible on MRI, a 2D registration strategy is adopted wherein corresponding planar *ex vivo* whole-mount histological sections (WMHS) and *in vivo* T2w MRI sections are non-linearly registered to one another [11]; correspondences between WMHS and T2w MRI sections having previously been established via visual inspection by a pathologist and a radiologist in unison.

Step Ic: Texture feature extraction: Multiple different classes of texture features (Gabor [94], Haar [98], first and second order texture [92, 93]) are extracted to distinguish between PZ and CG tumors on T2w MRI.

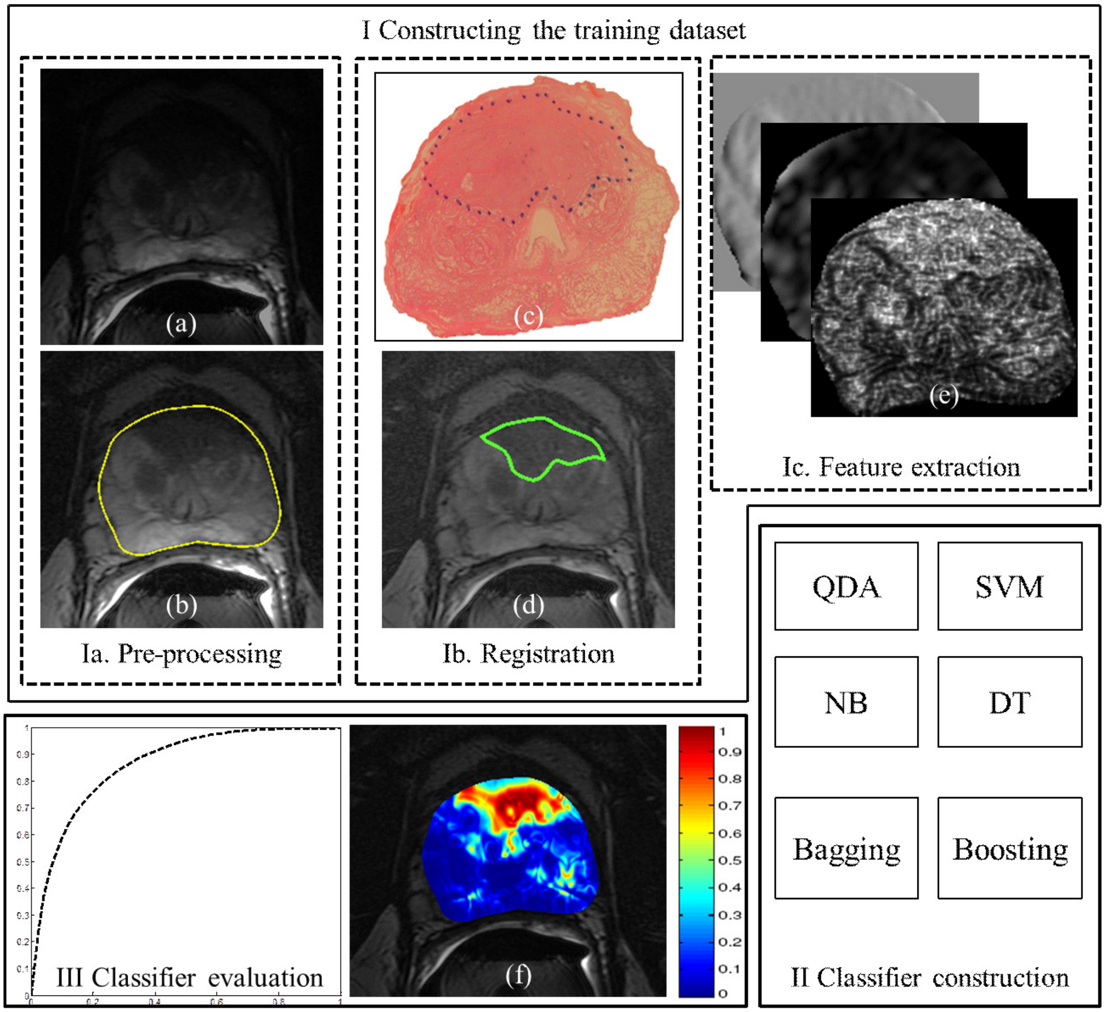


Figure 6.1: Overview of overall experimental design. Solid boxes represent the individual modules comprising our CAD scheme for CaP detection on T2w MRI. Step Ia involves pre-processing T2w MRI (original image in (a)) to segment the prostate ROI (outlined in yellow on (b)) and correct for intensity artifacts (note intensity differences between (a) and (b)). Step Ib then involves non-linearly registering WMHS (shown in (c)) with T2w MRI to determine ground truth CaP extent on T2w MRI (green outline in (d)). Step Ic involves extraction of texture features to quantify CaP presence on T2w MRI (representative features shown in (e)). The results of Step I are input to different classifiers in Step II. In Step III, classifier results from Step II (visualized as a probability image in (f), where red corresponds to a high likelihood of CaP presence) are evaluated on a per-voxel basis against CaP ground truth (Step Ib).

Step II: Classifier training for PZ and CG CaP separately: 12 classifiers were considered in this work, QDA [145], SVMs [117], naïve Bayes [116], and DTs [118], as well as their bagging and boosting variants, and were trained separately for classification of CG and PZ CaP, respectively. See Appendix C for a detailed description of the classifier techniques.

Step III: Classifier comparison and evaluation for PZ and CG separately: Results obtained via the 12 different classifiers are compared in terms of accuracy, model complexity, and execution time. Classifier accuracy is evaluated in terms of area overlap (on a per-voxel basis) with the surrogate for ground truth CaP extent (obtained in Step Ib).

6.2 Construction of CG- and PZ-specific CaP classifiers

6.2.1 Notation specific to this chapter

Appendix A summarizes commonly used notation and symbols appearing in this chapter. After feature extraction, every voxel $c \in C$ is associated with a 110-dimensional feature vector denoted $\widehat{\mathbf{F}}(c) = \{f_1(c), f_2(c), \ldots, f_{110}(c)\}$, for every $c \in C$. Similar to Chapter 5, the result of mRMR feature selection [144] was 2 distinct sets of features: one corresponding to CG CaP (denoted $\mathbf{F}^{CG}(c)$) and the other to PZ CaP (denoted $\mathbf{F}^{PZ}(c)$). Each set is comprised of texture features considered highly discriminatory for differentiating between CaP and benign regions in each of the two prostatic zones. The feature sets $\mathbf{F}^{CG}(c)$ and $\mathbf{F}^{PZ}(c)$ were each input to the different classification algorithms and their ensemble variants. The classifier construction and evaluation comprised a number of steps, described below.

6.2.2 Feature normalization

Normalization of features ensures that different feature values lie in a comparable range of intensity values when input to a classifier. Given a feature vector $\mathbf{F}(c)$, this can be done for each $f_i(c) \in \mathbf{F}(c)$ as follows,

$$f_i(c) = \frac{f_i(c) - \mu_{\mathbf{i}}}{\sigma_i}, \tag{6.1}$$

where μ_i is the mean and σ_i is the mean absolute deviation (MAD) corresponding to feature $i, i \in \{1, \ldots, N\}$. As a result of normalization, $\forall c \in C$, each feature in $\mathbf{F}^{CG}(c)$ and $\mathbf{F}^{PZ}(c)$ was transformed to have a mean of 0 and a MAD of 1.

6.2.3 Class balancing

A significant issue when training a supervised classifier is the *minority class problem* [147], wherein the target class (in this study ω_{+1}) has significantly fewer samples compared to the other class (ω_{-1}), i.e. $|\omega_{+1}| \ll |\omega_{-1}|$. Weiss et al [147] and Doyle et al [148] previously showed that using an imbalanced training set will likely result in a lower classifier accuracy compared to balanced training sets ($|\omega_{+1}| = |\omega_{-1}|$). The class balance problem was addressed for each of the base classifiers, as well as their ensemble variants. Note that class balancing and data sub-sampling was only applied to the training data in each case.

(a) QDA, DTs: For classifiers corresponding to these two families ($\mathbf{h}^{QDA}$, $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{DT}$, $\mathbf{h}^{Bag,DT}$, $\mathbf{h}^{Boost,DT}$), class imbalance was accounted for by randomized under-sampling of the majority class (ω_{-1}) such that $|\omega_{+1}| = |\omega_{-1}|$, i.e. an equal class balance was maintained when training the classifier.

(b) SVMs: Due to the complex nature of this algorithm, not only did a class balance have to be ensured in the training data, but the number of samples (voxels) used to train the classifier had to be reduced to ensure convergence within a reasonable amount of time. When training an SVM classifier, an equal number of voxels (not less than $0.2 \times |\omega_{+1}|$) were randomly sub-sampled from both ω_{+1} and ω_{-1} classes to form the training dataset. The number of samples was empirically decided based on a trade-off between execution time, classifier accuracy, and memory constraints specific to the SVM classifier. This procedure was adopted for all classifiers in the SVM family ($\mathbf{h}^{SVM}$, $\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Boost,SVM}$).

(c) Naïve Bayes: Training of the naïve Bayes classifier was implemented by directly estimating distributions for each of the classes, ω_{+1} and ω_{-1}, based on all the samples present. Such an estimate is most accurate when the maximal number of samples is utilized in calculating the distribution. Thus, no sub-sampling of the data was performed when constructing these classifiers ($\mathbf{h}^{Bay}$, $\mathbf{h}^{Bag,Bay}$, $\mathbf{h}^{Boost,Bay}$).

6.2.4 Classifier training

In order to avoid training bias, both three-fold cross-validation (3FCV) as well as leave-one-out cross-validation (LOOCV) were utilized; both implemented on a patient study basis. Feature selection and classifier construction were done separately for each set of training data so constructed, with corresponding testing data only used for evaluation of classifier performance. All classifications were performed and evaluated on a per-voxel basis.

Leave One Out Cross Validation (LOOCV): When classifying for PZ CaP via LOOCV, the PZ CaP data cohort was split into a training set of 15 patient studies with the 1 remaining study being used for testing. This process was repeated until each of 16 patient studies had been classified at least once. A single AUC value was calculated cumulatively over all 16 classification results; in addition the upper and lower bounds on the AUC were determined. Similarly in the case of CG CaP, the training set comprised 5 patient studies with 1 study being held out for testing at each iteration. The 6 classification results were then cumulatively evaluated to obtain a single AUC value along with the upper and lower bounds.

Three Fold Cross Validation (3FCV): This was performed as described previously in Section 5.5, and was done separately for the CG and PZ CaP detection tasks.

6.3 Evaluation of CG and PZ specific CaP classifiers

6.3.1 Classifier accuracy

Depending on the type of classifier used, the per-voxel classifier result $\mathbf{h}(c)$ can correspond to a probability value ($\mathbf{h}(c) \in [0, 1]$) or a hard decision ($\mathbf{h}(c) \in \{0, 1\}$).

In the case of $\mathbf{h}^{QDA}(c)$, $\mathbf{h}^{Bay}(c)$, $\mathbf{h}^{Bag,\beta}(c)$, $\mathbf{h}^{Boost,\beta}(c)$, $\beta \in \{QDA, Bay, SVM, DT\}$, which yield a probabilistic result, a binary prediction result at every $c \in C$ can be obtained by thresholding the associated probability value $\mathbf{h}(c) \in [0,1]$. Classifier evaluation was done via Receiver Operating Characteristic (ROC) curves [116], representing the trade-off between classification sensitivity and specificity. The vertical axis of the ROC curve is the true positive rate (TPR) or sensitivity, and the horizontal axis is the false positive rate (FPR) or 1-specificity, while each point on the curve corresponds to the sensitivity and specificity of detection at some $\rho \in [0,1]$. ROC curves were visualized by fitting a smooth polynomial through each set of sensitivity and specificity values calculated for each of the 25 3FCV runs, and averaging over the 25 curves generated for each classifier considered. The area under the ROC curve (AUC) is used as a measure of classification performance, as is commonly reported in the literature [51, 52, 83–86].

$\mathbf{h}_\rho(c)$ is defined as a binary prediction result at each threshold $\rho \in [0,1]$, such that

$$\mathbf{h}_\rho(c) = \begin{cases} 1 & \text{when } \mathbf{h}(c) > \rho, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

For the set of samples C, the corresponding detection result is given as $\Psi_\rho(C) = \{c|\mathbf{h}_\rho(c) = 1\}, c \in C$. Based on overlap with the ground truth $G(C)$, per-voxel detection sensitivity (SN), and specificity (SP) can be calculated at each threshold ρ as [149]

$$SN_\rho = 1 - \frac{|G(C) - \Psi_\rho(C)|}{|G(C)|} \text{ and } SP_\rho = 1 - \frac{|\Psi_\rho(C) - G(C)|}{|C - G(C)|}. \tag{6.3}$$

In the case of $\mathbf{h}^{SVM}(c)$ and $\mathbf{h}^{DT}(c)$, the output is a single hard partitioning of the sample $c \in C$ into one of the two classes under consideration. For these instances, a single detection result is calculated at a single threshold, given as $\Psi(C) = \{c|\mathbf{h}(c) = 1\}, c \in C$, based on which a single value for SP and SN can be calculated. It is assumed that the remaining points on this ROC curve are at $[0,0]$ and $[1,1]$, hence allowing the construction of a pseudo-ROC curve, and therefore calculation of an AUC value.

While analyzing ROC results, the 12 classifiers were segregated into 3 groups, (1) single classification strategies (comprising $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, $\mathbf{h}^{SVM}$, $\mathbf{h}^{DT}$), (2) bagging strategies (comprising $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,Bay}$, $\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Bag,DT}$), and (3) boosting strategies (comprising $\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,Bay}$, $\mathbf{h}^{Boost,SVM}$, $\mathbf{h}^{Boost,DT}$). Classifier comparisons were first made within each group (e.g. which of the single classification strategies $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, $\mathbf{h}^{SVM}$, and $\mathbf{h}^{DT}$ performed best), following which classifier performance across groups were examined.

6.3.2 Statistical testing

For the 3FCV procedure, each classifier yielded a set of 25 AUC values (corresponding to each cycle of the procedure) and for each classification task (CG and PZ CaP classification).

Multiple comparison testing to determine statistically significant differences in performance within groups (e.g. between all of $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, $\mathbf{h}^{SVM}$, $\mathbf{h}^{DT}$) was performed using the Kruskal-Wallis (K-W) one-way analysis of variance (ANOVA) [150]. The K-W ANOVA is a non-parametric alternative to the standard ANOVA test which does not assume normality of the distributions when testing. The null hypothesis for the K-W ANOVA was that the populations from which the AUC values originate have the same median. Based off the results of a K-W ANOVA, multiple comparison testing was performed to determine which groups (single classification strategies, bagging strategies, boosting strategies) show significant differences in performance.

Pairwise comparisons were performed for classifiers across groups (e.g. between $\mathbf{h}^{QDA}$ and $\mathbf{h}^{Bag,QDA}$) to identify statistically significant differences in performance. This was done using the non-parametric Wilcoxon rank-sum test [150]. The null hypothesis in such a case was that there were no statistically significant differences in AUC values between the 2 classifiers being compared.

The Bonferroni correction [150] was applied to correct the p-value within all statistical comparisons considered (whether pairwise or other).

6.3.3 Evaluating classifier variance

Studying the variance characteristics of ensemble classifiers (such as bagging) may allow us to better understand trends in classifier performance [70]. Variance between classifiers implies that different component classifiers make different decisions about a sample despite being trained on the same training set. Breiman [38] noted that improving the performance of the bagged classifier, $\mathbf{h}^{Bag}$, is dependent on increasing the variance between the component classifiers $h_t, t \in \{1, \ldots, T\}$, combined within the bagging framework (see Equation 10.9, T refers to the number of classifiers combined). In this work, the variance between component classifiers is approximated by calculating the *disagreement* between the weak classifiers that are combined within the bagging framework. Disagreement was calculated for every sample $c \in C$ as,

$$\delta(c) = 1 - \max\left(\frac{Q}{T}, \frac{T-Q}{T}\right), \tag{6.4}$$

where $Q < T$ is the number of component classifiers h_t that determine sample c as belonging to class ω_{+1}. Note that $0 \leq \delta(c) \leq 0.5$, where if $\delta(c) = 0$, all h_t have made the same decision about the class of sample c, and if $\delta(c) = 0.5$, 50% of the classifiers "disagree" about this decision. Plotting the distribution of $\delta(c)$ over all the classified samples can then help illustrate whether the majority of h_t agreed or disagreed in their decisions. Note that if this distribution is skewed towards $\delta(c) = 0$ (i.e. the component classifiers h_t largely tend to agree in their decisions), bagging these classifiers would be expected to demonstrate poor performance.

6.3.4 Model complexity

Each classifier/classifier ensemble was evaluated in terms of model complexity: (1) number of parameters for the classification algorithm, and (2) number of hierarchy levels within the algorithm. The term "hierarchy" refers to the number of classifier stages a sample undergoes. For instance, $\mathbf{h}^{Bag,QDA}$ has two hierarchy levels (since it involves two types of classifiers – Bagging and QDA) while $\mathbf{h}^{QDA}$ has only one.

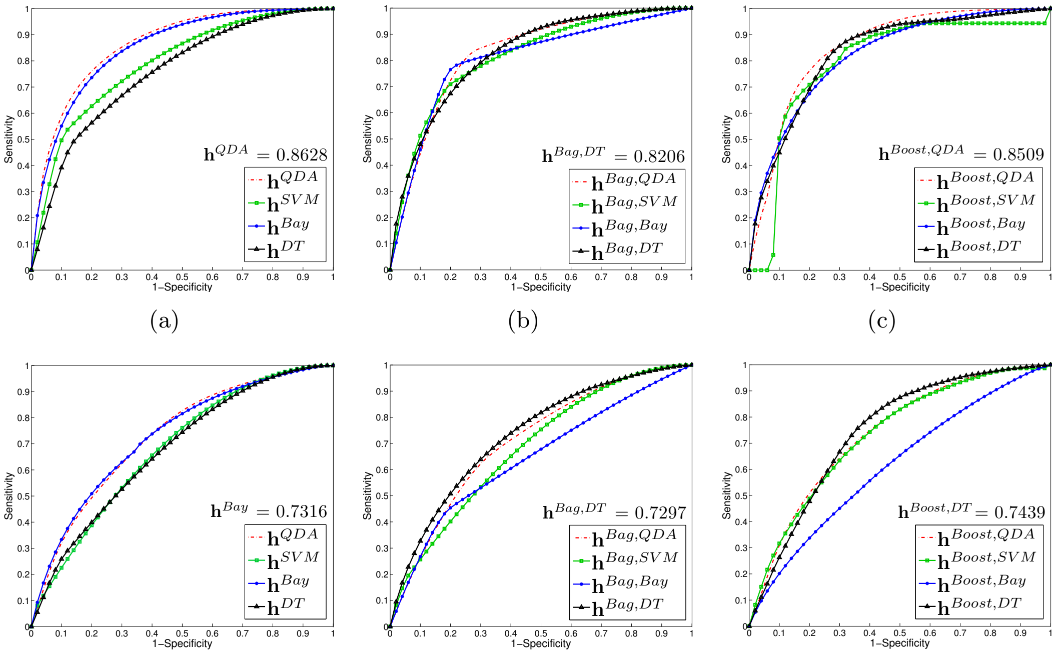
6.3.5 Computation time

For each of the classifiers compared, $\mathbf{h}^{\beta}, \mathbf{h}^{Bag,\beta}, \mathbf{h}^{Boost,\beta}$, $\beta \in \{QDA, Bay, SVM, DT\}$, the total amount of time required for (i) classifier construction, and (ii) for executing the constructed classifier on testing data, was recorded in seconds. The execution time for each classifier was averaged over 5 runs. All algorithms were implemented and evaluated using MATLAB®7.10 (The Mathworks, MA) on a 72 GB RAM, 2 quad core 2.33 GHz 64-bit Intel Core 2 processor machine.

6.4 Results

6.4.1 Classification accuracy

Figure 6.2 shows average ROC curves for CaP classification performance in the CG (top row) and PZ (bottom row), respectively. Figures 6.4 and 6.5 show corresponding



(a) (b) (c)

(d) (e) (f)

Figure 6.2: ROC curves obtained by averaging over 25 runs of 3FCV for CaP classification in (a)-(c) CG and (d)-(f) PZ. In each graph different colors correspond to different classifier strategies: (a), (d) $\mathbf{h}^{QDA}$ (red), $\mathbf{h}^{SVM}$ (green), $\mathbf{h}^{Bay}$ (blue), $\mathbf{h}^{DT}$ (black); (b), (e) $\mathbf{h}^{Bag,QDA}$ (red), $\mathbf{h}^{Bag,SVM}$ (green), $\mathbf{h}^{Bag,Bay}$ (blue), $\mathbf{h}^{Bag,DT}$ (black), and (c), (f) $\mathbf{h}^{Boost,QDA}$ (red), $\mathbf{h}^{Boost,SVM}$ (green), $\mathbf{h}^{Boost,Bay}$ (blue), $\mathbf{h}^{Boost,DT}$ (black).

boxplots of classifier AUC values obtained over the 25 3FCV runs, for the CG and PZ CaP classification tasks respectively. Table 6.1 summarizes the AUC values (obtained via LOOCV) along with the computation times for different methods for the CG CaP classification task. Also enumerated are the model parameters associated with each classifier. Table 6.2 similarly summarizes the corresponding results for the PZ CaP classification task (also via LOOCV).

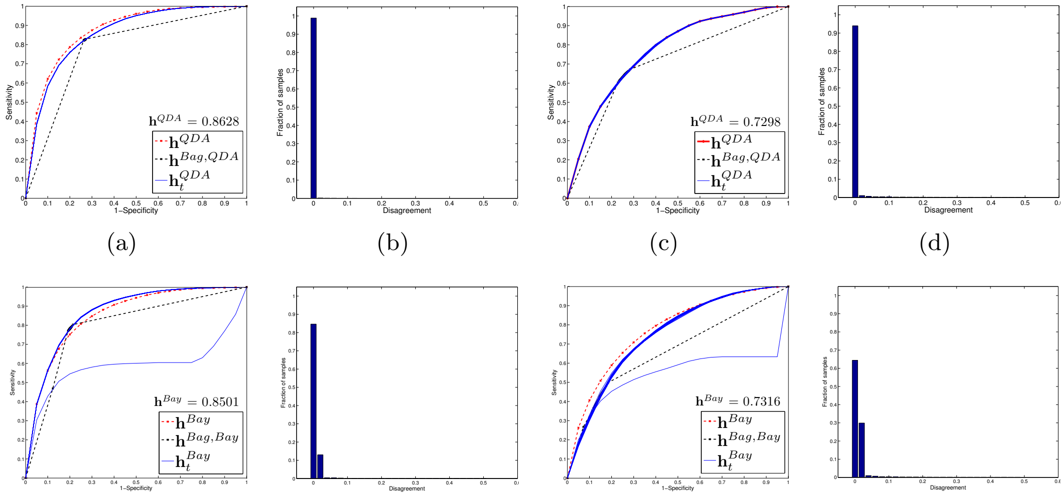
Comparing single classification strategies

Figures 6.2(a), (d) illustrate that both $\mathbf{h}^{QDA}$ (red) and $\mathbf{h}^{Bay}$ (blue) yielded the best performance for CaP classification in both the CG and PZ, with no statistically significant differences in their AUC values in either zone. In comparison, $\mathbf{h}^{SVM}$ (green) and $\mathbf{h}^{DT}$ (black) demonstrate relatively poorer performance. A multiple comparison test of AUC values (based off K-W ANOVA) determined that $\mathbf{h}^{SVM}$ and $\mathbf{h}^{DT}$ performed significantly worse in terms of AUC compared to $\mathbf{h}^{QDA}$ and $\mathbf{h}^{Bay}$; a finding that was consistent for both CG and PZ CaP detection tasks.

Comparing bagging strategies

Figures 6.2(b), (e) demonstrate that $\mathbf{h}^{Bag,DT}$ yielded a significantly improved CaP classification performance compared to all of $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,SVM}$, and $\mathbf{h}^{Bag,Bay}$, in both the CG and the PZ. A multiple comparison test of AUC values in the CG (based off K-W ANOVA), showed that there were no statistically significant differences between $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,SVM}$, and $\mathbf{h}^{Bag,Bay}$. In the PZ, no significant differences were observed between the AUC values for $\mathbf{h}^{Bag,QDA}$ and $\mathbf{h}^{Bag,Bay}$, while $\mathbf{h}^{Bag,SVM}$ was found to perform significantly worse compared to $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,Bay}$, and $\mathbf{h}^{Bag,DT}$. However, a significant improvement in classifier performance for $\mathbf{h}^{Bag,SVM}$ and $\mathbf{h}^{Bag,DT}$ was seen in both zones, compared to using $\mathbf{h}^{SVM}$ or $\mathbf{h}^{DT}$ (Wilcoxon test $p < 0.01$). In contrast, both $\mathbf{h}^{Bag,QDA}$ and $\mathbf{h}^{Bag,Bay}$ demonstrated significantly worse performance (Wilcoxon test $p < 0.01$) than they did as individual classifiers ($\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$), in both zones.

Figures 6.3(a) and (c) show the ROC curves for $\mathbf{h}^{QDA}$ (red), $\mathbf{h}^{Bag,QDA}$ (black), and those of the individual $h_t^{QDA}, t \in \{1, \ldots, T\}$, (which were combined to calculate



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6.3: Visualizing the variance within bagging when using QDA (top row) and naïve Bayes (bottom row) to construct the bagged classifier, respectively. (a), (e), show ROC curves corresponding to $\mathbf{h}^{\alpha}$ (red), $\mathbf{h}^{Bag,\alpha}$ (black), and h_t^{α} (blue), $\alpha \in \{QDA, Bay\}, t \in \{1, \ldots, T\}$, for the CG CaP classification task, while (c), (g) similarly show ROC curves for the PZ CaP classification task. Note that the blue curves show almost perfect overlap, implying little variance between the component classifiers h_t^{QDA} and h_t^{Bay}. (b), (d) show the distribution of $\delta(c)$ between $h_t^{QDA}, t \in \{1, \ldots, T\}$, for the CG and the PZ respectively. Similarly, (f), (h) show the distribution of $\delta(c)$ for $h_t^{Bay}, t \in \{1, \ldots, T\}$, in the CG and the PZ respectively. These distributions are seen to be highly skewed towards $\delta(c) = 0$ (X-axis), implying a majority of the component classifiers agree for most of the samples.

$\mathbf{h}^{Bag,QDA}$, shown in Figures 6.3(a) and (c) in blue), for the CG and PZ respectively. $\mathbf{h}^{QDA}$ performed better than $\mathbf{h}^{Bag,QDA}$ as well as any of $h_t^{QDA}, t \in \{1, \ldots, T\}$. By visualizing the distribution of disagreement ($\delta(c)$) over all the samples classified in a single CV run (Figures 6.3(b),(d)), it was seen that a majority of h_t^{QDA} agree for a majority ($> 90\%$) of the samples, in both the CG and the PZ; also reflected in Figures 6.3(a) and (c). Both these observations imply that there is low variance between all of $h_t^{QDA}, t \in \{1, \ldots, T\}$, resulting in poor performance when they are combined via bagging. Similar observations about the ROC curves and δ can be made in the case of $\mathbf{h}^{Bay}$ (red), $\mathbf{h}^{Bag,Bay}$ (black), and $h_t^{Bay}, t \in \{1, \ldots, T\}$, (which were combined to calculate $\mathbf{h}^{Bag,Bay}$, shown in Figures 6.3(e) and (g) in blue), shown in Figures 6.3(e)-(h). Note that the Y-axis of the graphs in Figures 6.3(b),(d),(f),(h) have been normalized by the total number of testing samples considered (n).

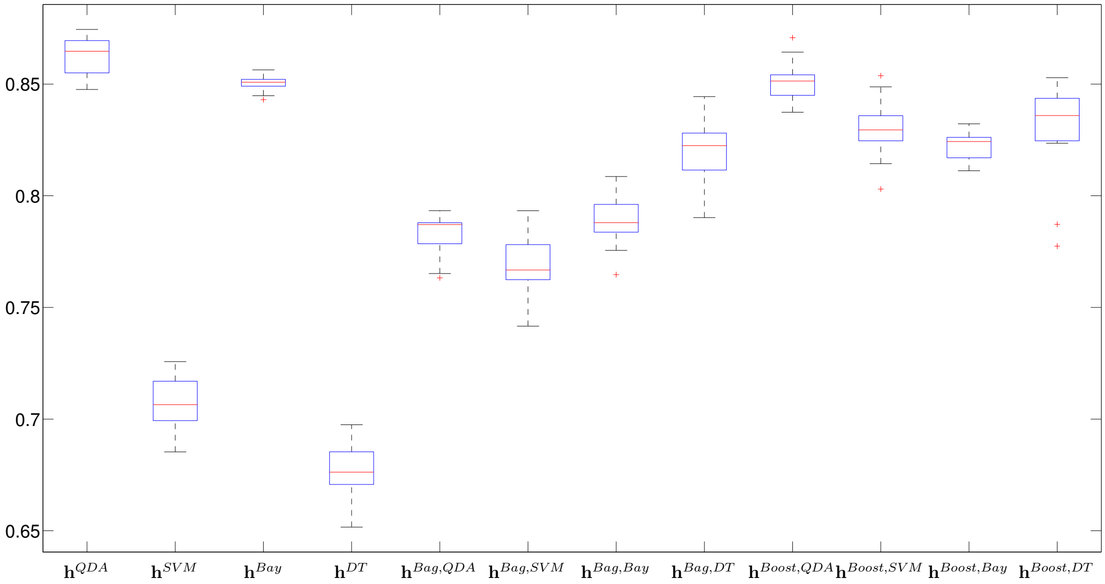


Figure 6.4: Box-and-whisker plot of AUC values for each of 12 classifiers compared in the CG CaP detection task, obtained across 25 runs of 3FCV. Note that the red line in the middle of each box reflects the median AUC value while the box is bounded by 25^{th} and 75^{th} percentile of AUC values. The whisker plot extends to the minimum and maximum AUC values (obtained across all 25 runs) outside the box and outliers are denoted via the red plus symbol.

Bauer and Kohavi [69] as well as Breiman [70] have previously noted that classifiers such as QDA and naïve Bayes have low variance (implying a lower performance in conjunction with bagging). In contrast, SVMs and DTs are known to have high variance [70], in turn suggesting that they would perform better in conjunction with bagging ($\mathbf{h}^{Bag,SVM}$ and $\mathbf{h}^{Bag,DT}$).

Comparing boosting strategies

In the CG, $\mathbf{h}^{Boost,QDA}$ yielded in the best CaP detection performance performance, while a multiple comparison test of AUC values (based off K-W ANOVA) showed that $\mathbf{h}^{Boost,SVM}$ and $\mathbf{h}^{Boost,DT}$ did not yield significantly different AUC values. In the PZ, all 3 of $\mathbf{h}^{Boost,SVM}$, $\mathbf{h}^{Boost,DT}$, and $\mathbf{h}^{Boost,QDA}$ performed comparably. $\mathbf{h}^{Boost,Bay}$ demonstrated significantly worse performance compared to $\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,DT}$, and $\mathbf{h}^{Boost,SVM}$ in both the CG and the PZ. Note that $\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,DT}$, and $\mathbf{h}^{Boost,SVM}$ yielded a marginal but significantly improved performance compared to $\mathbf{h}^{Bag,QDA}$,

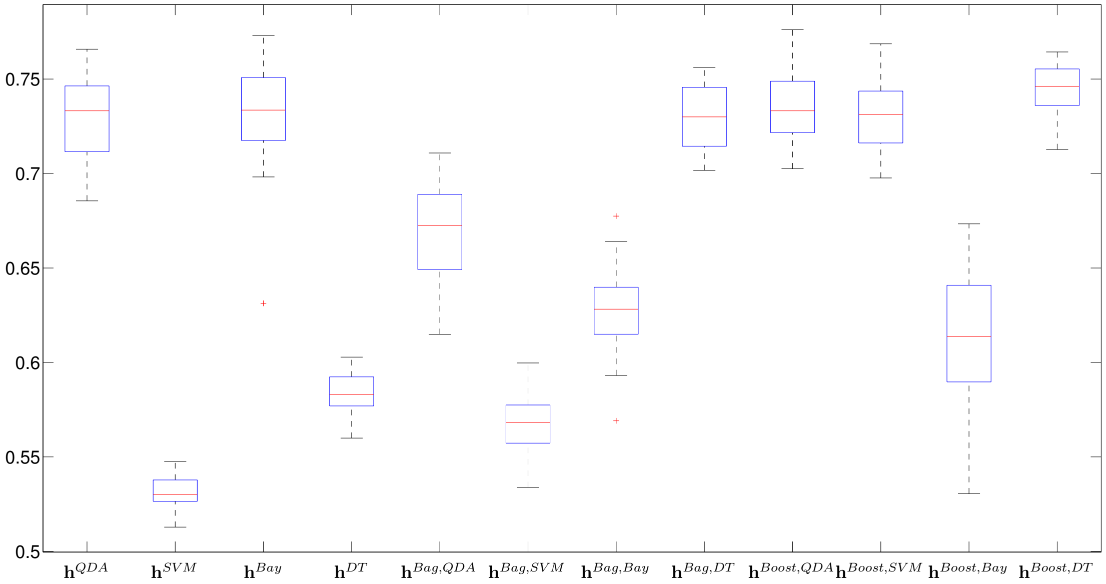


Figure 6.5: Box-and-whisker plot of AUC values for each of 12 classifiers compared in the PZ CaP detection task, obtained across 25 runs of 3FCV. Note that the red line in the middle of each box reflects the median AUC value while the box is bounded by 25^{th} and 75^{th} percentile of AUC values. The whisker plot extends to the minimum and maximum AUC values (obtained across all 25 runs) outside the box and outliers are denoted via the red plus symbol.

$\mathbf{h}^{Bag,DT}$, and $\mathbf{h}^{Bag,SVM}$ in both the CG and the PZ (Figures 6.2(c), (f)).

The poor performance of $\mathbf{h}^{Boost,Bay}$ can potentially be explained in terms of bias and variance requirements of ensemble frameworks. As discussed previously, optimal performance of bagging is highly dependent on the component classifiers exhibiting high variance [38,70]. In contrast, boosting is dependent on the component classifiers having low bias [70] and thus classifiers such as $\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,DT}$, and $\mathbf{h}^{Boost,SVM}$ show significant performance improvements compared to $\mathbf{h}^{QDA}$, $\mathbf{h}^{DT}$, and $\mathbf{h}^{SVM}$, respectively. In comparison, naïve Bayes, a classifier with high bias as well as low variance [70], causes both $\mathbf{h}^{Bag,Bay}$ and $\mathbf{h}^{Boost,Bay}$ to demonstrate a relatively poor performance compared to $\mathbf{h}^{Bay}$ in both zones. By contrast, the QDA classifier has low bias as well as low variance. Thus while $\mathbf{h}^{Boost,QDA}$ and $\mathbf{h}^{QDA}$ perform comparably, $\mathbf{h}^{Bag,QDA}$ performs significantly worse than either of them.

Classifier	AUC		Parameters	Total
	μ^{AUC}	Confidence bounds		execution time (seconds)
$\mathbf{h}^{QDA}$	**.8713**	.8006 - .9420	-	2
$\mathbf{h}^{SVM}$	.7112	.6284 - .7940	$\lambda = 1$	2388^{*}
$\mathbf{h}^{Bay}$	.8508	.7661 - .9355	-	$396^{\dagger}$
$\mathbf{h}^{DT}$	.6878	.6247 - .7509	-	43
$\mathbf{h}^{Bag,QDA}$	.7908	.7435 - .8381	$T = 50$	37.5
$\mathbf{h}^{Bag,SVM}$	.7846	.6987 - .8705	$T = 50, \lambda = 1$	5566^{*}
$\mathbf{h}^{Bag,Bay}$	.7865	.7179 - .8551	$T = 50$	$7680^{\dagger}$
$\mathbf{h}^{Bag,DT}$	.8310	.7596 - .9024	$T = 50$	186
$\mathbf{h}^{Boost,QDA}$	.8589	.7777 - .9400	$T = 50$	10.4
$\mathbf{h}^{Boost,SVM}$	.8316	.7624 - .9008	$T = 50$	13116^{*}
$\mathbf{h}^{Boost,Bay}$	.8270	.7351 - .9189	$T = 50$	$171^{\dagger}$
$\mathbf{h}^{Boost,DT}$	.8450	.7732 - .9168	$T = 50, L = 5$	38

Table 6.1: Mean AUC values and confidence bounds (obtained via LOOCV) for CG CaP classification, as well as parameter settings and total execution times for the different classifiers considered in this work. * SVM classifiers were constructed after significant sub-sampling of the training dataset while simultaneously ensuring class balance (see Section 6.2.3). † Naïve Bayes classifiers were constructed without any sub-sampling of the training dataset (see Section 6.2.3).

6.4.2 Classifier complexity

Tables 6.1 and 6.2 summarize the parameters that were set for each of the 12 classifiers considered. The SVM family of classifiers has a single parameter λ (to normalize the kernel representation) which was set at the default, as prescribed in previous work [52]. The remaining classifier families (QDA, naïve Bayes, DTs) did not have any parameters that required setting or tuning. All of $\mathbf{h}^{QDA}$, $\mathbf{h}^{DT}$, $\mathbf{h}^{Bay}$, and $\mathbf{h}^{SVM}$ have a single level of complexity, as they operate directly off the training data. Each of the ensemble schemes ($\mathbf{h}^{Bag,\beta}, \mathbf{h}^{Boost,\beta}, \beta \in \{QDA, Bay, DT, SVM\}$) have additional level of model complexity. $\mathbf{h}^{Boost,DT}$ has an additional parameter to be set reflecting the number

Classifier	**AUC**		**Parameters**	**Total**
	μ^{AUC}	**Confidence bounds**		**execution time (seconds)**
$\mathbf{h}^{QDA}$	.7349	.6864 - .7835	-	3
$\mathbf{h}^{SVM}$	.5328	.5186 - .5470	$\lambda = 1$	2496*
$\mathbf{h}^{Bay}$	.7386	.6796 - .7975	-	660†
$\mathbf{h}^{DT}$	.5870	.5559 - .6181	-	44.4
$\mathbf{h}^{Bag,QDA}$	.6722	.6278 - .7166	$T = 50$	54.8
$\mathbf{h}^{Bag,SVM}$	.5881	.5604 - .6158	$T = 50, \lambda = 1$	9768*
$\mathbf{h}^{Bag,Bay}$	.6294	.5734 - .6854	$T = 50$	14700†
$\mathbf{h}^{Bag,DT}$	.7501	.6956 - .8047	$T = 50$	193.3
$\mathbf{h}^{Boost,QDA}$	.7427	.6941 - .7912	$T = 50$	11.4
$\mathbf{h}^{Boost,SVM}$	.7377	.6927 - .7826	$T = 50, \lambda = 1$	18996*
$\mathbf{h}^{Boost,Bay}$	.6318	.5804 - .6831	$T = 50$	308.2†
$\mathbf{h}^{Boost,DT}$	**.7471**	.7117 - .7825	$T = 50, L = 5$	53

Table 6.2: Mean AUC values and confidence bounds (obtained via LOOCV) for PZ CaP classification, as well as parameter settings and computation times for the different classifiers considered in this work. * SVM classifiers were constructed after significant sub-sampling of the training dataset while simultaneously ensuring class balance (see Section 6.2.3). † Naïve Bayes classifiers were constructed without any sub-sampling of the training dataset (see Section 6.2.3).

of levels (L) for each node. L was set to 5, representing the best trade-off between execution time and accuracy, and was determined from the literature [52, 151].

6.4.3 Classifier execution time

Tables 6.1 and 6.2 summarize the total computation time for training and evaluating the different classifiers considered in this study for all 3 folds of a single 3FCV run (averaged over 5 such CV runs). All computations were performed on a 32 GB RAM, 2 quad core 2.33 GHz 64-bit Intel Core 2 cluster machine. $\mathbf{h}^{QDA}$ required the least amount of computation time, followed by $\mathbf{h}^{DT}$. $\mathbf{h}^{SVM}$ required the most time for training and evaluating the classifier amongst all algorithms; training and testing times were even longer for $\mathbf{h}^{Bag,SVM}$ and $\mathbf{h}^{Boost,SVM}$. SVM classifiers also required more careful memory management due to the thousands of samples being considered (voxel-wise classification) within a relatively complex algorithm (utilizing kernels).

Bagging was seen to typically increase computation time by a factor of 5, while boosting increased the computation time by a factor of 20. This may be because with bagging, the component classifiers are trained on smaller bootstrapped sample subsets, whereas with boosting, the component classifiers are trained on the entire set of training samples.

6.5 Discussion

The primary motivation for this work was to identify the appropriate classifier ensemble in the context of datasets encountered for most CAD problems; ones that are usually limited by small sample sizes and questionable class labels (SSS-QCL). In this work, we quantitatively compared 12 different classifier ensembles derived from 4 different classifier strategies; QDA, Bayesian learners, Decision Trees, and Support Vector Machines. The 12 classifier ensembles were compared in terms of accuracy, training and execution time, and model complexity for the computerized detection of prostate cancer from high resolution T2w MRI on a zonal basis. Most classifier comparison studies that have been previously reported in the machine learning and pattern recognition literature have typically not considered SSS-QCL datasets [69, 71–75]. A secondary motivation of this study was to investigate whether classifier trends previously reported on large databases

	Accuracy	**Execution time**	**Complexity**	**Overall**
Best	$\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$	$\mathbf{h}^{QDA}$,$\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,DT}$	$\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$	$\mathbf{h}^{QDA}$
Worst	$\mathbf{h}^{SVM}$,$\mathbf{h}^{DT}$	$\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Boost,SVM}$, $\mathbf{h}^{Bag,Bay}$	$\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Boost,SVM}$, $\mathbf{h}^{Boost,DT}$	$\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Boost,SVM}$

Table 6.3: Best and worst classifier ensembles in the context of classification accuracy, execution time, and model complexity, and overall.

of images with reliable ground truth were also valid for SSS-QCL datasets. While our findings are based on a specific dataset for a specific CAD problem, we believe that similar trends will be observable for other SSS-QCL CAD problems.

Our primary findings from this work were the following,

- For the 2 class SSS-QCL CAD problem considered in this study, the QDA classifier appears to offer an optimal trade-off between accuracy, training and testing time, and model complexity. For a majority of the classification tasks, the QDA classifier yielded the highest accuracy which, coupled with its low model complexity and execution time, made it the best classifier overall. In contrast, the SVM classifier demonstrated among the lowest classifier accuracies, was among the more complex classifiers, and took the longest to train and test.

- Boosting marginally but significantly outperformed bagging across most classifier strategies. However, the trade-off in execution time appears to negate some of its advantages. It is interesting to note that integrating boosting within the decision tree framework resulted in extremely high classification accuracy (PBTs yielded the highest AUC for PZ CaP classification).

- Satisfying the conditions of bias and variance are extremely crucial when constructing classifier ensembles. While SVMs and DTs show significant improvements within both bagging and boosting frameworks, Bayesian and QDA classifiers provided a more mixed performance as they suffered from low variance and/or high bias.

- Satisfying the independence assumption within the naïve Bayes classifier allowed for extremely efficient and relatively robust classifier construction. In fact, the individual naïve Bayes classifier (without use of bagging or boosting) provided the second highest classification performance within both the CG and PZ. It also provided the most consistent performance in the two zones, where other classifiers performed well in one zone but not the other.

- In the context of the specific problem of zone-wise CaP detection via T2w MRI, we achieved significantly high classification accuracies: an AUC of 0.8628 for CG CaP classification (via QDA), and AUC of 0.7439 for PZ CaP classification (via PBTs). These are comparable and in many cases better than detection rates reported by state-of-the-art CaP detection schemes for prostate T2w MRI [51, 52, 83–86].

- Not all classifiers performed as well in each zone, with significantly lower classifier performances observed in the PZ compared to the CG. This may reflect on the need for better training data in the case of PZ CaP, a problem compounded by the smaller size of PZ tumors that were observed our study.

We do however acknowledge a few limitations of our study. The cohort size was relatively limited (22 studies), but is comparable to other CaP detection studies in the literature (5-36 studies) [51, 52, 83–86]. However, the small cohort size is representative of the typical training set sizes previously reported for CAD schemes [85–88]. In this work we also attempted to minimize as many sources of error as possible, leveraging automated registration methods for determining disease extent on the imaging data, unlike previous work [51,52,83,85,86] where manual delineations of disease extent, which are known to be highly error prone, were employed. Additionally, extensive evaluation of the classifiers (both three-fold and leave-one-out cross validation were evaluated) was performed to ensure classifier robustness and generalizability.

When considering datasets where limited and/or erroneous labeled training data is available, semi-supervised and active learning methodologies [148] have begun to become extremely popular. These algorithms appear particularly suited to SSS-QCL problems as they may allow for improving the quality of labels of individual instances, or allow for more intelligent labeling of the data.

In the context of the specific SSS-QCL problem considered in this work, we limited ourselves to the use of T2w MRI as opposed to a multi-parametric MRI exam. There has been relatively little work on identifying MP MRI signatures for discriminating CG and PZ CaP on multi-parametric MRI. In contrast, the appearance of CG and PZ tumors on T2w MRI is well documented [23, 24]. Based on the results of this study, however, there is every indication to suggest that our conclusions regarding the classifiers will hold for multi-parametric MRI data.

In this study we limited ourselves to only 12 classifiers, primarily based off not wanting to dramatically increase the size of the study and number of subsequent experiments to be performed. However, based on the 4 distinct types of classifiers considered in this study, we believe our results may be generalized to other classifier families (e.g. relevance vector machines are similar to SVMs). While we have presented our findings based off only a single CAD problem and a single dataset, we believe that our findings, in addition to the lessons learnt in the context of SSS-QCL, are applicable to other medical imaging CAD problems.

Chapter 7

Application of consensus embedding for detection of CaP via multi-parametric MRI

In this chapter, we first demonstrate the application of consensus embedding for CaP detection from uni-modal (T2w) MRI. Following this, we present its application within a framework for multi-parametric data integration, which we have termed Enhanced Multi-Protocol Analysis via Intelligent Supervised Embedding (EMPrAvISE).

7.1 CaP detection on *ex vivo* prostate T2w MRI data

Two different prostates were imaged *ex vivo* using a 4 Tesla MRI scanner following surgical resection. The excised glands were then sectioned into 2D histological slices which were digitized using a whole slide scanner. Regions of cancer were determined via Haemotoxylin and Eosin (H&E) staining of the histology sections. The cancer areas were then mapped onto corresponding MRI sections via a deformable registration scheme [11]. Additional details of data acquisition are described in [51].

For this experiment, a total of 16 4 Tesla *ex vivo* T2-weighted MRI and corresponding digitized histology images were considered. The purpose of this experiment was to accurately identify cancerous regions on prostate MRI data via pixel-level classification, based on exploiting textural differences between diseased and normal regions on T2-weighted MRI [51]. For each MRI image, M embeddings, $\mathbb{R}^n_m, m \in \{1, \ldots, M\}$, were first computed (via $CreateEmbed$) along with their corresponding embedding strengths $\psi(\mathbb{R}^n_m)$ (based on clustering classification accuracy). Construction of the consensus embedding was performed via a supervised cross-validation framework, which utilized independent training and testing sets for selection of strong embeddings ($SelEmbed$). The algorithm proceeds as follows,

(a) Training (S^{tr}) and testing (S^{te}) sets of the data (MRI images) were created.

(b) For each element (image) of S^{tr}, strong embeddings were identified based on $\theta = 0.15 \times \max_M [\psi(\mathbb{R}^n_m)]$.

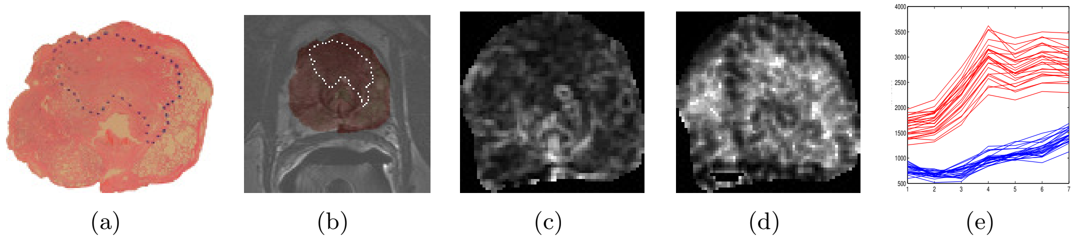
(c) Those embeddings voted as being strong across all the elements (images) in S^{tr} were then identified and selected.

(d) For the data (images) in S^{te}, corresponding embeddings were then combined (via *CalcConsEmbed*) to yield the final consensus embedding result.

A leave-one-out cross-validation strategy was employed in this experiment. A comparison is made between the pixel-level classifications (via replicated k-means clustering [8]) for (1) simple GE denoted as $\Psi(\mathbf{X}_{GE})$, and (2) consensus GE denoted as $\Psi(\widetilde{\mathbf{X}}_{GE})$. Note that this experimental setup is similar to those described in Chapter 3.

7.2 CaP detection on *in vivo* multi-parametric MRI data

The visual appearance of CaP on the different MRI protocols is summarized in Table 1.1 (based on radiologist and quantitative CAD-derived descriptors). A total of 5 image texture features were calculated from each of $\mathcal{C}^{T2}$ as well as $\mathcal{C}^{ADC}$. These include first and second order statistical features, as well as non-steerable gradient features. The extracted texture features and the corresponding intensity values were concatenated



(a) (b) (c) (d) (e)

Figure 7.1: (a) Original WMHS with CaP outline in blue (by a pathologist). (b) Overlay of deformed WMHS image $\mathcal{C}^H$ (via MACMI) onto $\mathcal{C}^{T2}$, allowing mapping of CaP extent (outlined in white). Representative texture features (derived within the prostate ROI alone) are also shown for (c) $\mathcal{C}^{T2}$ and (d) $\mathcal{C}^{ADC}$. Note the improvement in image characterization of CaP compared to original intensity information in (a) and (c), respectively. (e) Corresponding time-intensity curves for CaP (red) and benign (blue) regions are shown based on DCE MRI data. Note the differences in the uptake and wash-out characteristics between the red and blue curves.

to form the feature vectors $\mathbf{F}^{T2}(c) = [f^{T2}(c), f_{\phi}^{T2}(c)|\phi \in \{1, \ldots, 5\}]$ (from $\mathcal{C}^{T2}$) and $\mathbf{F}^{ADC}(c) = [f^{ADC}(c), f_{\phi}^{ADC}(c)|\phi \in \{1, \ldots, 5\}]$ (from $\mathcal{C}^{ADC}$), associated with every voxel $c \in C$. Representative feature images derived from $\mathcal{C}^{T2}$ and $\mathcal{C}^{ADC}$ are shown in Figures 7.1(c) and (d).

The wash-in and wash-out of the contrast agent within the gland is characterized by varying intensity values across the time-point images $\mathcal{C}^{T1,t}, t \in \{1, \ldots, 7\}$ (Figure 7.1(e)). This time-point information is directly concatenated to form a single feature vector $\mathbf{F}^{T1}(c) = [f^{T1,t}(c)|t \in \{1, \ldots, 6\}]$ associated with every voxel $c \in C$.

7.2.1 Overview of algorithm $EMPrAvISE$

The major steps of consensus embedding (Chapter 3) are summarized below within a single algorithm to represent and fuse multi-parametric MR data (existing in feature space $\mathbf{F}(c)$). Note that once the consensus embedding representation $\widetilde{\mathbb{R}}^n$ has been constructed, we may construct a classifier to distinguish the different object classes within $\widetilde{\mathbb{R}}^n$.

```
Algorithm $EMPrAvISE$
**Input**: $\mathbf{F}(c) \in \mathbb{R}^N$ for all objects $c$, $n$, $M, V, \theta$
**Output**: $\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n$
*begin*
    0. Construct feature space $\mathbf{F}(c) \in \mathbb{R}^N, \forall c \in C$ (via feature extraction);
    1. *for* $m = 1$ *to* $M$ *do*
    2.     Calculate $\mathbf{X}_m(c) = CreateWeakEmbed(\mathbf{F}(c)|\mathcal{F}_m, M, V), \forall c \in C$,
             hence yielding $\mathbb{R}^n_m$;
    3.     k=0;
    4.     Calculate $\psi^{Acc}(\mathbb{R}^n_m)$ (based on classification accuracy);
    5.     *if* $\psi^{Acc}(\mathbb{R}^n_m) > \theta$
    6.         k++;
    7.         $W_k(i,j) = \|\mathbf{X}_m(c) - \mathbf{X}_m(d)\|_2 \; \forall c, d$ with indices $i, j$;
    8.     *endif*
    9. *endfor*
    10. $\widetilde{W}(i,j) = \text{MEDIAN}_k \left[W_k(i,j)\right] \forall c, d$;
    11. Apply MDS to $\widetilde{W}$ to obtain $\widetilde{\mathbb{R}}^n$;
    12. Train a classifier on $\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n, \forall c \in C$, to distinguish object-class categories;
*end*
```

Every voxel $c \in C$ was characterized by a number of different multi-parametric feature vectors (summarized in Table 7.1). For the purposes of comparing $EMPrAvISE$ with an alternative data representation scheme, a multi-attribute vector $\mathbf{F}^{Feats}(c)$ is also constructed by directly concatenating the individual T2w, DCE, and ADC attributes.

7.2.2 Constructing the consensus embedding representation of multi-parametric MRI data

The algorithm $EMPrAvISE$ was applied to the feature vector $\mathbf{F}^{Feat}(c) \in \mathbb{R}^N, N = 18, |\mathbb{R}^N| = |C|$, i.e. for all voxels $c \in C$. We denote $\mathcal{F}$ as the superset of all multi-parametric features, such that $|\mathcal{F}| = N$. Note that $\mathcal{F} = \mathcal{F}_{T2} \cup \mathcal{F}_{T1} \cup \mathcal{F}_{ADC}$ where $\mathcal{F}_{T2}, \mathcal{F}_{T1}, \mathcal{F}_{ADC}$ are feature sets associated with the individual T2w, DCE, ADC protocols respectively. Feature space perturbation was implemented by first forming M bootstrapped subsets of features $\mathcal{F}_m \subset \mathcal{F}$. These features were randomly drawn from $\mathcal{F}$ such that (1) $|\mathcal{F}_u| = |\mathcal{F}_v| = V$, (2) $\mathcal{F}_u \cap \mathcal{F}_v \neq \emptyset$, (3) each of N features appears in at least one $\mathcal{F}_m$, and (4) one feature from each of $\mathcal{F}_{T2}, \mathcal{F}_{T1}, \mathcal{F}_{ADC}$ appears in each

Description		Data vectors	Classifier
Single Protocol	T2w	$\mathbf{F}^{T2}(c) = [f^{T2}(c), f^{T2}_{\phi}(c) \mid \phi \in \{1, \ldots, 5\}]$	$\mathbf{h}^{T2}(c)$
	DCE	$\mathbf{F}^{T1}(c) = [f^{T1,t}(c) \mid t \in \{1, \ldots, 6\}]$	$\mathbf{h}^{T1}(c)$
	ADC	$\mathbf{F}^{ADC}(c) = [f^{ADC}(c), f^{ADC}_{\phi}(c) \mid \phi \in \{1, \ldots, 5\}]$	$\mathbf{h}^{ADC}(c)$
Multi-parametric	Features	$\mathbf{F}^{Feat}(c) = [\mathbf{F}^{T2}(c), \mathbf{F}^{T1}(c), \mathbf{F}^{ADC}]$	$\mathbf{h}^{Feat}(c)$
	$EMPrAvISE$	$\mathbf{F}^{Em}(c) = [\widetilde{e}_v(c) \mid v \in \{1, \ldots, n\}]$	$\mathbf{h}^{Em}(c)$, $\mathbf{h}^{Em}_{MRF}$

Table 7.1: Different feature datasets and corresponding classifier strategies considered for multi-parametric data analysis.

$\mathcal{F}_m$, where $u, v, m \in \{1, \ldots, M\}$. The feature space associated with each feature subset $\mathcal{F}_m$ was then embedded in n-D space via GE [9], yielding M corresponding weak embeddings $\mathbb{R}^n_m$.

The corresponding M embedding strengths, $\psi^{Acc}(\mathbb{R}^n_m)$, were then calculated based on the supervised classification accuracy of a probabilistic boosting tree classifier [151] (additional details in Appendix C), using labels $l(c), \forall c \in C$. A leave-one-out cross-validation approach was utilized in the training and evaluation of this PBT classifier. Embeddings with $\psi^{Acc}(\mathbb{R}^n_m) > \theta$ were then selected as strong, and combined as described in Chapter 3. The final result of $EMPrAvISE$ is the consensus embedding vector $\mathbf{F}^{Em}(c) = [\widetilde{e}_v(c) | v \in \{1, \ldots, n\}] \in \widetilde{\mathbb{R}}^n, \forall c \in C$ (n, the intrinsic dimensionality, is estimated via the technique presented in [122]).

7.2.3 Classification of multi-parametric MRI via PBTs

A voxel-level probabilistic boosting tree classifier (PBT) classifier was constructed for each feature set, $\mathbf{F}^\beta(c)$, $\beta \in \{T1, T2,\ ADC, Feats, Em\}, \forall c \in C$, considered in Table 7.1. The PBT algorithm has recently demonstrated success in the context of multi-modal data analysis [152] as it leverages a powerful ensemble classifier (Adaboost) in conjunction with the robustness of decision tree classifiers [151] to allow for the computation of weighted probabilistic decisions for difficult to classify samples. As a result of PBT classification, we obtain a posterior conditional probability of belonging to the cancer class, denoted $\mathbf{h}^\beta(c) = p(l(c) = 1 | \mathbf{F}^\beta(c)) \in [0, 1]$, $\beta \in \{T1, T2, ADC, Feats, Em\}$, for every voxel $c \in C$.

Incorporating spatial constraints via Markov Random Fields

We have previously demonstrated the use of a novel probabilistic pairwise Markov model (PPMMs) to detect CaP lesions on prostate histopathology [36], via the incorporation of spatial constraints to a classifier output. PPMMs formulate Markov priors in terms of probability densities, instead of the typical potential functions [153], facilitating the creation of more sophisticated priors. We make use of this approach to similarly impose spatial constraints to the classifier output (per-voxel), with the objective of accurately segmenting CaP lesions on MRI.

7.2.4 Performance Evaluation Measures

We define $\mathbf{h}^{\beta}_{\rho}(c)$ as the binary prediction result for classifier $\mathbf{h}^{\beta}(c)$ at each threshold $\rho \in [0,1]$, such that $\mathbf{h}^{\beta}_{\rho}(c) = 1$ when $\mathbf{h}^{\beta}(c) \geq \rho$, 0 otherwise; $\forall \beta \in \{T1, T2, ADC, Feats, Em\}$. For every scene $\mathcal{C}$, threshold ρ, and classifier $\mathbf{h}^{\beta}(c)$, the set of voxels identified as CaP is denoted $\Psi^{\beta}_{\rho}(C) = \{c | \mathbf{h}^{\beta}_{\rho}(c) = 1\}, c \in C$, $\forall \beta \in \{T1, T2, ADC, Feats, Em\}$. We then perform ROC analysis as described in Section 6.3. Note that a leave-one-out cross validation strategy over the 39 slices was used to evaluate the performance of each of the classifiers constructed (Table 7.1).

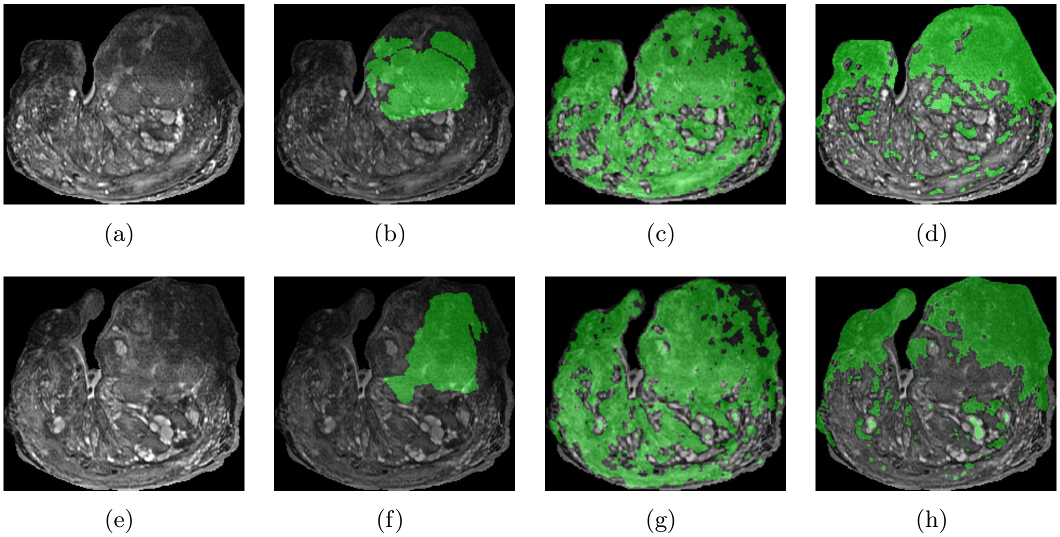


Figure 7.2: (a), (e) 2D sections from 3D prostate MRI data, and (b), (f) corresponding CaP masks superposed on the MRI, obtained via deformable registration with the corresponding histology slice (not shown) [11]. Corresponding CaP detection results via (c), (g) $\Psi(\mathbf{X}_{GE})$ (graph embedding), and (d), (h) $\Psi(\widetilde{\mathbf{X}}_{GE})$ (consensus embedding) are superposed back onto the original MRI sections ((a), (e)). In each of (b)-(d) and (f)-(h), green denotes the CaP segmentation region. Note the significantly fewer false positives in (d) and (h) compared to (c) and (g) respectively.

7.3 Results

7.3.1 Detecting CaP on *ex vivo* MRI data

Figure 7.2 shows qualitative results of detecting prostate cancer (CaP) on T2-weighted MRI, each row corresponding to a different 2D MRI image. Comparing the pixel-level CaP detection results (visualized in green) in Figures 7.2(c) and 7.2(g) to the green CaP masks in Figures 7.2(b) and 7.2(f), obtained by registering the MRI images with corresponding histology images [11] (not shown), reveals that $\Psi(\mathbf{X}_{GE})$ results in a large false positive error. In contrast, $\Psi(\widetilde{\mathbf{X}}_{GE})$ (Figures 7.2(d) and 7.2(h)) appears to better identify the CaP region when compared to the ground truth for CaP extent in Figures 7.2(b) and 7.2(f). Figure 7.3 illustrates the relative pixel-level prostate cancer detection accuracies averaged across 16 MRI slices for the 2 methods compared. $\Psi(\widetilde{\mathbf{X}}_{GE})$ was found to significantly ($p < 0.05$) outperform $\Psi(\mathbf{X}_{GE})$ in terms of accuracy and specificity of CaP segmentation over all 16 slices considered.

7.3.2 Detection of CaP on *in vivo* multi-parametric MRI

Comparison of EMPrAvISE against individual feature based classifiers

We first compared $\mathbf{h}^{Em}$ (via $EMPrAvISE$) against classifiers constructed using the different uni-modal feature sets corresponding to T2w, DCE, and DWI MRI data ($\mathbf{h}^{T2}, \mathbf{h}^{T1}, \mathbf{h}^{ADC}$). As may be gleaned from Table 7.3.2, $\mathbf{h}^{Em}$ yields a higher classification accuracy and AUC compared to $\mathbf{h}^{T2}, \mathbf{h}^{T1}, \mathbf{h}^{ADC}$.

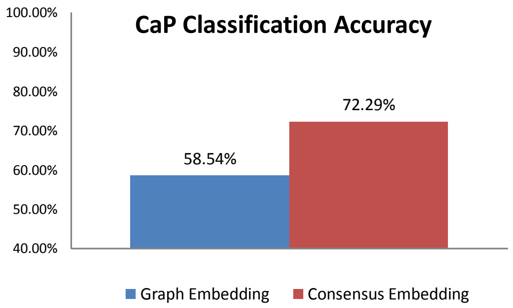


Figure 7.3: Pixel-level classification accuracy in identifying prostate cancer on T2-weighted MRI, averaged over 16 2D MRI slices for $\Psi(\mathbf{X}_{GE})$ (blue) and $\Psi(\widetilde{\mathbf{X}}_{GE})$ (red).

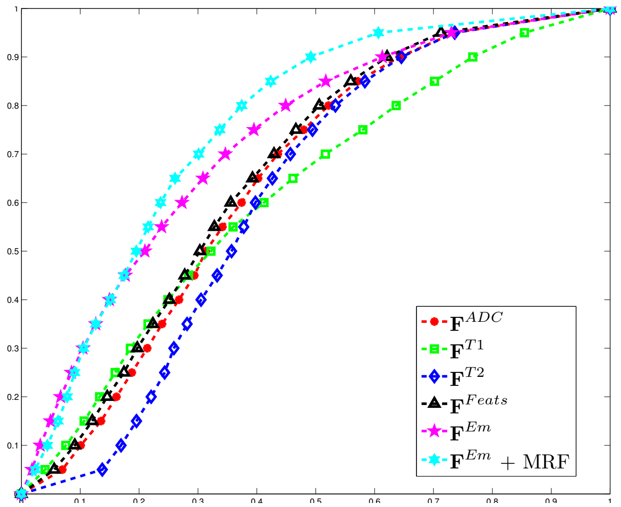


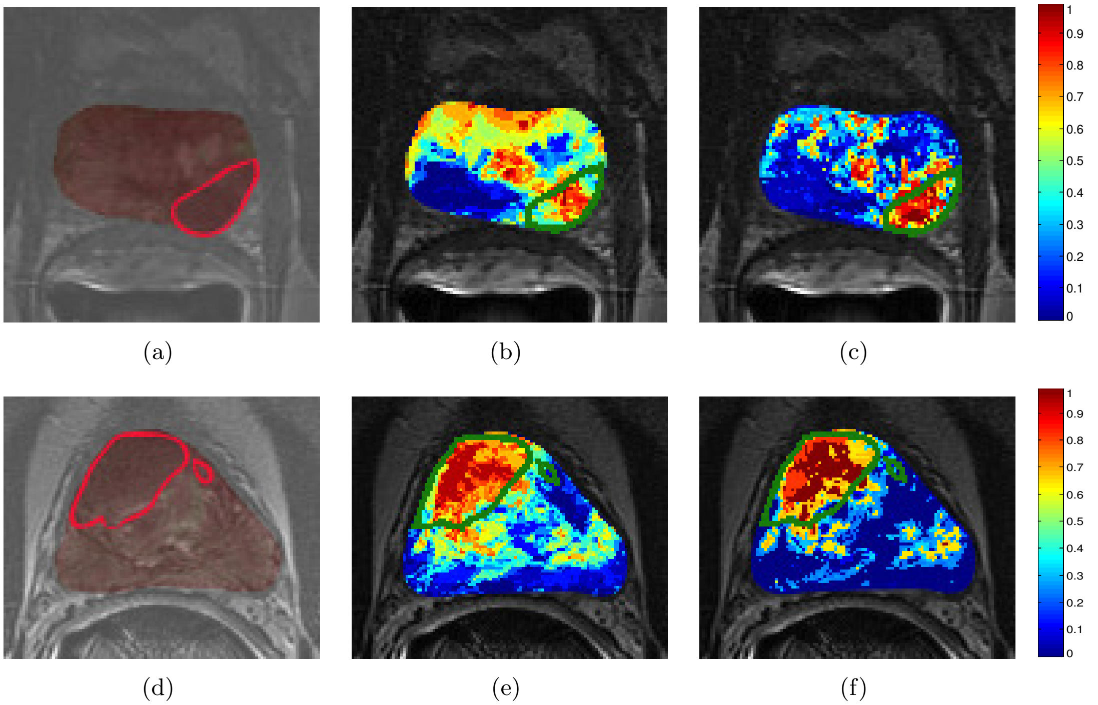
Figure 7.4: Average ROC curves across 39 leave-one-out cross validation runs. Different colored ROC curves correspond to different classifiers. The best performing classifier was $\mathbf{h}^{Em}_{MRF}(c)$, shown in light blue.

Comparison of EMPrAvISE against multi-modal classifier strategies

In this experiment, we compared the performance of $\mathbf{h}^{Em}$ with $\mathbf{h}^{Feats}$. Qualitative comparisons of the probability heatmaps so obtained are shown in Figure 7.5 (where red corresponds to a higher probability of CaP presence and blue corresponds to lower CaP probabilities). The ground truth spatial extent of CaP obtained by mapping disease extent from WMH onto MR imaging is outlined in red on Figures 7.5(a) and (d). It can be seen that $\mathbf{h}^{Em}$ (Figures 7.5(c) and (f)) demonstrates significantly more accurate and

Classifier	AUC	Accuracy
$\mathbf{h}^{T2}$	0.62±0.22	0.58±0.19
$\mathbf{h}^{T1}$	0.62±0.14	0.61±0.12
$\mathbf{h}^{ADC}$	0.65±0.21	0.62±0.19
$\mathbf{h}^{Feats}$	0.67±0.21	0.63±0.19
$\mathbf{h}^{Em}$	0.73±0.13	0.70±0.10
$\mathbf{h}^{Em}_{MRF}$ ($\mathbf{h}^{Em}$ + MRF)	**0.77±0.16**	**0.76±0.12**

Table 7.2: Summary of average and standard deviation of AUC and accuracy values for different classifiers averaged over the 39 leave-one-out cross-validation runs, for the different classifier strategies in Table 7.1.



(a) (b) (c)

(d) (e) (f)

Figure 7.5: Representative results are shown for 2D slices from 2 different studies (on each row). (a), (d) CaP extent outline (in red) delineated on WMHS-T2w MRI overlay (via MACMI). Probability heatmaps are shown for (b), (e) $\mathbf{h}^{Feats}$, and (c), (f) $\mathbf{h}^{Em}$. On each probability heatmap, red corresponds to a higher probability of CaP presence, and the mapped CaP extent (from WMHS) is delineated in green. Note that $EMPrAvISE$ ((c), (f)) is far more accurate, with significantly fewer false positives and false negatives compared to either of (b), (e).

specific predictions of CaP presence compared to $\mathbf{h}^{Feats}$ (Figures 7.5(b) and (e)). This is also reflected in the quantitative evaluation, with $\mathbf{h}^{Em}$ resulting in an AUC of 0.73 (purple curve, Figure 7.4) compared to an AUC of 0.67 for $\mathbf{h}^{Feats}$ (black curve, Figure 7.4). Additionally, we see that classification based on multi-parametric integration ($\mathbf{F}^{Feats}$, $\mathbf{F}^{Em}$) outperforms classification based on the individual protocols ($\mathbf{F}^{T1}$, $\mathbf{F}^{T2}$, $\mathbf{F}^{ADC}$). Our quantitative results corroborate findings in the clinical literature which suggest that the combination of multiple imaging protocols yield superior diagnostic accuracy compared to any single protocol [18,154,155].

Markov Random Fields in conjunction with EMPrAvISE

Figure 7.6 illustrates results of applying MRFs to the probability heatmaps obtained via $EMPrAvISE$ ($\mathbf{h}^{Em}$) to yield $\mathbf{h}^{Em}_{MRF}$. At the operating point of the ROC curve,

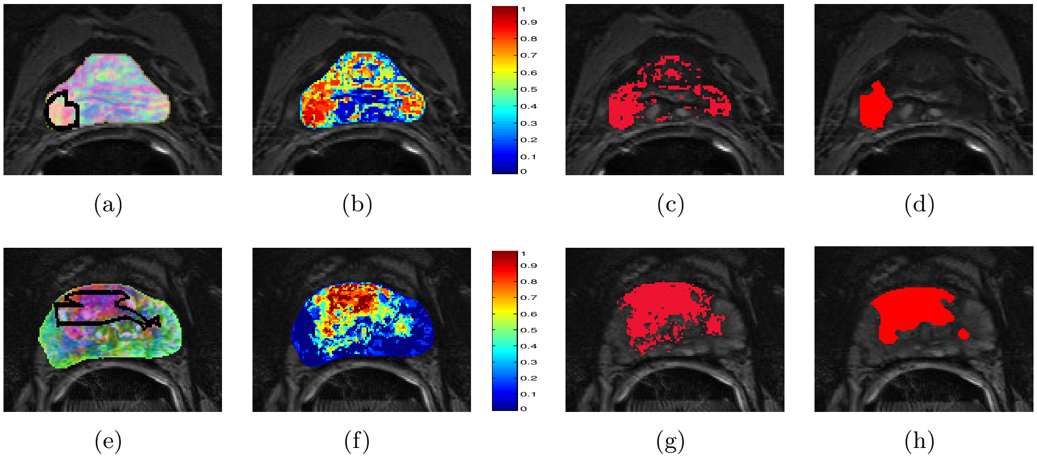


Figure 7.6: (a), (e) RGB representation of the consensus embedding (calculated via *EMPrAvISE*) with the CaP ground truth region superposed in black (obtained via registration with corresponding WMHS). (b), (f) Probability heatmap for $\mathbf{h}^{Em}$, where red corresponds to a higher probability for presence of CaP. Note the significantly higher accuracy and specificity of CaP segmentation results via application of MRFs in (d), (h) $\Psi^{Em}_{MRF,\vartheta}(C)$ compared to (c), (g) $\Psi^{Em}_{\vartheta}(C)$ (obtained by thresholding the heatmaps in (b), (f) at the operating point threshold ϑ).

$\Psi^{Em}_{\vartheta}(C)$ can be seen to have a number of extraneous regions (Figures 7.6(c) and (g)). In contrast, $\Psi^{Em}_{MRF,\vartheta}(C)$ results in a more accurate and specific CaP detection result (Figures 7.6(d) and (h)). Also shown are RGB colormap representations based on scaling the values in $\widetilde{e}_1(c), \widetilde{e}_2(c), \widetilde{e}_3(c)$ (from $\mathbf{F}^{Em}(c)$) into the RGB colorspace (Figures 7.6(a), (e)). Similarly colored regions are those that are similar in the consensus embedding space $\widetilde{\mathbb{R}}^n$. Note relatively uniform coloring within ground truth CaP areas in Figures 7.6(a) and (e), suggesting that *EMPrAvISE* is able to accurately represent the data in a reduced dimensional space while preserving disease-pertinent information.

	$\mathbf{h}^{T2}/\mathbf{h}^{Em}_{MRF}$	$\mathbf{h}^{T1}/\mathbf{h}^{Em}_{MRF}$	$\mathbf{h}^{ADC}/\mathbf{h}^{Em}_{MRF}$	$\mathbf{h}^{Feats}/\mathbf{h}^{Em}_{MRF}$	$\mathbf{h}^{Em}/\mathbf{h}^{Em}_{MRF}$
AUC	2.15e-07	1.40e-05	1.33e-04	5.86e-06	2.43e-04
Accuracy	9.64e-08	3.16e-08	1.89e-05	3.32e-05	3.32e-05

Table 7.3: p values for a paired Students t-test comparing the improvement in CaP detection performance (in terms of AUC and accuracy) of $\mathbf{h}^{Em}_{MRF}$ over $\mathbf{h}^{T2}, \mathbf{h}^{T1}, \mathbf{h}^{ADC}, \mathbf{h}^{Feats}$, and $\mathbf{h}^{Em}$ respectively. Improvements in accuracy and AUC for $\mathbf{h}^{Em}_{MRF}$ were found to be statistically significantly better ($p < 0.01$) compared to each of $\mathbf{h}^{T2}, \mathbf{h}^{T1}, \mathbf{h}^{ADC}, \mathbf{h}^{Feats}$, and $\mathbf{h}^{Em}$ respectively; the null hypothesis being that no improvement was seen via $\mathbf{h}^{Em}_{MRF}$ in each comparison.

The ROC curves in Figure 7.4 further demonstrate the improvements in CaP detection accuracy via $\mathbf{h}^{Em}_{MRF}$ (light blue curve, AUC = 0.77). These improvements in AUC and classification accuracy were found to be statistically significant ($p < 0.01$) in a paired two-tailed Students' t-test across the 39 leave-one-out cross-validation runs (Table 7.3), with the null hypothesis being that no improvement was offered by $\mathbf{h}^{Em}_{MRF}$.

7.4 Discussion

We have presented application of consensus embedding within a novel multi-parametric data representation and integration framework termed Enhanced Multi-Protocol Analysis via Intelligent Supervised Embedding or EMPrAvISE. EMPrAvISE makes use of dimensionality reduction and a supervised ensemble of embeddings to (1) accurately capture the maximum available class information from the data, and (2) account for differing dimensionalities and scales in the data. The spirit behind using an ensemble of embeddings is to exploit the variance among multiple uncorrelated embeddings in a manner similar to ensemble classifier schemes. We have demonstrated the application of EMPrAvISE to the detection of prostate cancer on (a) 4 T *ex vivo* T2w MRI, and (b) 3 Tesla *in vivo* multi-parametric (T2w, DCE, DWI) MRI. The low-dimensional data representation via EMPrAvISE was found to be superior for classification as compared to (1) the individual protocols, and (2) concatenation of multi-parametric features, and (3) simple DR of a high-dimensional feature space. We made use of a probabilistic pairwise Markov Random Field algorithm to complement the result of EMPrAvISE (AUC = 0.77) via the incorporation of spatial constraints. Sources of error within our study may exist due to (1) approximate calculation of slice correspondences between MRI and WMHS, and (2) registration-induced errors in the mapping of ground truth CaP extent from WMHS onto MRI. Therefore, our results could prove more (or less) accurate than reported, based on the margin of error in these 2 methods. However, we also note that there is currently no exact, error-free method to determine the ground truth CaP extent on MRI. Future work will hence focus on validation of our approach on a larger cohort of data. We also intend to explore the application of both EMPrAvISE and consensus embedding in the context of other domains.

Chapter 8

Concluding Remarks

In this thesis, we have presented a suite of novel computerized techniques for quantitative examination of prostate cancer presence and extent *in vivo*, based on MR imaging data. Specific goals accomplished include, (a) accurate quantification of disease presence based on MR appearance, with further stratification of such information based on spatial location of cancer within the gland, (b) comprehensive evaluation of a variety of different classifier techniques to determine the most optimal classifier for CaP detection on MRI, (c) theoretic and algorithmic development of a novel representation technique known as *consensus embedding*, and (d) application of consensus embedding to detection of CaP presence and extent at high resolution (per-voxel) on MR imaging data. We now summarize the major findings structured in terms of each of the specific aims considered in this work.

This work represents the first attempt at quantitatively defining quantitative imaging signatures (QISes) for CG and PZ tumors on T2w MRI. Our results showed that the QISes for each of CG and PZ CaP comprised largely non-overlapping textural attributes. Further, for the 2 class zone-wise CaP detection problem considered in by us, the quadratic discriminant analysis (QDA) classifier appears to offer an optimal trade-off between accuracy, training and testing time, and model complexity. We achieved significantly high classification accuracies, which are comparable and in many cases better than detection rates reported by state-of-the-art CaP detection schemes for prostate T2w MRI. The relatively high accuracy associated with the zone-specific QISes in detecting CaP imply that we have largely optimized the features and classifier for CaP detection on T2w MRI alone. These results also appear to confirm current clinical intuition, which suggests that CG and PZ CaP may have inherently differing appearances on MRI.

In this work, we have presented a novel dimensionality reduction scheme called *consensus embedding* which can be used in conjunction with a variety of DR methods for a wide range of high-dimensional biomedical data classification and segmentation problems. Consensus embedding exploits the variance within multiple base embeddings and combines them to produce a single stable solution that is superior to any of the individual embeddings, from a classification perspective. Results of quantitative and qualitative evaluation in over 200 experiments on toy, synthetic, and clinical images in terms of detection and classification accuracy demonstrated that consensus embedding shows significant improvements compared to traditional DR methods such as PCA and GE. We have presented a specific application of consensus embedding within a novel multi-parametric data representation and integration framework termed Enhanced Multi-Protocol Analysis via Intelligent Supervised Embedding or EMPrAvISE. The low-dimensional data representation via EMPrAvISE was found to be superior for representation, fusion, and classification as compared to alternative state-of-the-art schemes.

Chapter 9

Future Work

Our results present several avenues for future work. When examining quantitative signatures for CaP on MRI, we did not consider textural differences which may exist between different Gleason grades of CaP. Additional data collection may allow for exploring such differences in more detail. The findings of our classifier comparison study are intended to generalize to other medical imaging CAD problems. Indeed, the area of medical imaging is a plentiful source of similar problems concerning limited datasets with erroneous class labels; the comparison of different classifiers and machine learning methods is an area of great potential for research.

While consensus embedding has been demonstrated to have wide application in imaging and non-imaging domains, many additional extensions of classifier theory may be applied within the area of dimensionality reduction. These can include areas such as active learning or semi-supervised learning. Ensuring computational feasibility when constructing an embedding, without falling prey to the out-of-sample extension problem, is another area that requires additional research.

Currently, the topical nature of the prostate cancer problem has led to growing awareness of the requirement for better screening, diagnostic, prognostic, and treatment procedures for this disease. A significant need hence exists for tools to stage and visualize prostate cancer early using non-invasive MR imaging data. In this context, the methods and results presented in this thesis have significant translational and commercial implications for patient care. Continuing development of quantitative signatures of qualitative image features may hence form a significant precursor to developing personalized care procedures for prostate cancer.

Chapter 10

Appendices

Appendix A: Glossary of notation, symbols, and abbreviations commonly used in this thesis

CaP	Prostate cancer
MRI	Magnetic Resonance Imaging
T2w	T2-weighted
DCE	Dynamic Contrast Enhanced
DWI	Diffusion-weighted Imaging
ADC	Apparent Diffusion Coefficient
GE	Graph Embedding
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
LLE	Locally Linear Embedding
ISOMAP	Isometric Mapping
DR	Dimensionality Reduction
MDS	Multidimensional Scaling
CG	Central Gland
PZ	Peripheral Zone
SI	Signal Intensity
CAD	Computer-Aided Diagnosis
COD	Combination of Data
COI	Combination of Interpretations
SVM	Support Vector Machines
PBT	Probabilistic Boosting Trees

SSS	Small Sample Size
QCL	Questionable Class Labels
mRMR	Minimum Redundancy Maximum Relevance

c	Samples in set C
$\mathbf{F}(c) \in \mathbb{R}^N$	N-dimensional feature vector (superscript to differentiate)
$\mathcal{C}$	Image scene (superscript to differentiate T1, T2, ADC, histology)
$\mathcal{F}$	Set of all feature vectors
$l(c)$	Class label of sample c
ω_{+1}, ω_{-1}	Classes associated with $l(c) = 1, l(c) = 0$
$\mathbf{h}$	Classifier (superscript to differentiate)
$\mathbf{h}^{Bag}$	Bagged classifier
$\mathbf{h}^{Boost}$	Boosted classifier
Ψ	Detection/classification result
$G(C)$	Samples labeled with $l(c) = 1$ in C (ground truth)
$\mathbf{X}(c) \in \mathbb{R}^n$	n-dimensional embedding vector
$\widetilde{\mathbf{X}}(c) \in \widetilde{\mathbb{R}}^n$	n-dimensional consensus embedding vector
W	Confusion matrix
$\mathbf{Q}^{CG}/\mathbf{F}^{CG}$	CG-specific feature set
$\mathbf{Q}^{PZ}/\mathbf{F}^{PZ}$	PZ-specific feature set

Appendix B: Relationship between classifier accuracy and embedding strength

While embedding strength may be seen as a generalized concept for evaluating embeddings, in this work we have examined applications of DR and consensus embedding to classifying biomedical data. We now derive a direct relationship between embedding strength and classification accuracy, presented in Theorem 1 below.

For the purposes of the following discussion, all objects $c, d, e \in C$ are considered to be associated with class labels $l(c), l(d), l(e) \in \{\omega_1, \omega_2\}$, respectively, such that if $l(c) = l(d) = \omega_1$ and $l(e) = \omega_2$ then $\Lambda^{cd} < \Lambda^{ce}$ and $\Lambda^{cd} < \Lambda^{de}$. Note that ω_1, ω_2 are binary class labels that can be assigned to all objects $c \in C$.

Definition 6 *An unique triplet $(c, d, e) \in C$ with $l(c), l(d), l(e) \in \{\omega_1, \omega_2\}$ is called a class triplet, $(c, d, e)_l$, if either $l(c) \neq l(d)$, or $l(d) \neq l(e)$, or $l(c) \neq l(e)$.*

Thus, in a *class triplet* of objects, two out of three objects have the same class label but the third has a different class label, e.g. if $l(c) = l(d) = \omega_1$ and $l(e) = \omega_2$. Further, in the specific case that $\Delta(c, d, e) = 1$ for a class triplet $(c, d, e)_l$, it will be denoted as $\Delta^l(c, d, e)$. For the above example of a class triplet, we know that $\Lambda^{cd} < \Lambda^{ce}$ and $\Lambda^{cd} < \Lambda^{de}$ (see above). If $\Delta(c, d, e) = 1$, $\delta^{cd} < \delta^{ce}$ and $\delta^{cd} < \delta^{de}$. This implies that even after projection from $\mathbb{R}^N$ to $\mathbb{R}^n$, the class-based pairwise relationships within the data are accurately preserved (a classifier can be constructed which will correctly classify objects c, d, e in $\mathbb{R}^n$).

Consider that if $\frac{R}{S}$ objects have class label ω_1, then $\frac{(S-1)R}{S}$ objects will have class label ω_2. Based on the total number of unique triplets Z, the total number of triplets which are not class triplets is,

$$Y = \frac{\frac{R}{S}!}{3!(\frac{R}{S} - 3)!} + \frac{\frac{(S-1)R}{S}!}{3!(\frac{(S-1)R}{S} - 3)!} \tag{10.1}$$

Y will be a constant for a given set of objects C, and is based on forming unique triplets (c, d, e) where $l(c) = l(d) = l(e)$ (triplets which are not class triplets). $U = (Z - Y)$ will correspond to the number of class triplets that may be formed for set C. If all U class triplets have $\Delta^l(c, d, e) = 1$, then it is possible to construct U classifiers which correctly classify the corresponding objects in these class triplets.

Definition 7 *Given U unique class triplets $(c, d, e)_l \in C$ and an embedding $\mathbb{R}^n$ of all objects $c, d, e \in C$, the associated classification accuracy $\phi^{Acc}(\mathbb{R}^n) = \frac{\sum_C \Delta^l(c,d,e)}{U}$*

As illustrated previously, class triplets $(c, d, e)_l$ for which $\Delta^l(c, d, e) = 1$ will correspond to those objects which will be classified correctly in $\mathbb{R}^n$. Therefore, the classification accuracy $\phi^{Acc}(\mathbb{R}^n)$ may simply be defined as the fraction of class triplets $(c, d, e)_l \in C$ for which $\Delta^l(c, d, e) = 1$.

Theorem 1 *For any $\mathbb{R}^n$, the corresponding $\psi^{ES}(\mathbb{R}^n)$ increases monotonically as a function of $\phi^{Acc}(\mathbb{R}^n)$.*

Proof.

$$\begin{aligned} \text{By definition, } \sum_C \Delta(c, d, e) &\geq \sum_C \Delta^l(c, d, e) \\ \text{Dividing by } Z = U + Y \text{ on either side, } \psi^{ES}(\mathbb{R}^n) &\geq \frac{\sum_C \Delta^l(c, d, e)}{U + Y} \\ \text{Inverting, } \frac{1}{\psi^{ES}(\mathbb{R}^n)} &\leq \frac{1}{\phi^{Acc}(\mathbb{R}^n)} + \frac{Y}{\sum_C \Delta^l(c, d, e)} \end{aligned}$$

As $\frac{Y}{\sum_C \Delta^l(c,d,e)}$ is a constant, $\psi^{ES}(\mathbb{R}^n)$ increases monotonically with $\phi^{Acc}(\mathbb{R}^n)$. □

Thus an embedding $\mathbb{R}^n$ with a high embedding strength will have a high classification accuracy. Practically, this implies that $\psi^{ES}(\mathbb{R}^n)$ may be estimated via any measure of object-class discrimination such as classification accuracy or cluster-validity measures. We have exploited this relationship in our algorithmic implementation.

Appendix C: Review of classifier ensemble schemes used in this thesis

In the following sections, we denote a set of samples $c \in C$ where $|C|$ is the cardinality of any set C. Every sample $c \in C$ is associated with a N-dimensional feature vector denoted $\mathbf{F}(c) = \{f_1(c), f_2(c), \ldots, f_N(c)\}$. We define the set of all feature vectors as $\mathcal{F} = \{\mathbf{F}(c_1); \ldots; \mathbf{F}(c_n)\}$ $(n = |C|)$. Every sample $c \in C$ is also associated with a class label $l(c) \in \{0, 1\}$. For ease of notation, all $c \in C$ with $l(c) = 1$ will be considered to belong to class ω_{+1} (target), while those with $l(c) = 0$ belong to class ω_{-1} (non-target). We denote a classifier as $\mathbf{h}^{\varphi}$ (where the superscript φ is used to differentiate between the algorithms considered in this work). Note that while QDA, naïve Bayes, SVMs, and DTs are considered single classification strategies (or “base” classifiers), bagging and boosting are considered ensemble classifiers (as they combine multiple base classifiers to obtain a single result). Appendix A may be referenced for a summary of notation and symbols appearing herein.

Quadratic Discriminant Analysis (QDA)

While DA is typically considered to have low model complexity and require minimal training within classifier comparison studies (due to simplicity), the use of a quadratic function in conjunction with DA (as opposed to a linear discriminator) ensures robustness while not compromising on the overall speed of such methods [145]. QDA has previously been used with considerable success in discriminating classes for problems related to bioinformatics [156] as well as medical imaging data [82, 157].

The QDA classifier [145] aims to find a transformation of the input features that results in the best possible discrimination between the classes in the dataset. Given the set of samples C with associated feature set $\mathcal{F}$, QDA solves for

$$\mathcal{Y} = \mathcal{F}^{\mathrm{T}} A \mathcal{F} + B^{\mathrm{T}} \mathcal{F}, \tag{10.2}$$

where $\mathcal{Y} = \{\mathbf{Y}(c_1); \ldots; \mathbf{Y}(c_n)\}$ is the result of QDA, and A, B parametrize the transformation. Note that $\mathbf{Y}(c_1)$ is the transformed vector corresponding to $\mathbf{F}(c_1)$.

Based on calculating the means $\mu_{c\in\omega_{+1}}, \mu_{c\in\omega_{-1}}$ and covariances $\Sigma_{c\in\omega_{+1}}, \Sigma_{c\in\omega_{-1}}$ of the 2 classes in the dataset, Equation 10.2 can be solved in terms of the following log likelihood ratio,

$$\log(\mathbf{H}^{QDA}) = \frac{(\mathcal{F} - \mu_{c\in\omega_{+1}})^{\mathrm{T}} \Sigma^{-1}_{c\in\omega_{+1}} (\mathcal{F} - \mu_{c\in\omega_{+1}})}{(\mathcal{F} - \mu_{c\in\omega_{-1}})^{\mathrm{T}} \Sigma^{-1}_{c\in\omega_{-1}} (\mathcal{F} - \mu_{c\in\omega_{-1}})}. \tag{10.3}$$

The result of QDA classification is a per-sample probability of belonging to the target class ω_{+1}, given by $\mathbf{h}^{QDA}(c) \in \mathbf{H}^{QDA}$ (bounded in [0,1]).

Naïve Bayes

Bayesian learning is a method of statistical inference in which evidence or observations are used to update the probability that a hypothesis may be true [116]. In spite of its simplicity and lack of model parameters, the naïve Bayes classifier has been shown be extremely accurate and robust [69,158,159] when (a) sufficient training data is available, and (b) independence assumptions of its constituent features have been satisfied.

The naïve form of Bayes' rule for a binary class problem defines the likelihood of observing class ω_{+1} given the multivariate feature vector $\mathbf{F}(c)$ as,

$$P(\omega_{+1}|\mathbf{F}(c)) = \frac{P(\omega_{+1})p(\mathbf{F}(c)|\omega_{+1})}{P(\omega_{+1})p(\mathbf{F}(c)|\omega_{+1}) + P(\omega_{-1})p(\mathbf{F}(c)|\omega_{-1})}, \tag{10.4}$$

where the prior probabilities of occurrence of the two classes are $P(\omega_{+1})$ and $P(\omega_{-1})$, and $p(\mathbf{F}(c)|\omega_{+1})$ and $p(\mathbf{F}(c)|\omega_{-1})$ represent the *a priori* class conditional distributions of $\mathbf{F}(c)$. Estimation of $p(\mathbf{F}(c)|\omega_{+1})$ and $p(\mathbf{F}(c)|\omega_{-1})$ is difficult when $\mathbf{F}(c)$ is of high dimensionality. Dimensionality reduction methods such as independent component analysis [160] (ICA) allow for reduction of the number of components of the feature vector $\mathbf{F}$ (of dimensionality N) to a vector $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_M\}$ (of dimensionality $M \ll N$). This is done by calculating a linear transformation $\boldsymbol{\phi} = W\mathbf{F}$ such that each dimension of the resulting $\boldsymbol{\phi}$ is statistically independent. The independence between the dimensions means that the corresponding class conditional probability can be written as a product of the component probabilities for each dimension. Thus for any sample $c \in C$,

$$p(\boldsymbol{\phi}(c)|\omega_{+1}) = p(\phi_1(c)|\omega_{+1}) \ldots p(\phi_M(c)|\omega_{+1})) = \prod_{i}^{M} p(\phi_i(c)|\omega_{+1}). \tag{10.5}$$

The M independent components $\phi_i(c)$ of pixels $c \in \omega_{+1}$ are used to generate M *a priori* distributions $p(\phi_i(c)|\omega_{+1})$ for the target class ω_{+1}, while pixels $c \in \omega_{-1}$ are used to generate distributions for class ω_{-1}. The *a posteriori* probability distribution $P(\omega_{+1}|\boldsymbol{\phi}(c))$ of observing class ω_{+1} for the linearly independent feature vector $\boldsymbol{\phi}(c)$ at each pixel $c \in C$ can now be expressed as,

$$P(\omega_{+1}|\boldsymbol{\phi}(c)) = \frac{P(\omega_{+1})\prod_i^M p(\phi_i(c)|\omega_{+1})}{P(\omega_{+1})\prod_i^M p(\phi_i(c)|\omega_{+1}) + P(\omega_{-1})\prod_i^M p(\phi_i(c)|\omega_{-1})}. \tag{10.6}$$

The probabilistic result of naïve Bayes' classification is denoted $\mathbf{h}^{Bay}(c) = P(\omega_{+1}|\boldsymbol{\phi}(c)) \in [0,1]$.

Support Vector Machines (SVMs)

The SVM algorithm [117] is one of the most popular classifier schemes for medical imaging and CAD applications [32,34,52,83–86]. This classifier makes use of the "kernel trick" [117] wherein data is projected into higher dimensions within which a hyperplane can be found to separate the classes. While SVMs are powerful, non-linear classifiers, they require significant training and tend to be computationally expensive [161].

Instead of minimizing an objective function based on the training samples, SVMs focus on the training examples that are most difficult to classify. These "borderline" training examples are called *support vectors*. The general form of the decision function of the SVM classifier for a new test sample d is given by,

$$V(d) = \sum_{u=1}^{n} \xi_u l(c_u)\mathbf{K}(\mathbf{F}(d), \mathbf{F}(c_u)) + b, \tag{10.7}$$

where $\mathbf{F}(c_u)$, $u \in \{1, 2, ..., n\}$, denote the support vectors, ξ_u correspond to weights (model parameters) for $\mathbf{F}(c_u)$, $\mathbf{K}(\cdot,\cdot)$ is a positive definite symmetric function called the kernel function, and b is the bias. Both bias and model parameters are estimated on training data. The kernel function $\mathbf{K}(\cdot,\cdot)$ defines the nature of the decision surface. In our implementation, the popular radial basis function was used, $\mathbf{K}(\mathbf{F}(c), \mathbf{F}(d)) = e^{(-\lambda||\mathbf{F}(c)-\mathbf{F}(d)||^2)}$, where λ is a real number that normalizes the inputs.

The decision function $V(c)$ resulting from Equation 10.7 defines a hard SVM classification result where,

$$\mathbf{h}^{SVM}(c) = \begin{cases} 1 & \text{when } V(c) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10.8}$$

Decision Trees (DTs)

Decision trees are a popular algorithm for machine learning [73–75, 118] in that they provide an easily interpretable model of the data. They are generally robust and efficient, though their generalizability reduces as additional features are included within the model. More and more, DTs have begun to be used in conjunction with ensemble algorithms [73–75], rather than as a stand-alone classifier as this tends to improve their stability and generalizability.

The most popular DT is the C4.5 classifier [118]. The rules generated by this approach are in conjunctive form such as "if X and Y then Z" where both X and Y are the rule antecedents, while Z is the rule consequence. During training, the tree rules are generated using an iterative selection of individual features that are the most salient at each node in the tree. Every path from the root to the leaf is converted to an initial rule by regarding all the conditions appearing in the path as the conjunctive rule antecedents while regarding the class label held by the leaf as a rule consequence. Tree pruning is then done by using both greedy elimination and minimum description length rules [118], which remove antecedents that are not sufficiently discriminatory. The final result of the DT algorithm is a hard classification $\mathbf{h}^{DT}(c) \in \{0, 1\}$.

Bagging

Bootstrap **Agg**regation (Bagging) was first proposed by Brieman [38] in order to overcome issues with the stability of classification procedures (which can cause erroneous results). Since its introduction, bagging has offered a relatively simple way of improving classifier model stability without requiring significant additional computation time [71,73,156]. The relatively naïve averaging approach adopted by bagging yields an optimal performance when its component classifiers satisfy certain assumptions about bias and variance.

The bagging algorithm consists of first generating multiple subsets of the original data set via bootstrap sampling (with replication) i.e. for each trial $t \in \{1, 2, .., T\}$, a training subset $S_t \subset C$ is sampled with replacement. Based on each bootstrapped training set S_t, a component classification result $h_t^{\beta}(c)$, $\forall t \in \{1, 2, .., T\}$, is generated for all samples $c \in (C - S_t)$, where $\beta \in \{QDA, Bay, SVM, DT\}$, corresponds to the learning algorithm used to construct the classifier. The bagged classification result $\mathbf{h}^{Bag,\beta}(c)$ is typically obtained as the majority vote of the classification decisions across the T component learners $h_t^{\beta}(c)$. Bagging improves classification accuracy by exploiting the variance in the component learners [116], and will thus improve in classification accuracy only if perturbing the training sets can cause significant changes in the corresponding predictions. The bagged classification result is given as,

$$\mathbf{h}^{Bag,\beta}(c) = \frac{1}{T}\sum_{t=1}^{T} h_t^{\beta}(c) \in [0, 1], \beta \in \{QDA, Bay, SVM, DT\}. \qquad (10.9)$$

Boosting

Boosted ensembles [34, 71, 73] refer to a category of classifiers that iteratively weight misclassified instances with the objective of ensuring they are classified correctly [39]. While boosting adds significantly more computational complexity compared to bagging and can be considered to be susceptible to outliers [75], its targeted approach ensures that it reduces over-fitting significantly [40]. This has led to the development of different variants of the boosting algorithm [39, 70, 151, 162] to address these problems. In this study, two popular variants of the boosting strategy are considered – Adaboost [39] and Probabilistic Boosting Trees [151].

Adaboost

Freund and Schapire proposed an **ada**ptive **boost**ing algorithm (AdaBoost) [39], wherein the goal was to significantly reduce the learning error of any algorithm (called a "component learner") whose performance is a little better than random guessing. Unlike bagging [38], boosting maintains a weight for each training instance – the higher the weight, the more the instance influences the learned classifier at each iteration.

In the following description, $w_{t,c}$ denotes the weight of training instance $c \in R_t$ at trial t, $t \in \{1, 2, .., T\}$. In essence, this training set R_t comprises n samples, but each sample $c \in R_t$ is described by $Z < N$ features i.e. the feature space is sub-sampled to construct a training subset. Initially for every $c \in R_t$, $w_{1,c}{=}\frac{1}{|R_1|}$. At each trial $t \in \{1, 2, .., T\}$, a component classifier $h_t^\alpha(c)$, $\alpha \in \{QDA, Bay, SVM\}$, is constructed from the given instances $c \in R_t$ under the distribution $w_{t,c}$. The error ϵ_t of this classifier is also measured with respect to the weights and is the sum of the weights of the training instances that are misclassified by $h_t^\alpha(c)$. At each iteration, the weights of samples in the training dataset are changed based on the performance of the classifiers that have been generated thus far. Thus, the weight vector for the following trial $(t + 1)$ is generated by amplifying $w_{t+1,c}$ for $c \in R_{t+1}$ such that $h_{t+1}^\alpha(c)$ classifies correctly by the factor $\gamma_t{=}\frac{\epsilon_t}{1-\epsilon_t}$. The weights are re-normalized so that $\sum_c w_{t+1,c}{=}1$. When $\epsilon_t \geq 0.5$, the trials are terminated. The final classifier is obtained as a weighted combination of the decisions of the individual component learners.

In this work, Adaboost was implemented in conjunction with the QDA, naïve Bayes, and SVM classifiers. The probabilistic boosted classifier $\mathbf{h}^{Boost,\alpha}$ is obtained as,

$$\mathbf{h}^{Boost,\alpha}(c) = \sum_{t=1}^{T} h_t^\alpha(c) \log(\frac{1}{\gamma_t}) \in [0, 1], \alpha \in \{QDA, Bay, SVM\}. \qquad (10.10)$$

Probabilistic Boosting Trees

The PBT algorithm [151] directly integrates Adaboost [39] into the construction of the decision tree to allow for the computation of weighted probabilistic decisions for difficult to classify samples. This is done by iteratively generating a tree structure of length L

in the training stage, where each node of the tree is boosted with T weak classifiers. During testing, the conditional probability of the object $c \in C$ is calculated at each node based on the learned hierarchical tree. A discriminative model is obtained at the top of the tree by combining the probabilities associated with propagation of the object at various nodes, yielding a posterior conditional probability of the sample belonging to ω_{+1}, $\mathbf{h}^{Boost,DT}(c) = p(\omega_{+1}|\mathbf{F}(c)) \in [0, 1]$.

References

[1] S. E. Viswanath, N. B. Bloch, J. C. Chappelow, R. Toth, N. M. Rofsky, E. M. Genega, R. E. Lenkinski, and A. Madabhushi. Central gland and peripheral zone prostate tumors have significantly different quantitative imaging signatures on 3 tesla endorectal, in vivo t2-weighted mr imagery. *J Magn Reson Imaging*, page Accepted, 2012.

[2] S. Viswanath and A. Madabhushi. Consensus embedding: theory, algorithms and application to segmentation and classification of biomedical data. *BMC Bioinformatics*, 13(1):26, 2012.

[3] Satish Viswanath, B. Nicolas Bloch, Jonathan Chappelow, Neil Rofsky, Elizabeth Genega, Robert Lenkinski, , and Anant Madabhushi. Comparing Classifier Ensembles for Discriminating Central Gland and Peripheral Zone Tumors on In Vivo 3 Tesla T2-weighted Prostate MRI. *Med Phys*, page Under review, 2012.

[4] S. Viswanath and A. Madabhushi. An Intelligent Data Embedding Aggregation Scheme for High Resolution Prostate Cancer Detection via Multi-Parametric MRI. *IEEE Trans. Med. Imag.*, page Under review, 2012.

[5] Satish Viswanath, Mark Rosen, and Anant Madabhushi. A consensus embedding approach for segmentation of high resolution in vivo prostate magnetic resonance imagery. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 6915, page 69150U. SPIE, 2008.

[6] Satish Viswanath, B.N. Bloch, M. Rosen, J. Chappelow, R. Toth, N. Rofsky, R. E. Lenkinski, E Genega, A. Kalyanpur, and A. Madabhushi. Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol in vivo 3 Tesla MRI. In *SPIE Medical Imaging : Computer-Aided Diagnosis*, volume 7260, page 72603I. SPIE, 2009.

[7] S Viswanath, B. N. Bloch, J Chappelow, P Patel, N. Rofsky, R. Lenkinski, E. Genega, and A Madabhushi. Enhanced multi-protocol analysis via intelligent supervised embedding (EMPrAvISE): detecting prostate cancer on multi-parametric MRI. In *SPIE Medical Imaging*, volume 7963, page 79630U. SPIE, 2011.

[8] A.L.N. Fred and A.K. Jain. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):835–850, 2005.

[9] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[10] R.K.S. Kwan, A.C. Evans, and G.B. Pike. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Trans. Med. Imag.*, 18(11):1085–1097, 1999.

[11] Jonathan Chappelow, B. Nicolas Bloch, Neil Rofsky, Elizabeth Genega, Robert Lenkinski, William DeWolf, and Anant Madabhushi. Elastic registration of multimodal prostate MRI and histology via multiattribute combined mutual information. *Medical Physics*, 38(4):2005–2018, 2011.

[12] M. J. Chelsky, M. D. Schnall, E. J. Seidmon, and H. M. Pollack. Use of endorectal surface coil magnetic resonance imaging for local staging of prostate cancer. *J Urol*, 150(2 Pt 1):391–5, 1993.

[13] G. J. Jager, E. T. Ruijter, C. A. van de Kaa, J. J. de la Rosette, G. O. Oosterhof, J. R. Thornbury, and J. O. Barentsz. Local staging of prostate cancer with endorectal MR imaging: correlation with histopathology. *AJR Am J Roentgenol*, 166(4):845–52, 1996.

[14] M. Bezzi, H. Y. Kressel, K. S. Allen, M. L. Schiebler, H. G. Altman, A. J. Wein, and H. M. Pollack. Prostatic carcinoma: staging with MR imaging at 1.5 T. *Radiology*, 169(2):339–46, 1988.

[15] M. R. Engelbrecht, G. J. Jager, R. J. Laheij, A. L. Verbeek, H. J. van Lier, and J. O. Barentsz. Local staging of prostate cancer using magnetic resonance imaging: a meta-analysis. *European Radiology*, 12(9):2294–302, 2002.

[16] M. L. Schiebler, M. D. Schnall, H. M. Pollack, R. E. Lenkinski, J. E. Tomaszewski, A. J. Wein, R. Whittington, W. Rauschning, and H. Y. Kressel. Current role of MR imaging in the staging of adenocarcinoma of the prostate. *Radiology*, 189(2):339–352, 1993.

[17] K. K. Yu and H. Hricak. Imaging prostate cancer. *Radiol Clin North Am*, 38(1):59–85, viii, 2000.

[18] K. Kitajima, Y. Kaji, Y. Fukabori, K. I. Yoshida, N. Suganuma, and K. Sugimura. Prostate cancer detection with 3 T MRI: Comparison of diffusion-weighted imaging and dynamic contrast-enhanced MRI in combination with T2-w imaging. *J Magn Reson Imaging*, 31(3):625–631, 2010.

[19] A. Maio and M. D. Rifkin. Magnetic resonance imaging of prostate cancer: update. *Top Magn Reson Imaging*, 7(1):54–68, 1995.

[20] J. E. McNeal. Regional morphology and pathology of the prostate. *Am J Clin Pathol*, 49(3):347–57, 1968.

[21] M. L. Schiebler, J. E. Tomaszewski, M. Bezzi, H. M. Pollack, H. Y. Kressel, E. K. Cohen, H. G. Altman, W. B. Gefter, A. J. Wein, and L. Axel. Prostatic carcinoma and benign prostatic hyperplasia: correlation of high-resolution MR and histopathologic findings. *Radiology*, 172(1):131–7, 1989.

[22] Jurgen J. Futterer and Jelle O. Barentsz. 3T MRI of prostate cancer. *Applied Radiology*, 39(1):25–32, 2009.

[23] O. Akin, E. Sala, C. S. Moskowitz, K. Kuroiwa, N. M. Ishill, D. Pucar, P. T. Scardino, and H. Hricak. Transition zone prostate cancers: features, detection, localization, and staging at endorectal MR imaging. *Radiology*, 239(3):784–92, 2006.

[24] B. Hamm, G. Krestin, M. Laniado, V. Nicolas, and M. Taupitz. *MR Imaging of the Abdomen and Pelvis.* Thieme, New York, 2nd edition, 2009.

[25] B. A. Shannon, J. E. McNeal, and R. J. Cohen. Transition zone carcinoma of the prostate gland: a common indolent tumour type that occasionally manifests aggressive behaviour. *Pathology*, 35(6):467–71, 2003.

[26] A. Oto, A. Kayhan, Y. Jiang, M. Tretiakova, C. Yang, T. Antic, F. Dahi, A. L. Shalhav, G. Karczmar, and W. M. Stadler. Prostate cancer: differentiation of central gland cancer from benign prostatic hyperplasia by using diffusion-weighted and dynamic contrast-enhanced MR imaging. *Radiology*, 257(3):715–23, 2010.

[27] B. Nicolas Bloch, Robert E. Lenkinski, and Neil M. Rofsky. The role of magnetic resonance imaging (MRI) in prostate cancer imaging and staging at 1.5 and 3 Tesla: The Beth Israel Deaconess Medical Center (BIDMC) approach. *Cancer Biomark*, 4(4):251–262, 2008.

[28] C. K. Kim and B. K. Park. Update of prostate magnetic resonance imaging at 3 T. *J Comput Assist Tomogr*, 32(2):163–72, 2008.

[29] J. H. Kim, J. K. Kim, B. W. Park, N. Kim, and K. S. Cho. Apparent diffusion coefficient: prostate cancer versus noncancerous tissue according to anatomical region. *J Magn Reson Imaging*, 28(5):1173–9, 2008.

[30] P. Swindle, S. Ramadan, P. Stanwell, S. McCredie, P. Russell, and C. Mountford. Proton magnetic resonance spectroscopy of the central, transition and peripheral zones of the prostate: assignments and correlation with histopathology. *MAGMA*, 21(6):423–34, 2008.

[31] Allan J. Frame, Peter E. Undrill, Michael J. Cree, John A. Olson, Kenneth C. McHardy, Peter F. Sharp, and John V. Forrester. A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms. *Comput Biol Med*, 28(3):225–238, 1998.

[32] J. Juntu, J. Sijbers, S. De Backer, J. Rajan, and D. Van Dyck. Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. *J Magn Reson Imaging*, 31(3):680–9, 2010.

[33] Y. M. Kadah, A. A. Farag, J. M. Zurada, A. M. Badawi, and A. B. M. Youssef. Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images. *IEEE Trans. Med. Imag.*, 15(4):466–478, 1996.

[34] Liyang Wei, Yongyi Yang, R. M. Nishikawa, and Yulei Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Trans. Med. Imag.*, 24(3):371–380, 2005.

[35] S. Herlidou-Meme, J. M. Constans, B. Carsin, D. Olivie, P. A. Eliat, L. Nadal-Desbarats, C. Gondry, E. Le Rumeur, I. Idy-Peretti, and J. D. de Certaines. MRI texture analysis on texture test objects, normal brain and intracranial tumors. *Magnetic Resononance Imaging*, 21(9):989–993, 2003.

[36] James P. Monaco, John E. Tomaszewski, Michael D. Feldman, Ian Hagemann, Mehdi Moradi, Parvin Mousavi, Alexander Boag, Chris Davidson, Purang Abolmaesumi, and Anant Madabhushi. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models. *Medical Image Analysis*, 14(4):617–629, 2010.

[37] Anant Madabhushi, Jianbo Shi, Michael Feldman, Mark Rosen, and John Tomaszewski. Comparing Ensembles of Learners: Detecting Prostate Cancer from High Resolution MRI. In *Proc 2nd Int'l Computer Vision Approaches to Medical Image Analysis (CVAMIA) Workshop (held in conjunction with ECCV)*, pages 25–36, 2006.

[38] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[39] Yoav Freund and Robert Schapire. Experiments with a New Boosting Algorithm. In *Proc Int'l Conf Mach Learn*, pages 148–156, 1996.

[40] T. Dietterich. Ensemble methods in machine learning. In *Proc. 1st Int'l Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[41] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imag.*, 23(7):903–921, 2004.

[42] Jr. Allsbrook, W. C., K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol*, 32(1):81–8, 2001.

[43] A. De la Taille, A. Viellefond, N. Berger, E. Boucher, M. De Fromont, A. Fondimare, V. Molinie, D. Piron, M. Sibony, F. Staroz, M. Triller, E. Peltier, N. Thiounn, and M. A. Rubin. Evaluation of the interobserver reproducibility of gleason grading of prostatic adenocarcinoma using tissue microarrays. *Hum Pathol*, 34(5):444–9, 2003.

[44] A. A. Renshaw, D. Schultz, K. Cote, M. Loffredo, D. E. Ziemba, and A. V. D'Amico. Accurate Gleason grading of prostatic adenocarcinoma in prostate needle biopsies by general pathologists. *Arch Pathol Lab Med*, 127(8):1007–8, 2003.

[45] J. Melia, R. Moseley, R. Y. Ball, D. F. R. Griffiths, K. Grigor, P. Harnden, M. Jarmulowicz, L. J. McWilliam, R. Montironi, M. Waller, S. Moss, and M. C. Parkinson. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology*, 48(6):644–654, 2006.

[46] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory*, 14(1):55–63, 1968.

[47] Gaoyu Xiao, B. Nicolas Bloch, Jonathan Chappelow, Elizabeth M. Genega, Neil M. Rofsky, Robert E. Lenkinski, John Tomaszewski, Michael D. Feldman, Mark Rosen, and Anant Madabhushi. Determining histology-MRI slice correspondences for defining MRI-based disease signatures of prostate cancer. *Comput Med Imag Graphics*, 35(7-8):568–78, 2011.

[48] G. Lee, S. Doyle, J. Monaco, A. Madabhushi, M. D. Feldman, S. R. Master, and J. E. Tomaszewski. A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: Preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology. In *Proc. ISBI*, pages 77–80, 2009.

[49] S. Viswanath, B.N. Bloch, M. Rosen, J. Chappelow, R. Toth, N. Rofsky, R. E. Lenkinski, E Genega, A. Kalyanpur, and A. Madabhushi. Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol in vivo 3 Tesla MRI. In *SPIE Medical Imaging : Computer-Aided Diagnosis*, volume 7260, page 72603I, 2009.

[50] J. Chappelow, B.N. Bloch, N. Rofsky, E. Genega, R. Lenkinski, W. DeWolf, S. Viswanath, and A. Madabhushi. COLLINARUS: Collection of Image-derived Non-linear Attributes for Registration Using Splines. In *Proc. SPIE*, volume 7259, 2009.

[51] A. Madabhushi, M. Feldman, D. Metaxas, J. Tomaszeweski, and D. Chute. Automated Detection of Prostatic Adenocarcinoma from High-Resolution Ex Vivo MRI. *IEEE Trans. Med. Imag.*, 24(12):1611–1625, 2005.

[52] P Tiwari, S Viswanath, J Kurhanewicz, A Sridhar, and A Madabhushi. Multimodal wavelet embedding representation for data combination (MaWERiC): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection. *NMR Biomed*, page Accepted, 2011.

[53] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Pac Symp Biocomput*, pages 300–11, 2004.

[54] R. Verma, E. Zacharaki, Y. Ou, H. Cai, S. Chawla, S. Lee, E. Melhem, R. Wolf, and C. Davatzikos. Multiparametric Tissue Characterization of Brain Neoplasms and Their Recurrence Using Pattern Classification of MR Images. *Academic Radiology*, 15(8):966–977, 2008.

[55] S. Viswanath, B. Bloch, E. Genega, N. Rofsky, R. Lenkinski, J. Chappelow, R. Toth, and A. Madabhushi. A Comprehensive Segmentation, Registration, and Cancer Detection Scheme on 3 Tesla In Vivo Prostate DCE-MRI. In *Proc. MICCAI*, pages 662–669, 2008.

[56] G. Lee, C. Rodriguez, and A. Madabhushi. Investigating the Efficacy of Nonlinear Dimensionality Reduction schemes in Classifying Gene- and Protein-Expression Studies. 5(3):1–17, 2008.

[57] Richard Bellman. *Adaptive control processes: a guided tour.* Princeton University Press, 1961.

[58] I.T. Jolliffe. *Principal Component Analysis.* Springer, 2002.

[59] Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):796–809, 2008.

[60] L. Saul and S. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[61] J.B. Tenenbaum, V. Silva, and J.C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[62] K. Dawson, R. Rodriguez, and W. Malyj. Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC Bioinformatics*, 6(1):195, 2005.

[63] A. Madabhushi, J. Shi, M. Rosen, J. E. Tomaszeweski, and M. D. Feldman. Graph embedding to improve supervised classification and novel class detection: application to prostate cancer. In *Proc. 8th Int'l Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 729–37, 2005.

[64] C. Varini, A. Degenhard, and T. Nattkemper. Visual exploratory analysis of DCE-MRI data in breast cancer by dimensional data reduction: A comparative study. *Biomedical Signal Processing and Control*, 1(1):56–63, 2006.

[65] J.R. Quinlan. *The effect of noise on concept learning.* Morgan Kaufmann, 1986.

[66] Mukund Balasubramanian, Eric L. Schwartz, Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. The isomap algorithm and topological stability. *Science*, 295(5552):7a–, 2002.

[67] H. Chang and D. Yeung. Robust locally linear embedding. *Pattern Recognition*, 39(6):1053–1065, 2006.

[68] C. Shao, H. Huang, and L. Zhao. P-ISOMAP: A New ISOMAP-Based Data Visualization Algorithm with Less Sensitivity to the Neighborhood Size. *Dianzi Xuebao(Acta Electronica Sinica)*, 34(8):1497–1501, 2006.

[69] Eric Bauer and Ron Kohavi. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Mach Learn*, 36(1):105–139, 1999.

[70] L. Breiman. Arcing Classifiers. *The Annals of Statistics*, 26(3):801–824, 1998.

[71] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Mach Learn*, 40(3):203–228, 2000.

[72] Q. L. Tran, K. A. Toh, D. Srinivasan, K. L. Wong, and Low Shaun Qiu-Cen. An empirical comparison of nine pattern classifiers. *IEEE Trans. Syst., Man, Cybern.*, 35(5):1079–1091, 2005.

[73] Thomas G. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach Learn*, 40(2):139–157, 2000.

[74] M. Hamza and D. Larocque. An empirical comparison of ensemble methods based on classification trees. *JSCS*, 75(8):629–643, 2005.

[75] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *JAIR*, 11(1):169–198, 1999.

[76] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[77] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *JASA*, 97(457):77–87, 2002.

[78] Tao Li, Chengliang Zhang, and Mitsunori Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinform*, 20(15):2429–2437, 2004.

[79] Jae Won Lee, Jung Bok Lee, Mira Park, and Seuck Heun Song. An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data An*, 48(4):869–885, 2005.

[80] Georges Natsoulis, Laurent El Ghaoui, Gert R.G. Lanckriet, Alexander M. Tolley, Fabrice Leroy, Shane Dunlea, Barrett P. Eynon, Cecelia I. Pearson, Stuart Tugendreich, and Kurt Jarnagin. Classification of a large microarray data set: Algorithm comparison and analysis of drug signatures. *Genome Res*, 15(5):724–736, 2005.

[81] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proc 23rd ICML*, pages 161–168, 2006.

[82] T. Schmah, G. Yourganov, R. S. Zemel, G. E. Hinton, S. L. Small, and S. C. Strother. Comparing classification methods for longitudinal fMRI studies. *Neural Comput*, 22(11):2729–62, 2010.

[83] Ian Chan, William Wells III, Robert V. Mulkern, Steven Haker, Jianqing Zhang, Kelly H. Zou, Stephan E. Maier, and Clare M. C. Tempany. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Medical Physics*, 30(9):2390–2398, 2003.

[84] Pieter C Vos, Thomas Hambrock, Jelle O Barenstz, and Henkjan J Huisman. Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI. *Phys Med Biol*, 55(6):1719, 2010.

[85] Sedat Ozer, Deanna L. Langer, Xin Liu, Masoom A. Haider, Theodorus H. van der Kwast, Andrew J. Evans, Yongyi Yang, Miles N. Wernick, and Imam S. Yetik. Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI. *Medical Physics*, 37(4):1873–1883, 2010.

[86] R. Lopes, A. Ayache, N. Makni, P. Puech, A. Villers, S. Mordon, and N. Betrouni. Prostate cancer characterization on MR images using fractal features. *Med Phys*, 38(1):83–95, 2010.

[87] D. R. Busch, W. Guo, R. Choe, T. Durduran, M. D. Feldman, C. Mies, M. A. Rosen, M. D. Schnall, B. J. Czerniecki, J. Tchou, A. DeMichele, M. E. Putt, and A. G. Yodh. Computer aided automatic detection of malignant lesions in diffuse optical mammography. *Med Phys*, 37(4):1840–9, 2010.

[88] K. Yan, T. Podder, L. Li, J. Joseph, D. R. Rubens, E. M. Messing, L. Liao, and Y. Yu. A real-time prostate cancer detection technique using needle insertion force and patient-specific criteria during percutaneous intervention. *Med Phys*, 36(9):4184–90, 2009.

[89] Keinosuke Fukunaga. *Introduction to statistical pattern recognition.* Computer science and scientific computing. Academic Press, 1990.

[90] S E Seltzer, D J Getty, C M Tempany, R M Pickett, M D Schnall, B J McNeil, and J A Swets. Staging prostate cancer with MR imaging: a combined radiologist-computer system. *Radiology*, 202(1):219–226, 1997.

[91] Dongjiao Lv, Xuemei Guo, Xiaoying Wang, Jue Zhang, and Jing Fang. Computerized characterization of prostate cancer by fractal analysis in MR images. *Journal of Magnetic Resonance Imaging*, 30(1):161–168, 2009. 10.1002/jmri.21819.

[92] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Trans. Syst., Man, Cybern.*, 3(6):610–621, 1973.

[93] J.C. Russ. *The Image Processing Handbook.* CRC Press, 5th edition, 2007.

[94] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):55–73, 1990.

[95] P. C. Vos, T. Hambrock, C. A. Hulsbergen-van de Kaa, J. J. Futterer, J. O. Barentsz, and H. J. Huisman. Computerized analysis of prostate lesions in the peripheral zone using dynamic contrast enhanced MRI. *Med Phys*, 35(3):888–99, 2008.

[96] P. C. Vos, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Phys Med Biol*, 57(6):1527–1542, 2012.

[97] Y. Artan, M. A. Haider, D. L. Langer, T. H. van der Kwast, A. J. Evans, Y. Yang, M. N. Wernick, J. Trachtenberg, and I. S. Yetik. Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields. *IEEE Trans. Image Process.*, 19(9):2444–55, 2010.

[98] Christoph Busch. Wavelet based texture segmentation of multi-modal tomographic images. *Comput Graph*, 21(3):347–358, 1997.

[99] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. 2nd European Conf. Computational Learning Theory*, pages 23–37, 1995.

[100] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[101] X. Fern and C. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proc. 20th Int'l Conf. Machine Learning*, pages 186–193, 2003.

[102] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889–899, 2006.

[103] O. Samko, A. Marshall, and P. Rosin. Selection of the optimal parameter value for the Isomap algorithm. *Pattern Recognition Letters*, 27(9):968–979, 2006.

[104] O. Kouropteva, O. Okun, and M. Pietikainen. Selection of the Optimal Parameter Value for the Locally Linear Embedding Algorithm. In *Proc. 1st Int'l Conf. Fuzzy Systems and Knowledge Discovery*, pages 359–363, 2002.

[105] V. de Silva and J. Tenenbaum. Global Versus Local Methods in Nonlinear Dimensionality Reduction. In *Proc. 15th Conf. Adv. Neural Information Processing Systems (NIPS)*, pages 705–712, 2003.

[106] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Proc. 17th Conf. Adv. Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2004.

[107] X. Geng, D. C. Zhan, and Z. H. Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. Syst., Man, Cybern. B*, 35(6):1098–107, 2005.

[108] D. de Ridder, O. Kouropteva, O. Okun, M. Pietikainen, and R. Duin. Supervised locally linear embedding. In *Proc. Artificial Neural Networks and Neural Information Processing*, pages 333–341, 2003.

[109] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: An embedding method for efficient nearest neighbor retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):89–104, 2008.

[110] Neil Lawrence. Spectral Dimensionality Reduction via Maximum Entropy. In *Proc. 14th Intn'l Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 51–59, 2011.

[111] Kai Mao, Feng Liang, and Sayan Mukherjee. Supervised Dimension Reduction Using Bayesian Mixture Modeling. In *Proc. 13th Intn'l Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 501–508, 2010.

[112] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. 11th Annual Conf. Computational Learning Theory*, pages 92–100, 1998.

[113] C. Hou, C. Zhang, Y. Wu, and F. Nie. Multiple view semi-supervised dimensionality reduction. *Pattern Recognition*, 43(3):720–730, 2009.

[114] C Wachinger, M Yigitsoy, and N Navab. Manifold Learning for Image-Based Breathing Gating with Application to 4D Ultrasound. In *Proc. 13th Int'l Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 26–33, 2010.

[115] Christian Wachinger and Nassir Navab. Manifold Learning for Multi-Modal Image Registration. In *Proc. 11th British Machine Vision Conference (BMVC)*, pages 82.1–82.12, 2010.

[116] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley, New York, 2nd edition, 2001.

[117] V. N. Vapnik. An overview of statistical learning theory. *IEEE Trans. Neural Netw.*, 10(5):988–99, 1999.

[118] J. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[119] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13(1):21–27, 1967.

[120] P. Tiwari, M. Rosen, and A. Madabhushi. Consensus-locally linear embedding (C-LLE): application to prostate cancer detection on magnetic resonance spectroscopy. In *Proc. 11th Int'l Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 330–8, 2008.

[121] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832 –844, 1998.

[122] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Proc. 17th Conf. Adv. Neural Information Processing Systems (NIPS)*, pages 777–784, 2005.

[123] L.I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004.

[124] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

[125] Jagdish K. Patel and Campbell B. Read. *Handbook of the normal distribution*. Marcel Dekker, 1996.

[126] C. Yang, R. Duraiswami, N. A. Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *Proc. 9th IEEE Int'l Conf. Computer Vision (ICCV)*, pages 664–671, 2003.

[127] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[128] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article6, 2006.

[129] J. Carballido-Gamio, S.J. Belongie, and S. Majumdar. Normalized cuts in 3-D for spinal MRI segmentation. *IEEE Trans. Med. Imag.*, 23(1):36–44, 2004.

[130] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13:51–60, 2002.

[131] Haitao Zhao. Combining labeled and unlabeled data with graph embedding. *Neurocomputing*, 69(16):2385 – 2389, 2006.

[132] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.

[133] Maher Moakher and Philipp G. Batchelor. *Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization.* 2006.

[134] Y Bengio, J Paiement, P Vincent, O Delalleau, N Le Roux, and M Ouimet. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *Proc. 16th Conf. Adv. Neural Information Processing Systems (NIPS)*, pages 177–184, 2004.

[135] Julie Bulman, Robert Toth, Amish Patel, B Nicolas Bloch, Colm J McMahon, Long Ngo, Anant Madabhushi, and Neil M Rofsky. Automated Computer-derived Prostate Volumes from MR Imaging Data: Comparison with Radiologist-derived MR Imaging and Pathologic Specimen Volumes. *Radiology*, 262(1):144–51, 2012.

[136] A. Madabhushi and J. K. Udupa. New methods of MR image intensity standardization via generalized scale. *Medical Physics*, 33(9):3426–34, 2006.

[137] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imag.*, 17(1):87–97, 1998.

[138] C. Bartolozzi, I. Menchi, R. Lencioni, S. Serni, A. Lapini, G. Barbanti, A. Bozza, A. Amorosi, A. Manganelli, and M. Carini. Local staging of prostate carcinoma with endorectal coil MRI: correlation with whole-mount radical prostatectomy specimens. *European Radiology*, 6:339–345, 1996.

[139] T. R. Reed and J. M. H. Dubuf. A Review of Recent Texture Segmentation and Feature Extraction Techniques. *CVGIP: Img Understan*, 57(3):359–372, 1993.

[140] S. Herlidou, Y. Rolland, J. Y. Bansard, E. Le Rumeur, and J. D. de Certaines. Comparison of automated and visual texture analysis in MRI: Characterization of normal and diseased skeletal muscle. *Magn Reson Imaging*, 17(9):1393–1397, 1999.

[141] Bonilha L. Li L. Cendes F. Castellano, G. Texture analysis of medical images. *Clinical Radiology*, 59:1061–1069, 2004.

[142] V. A. Kovalev, F. Kruggel, H. J. Gertz, and D. Y. von Cramon. Three-dimensional texture analysis of MRI brain datasets. *IEEE Trans. Med. Imag.*, 20(5):424–433, 2001.

[143] T. Randen and J. H. Husoy. Filtering for texture classification: a comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(4):291–310, 1999.

[144] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

[145] Geoffrey J. McLachlan. *Discriminant analysis and statistical pattern recognition.* Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J., 2004.

[146] R. Toth, B. N. Bloch, E. M. Genega, N. M. Rofsky, R. E. Lenkinski, M. A. Rosen, A. Kalyanpur, S. Pungavkar, and A. Madabhushi. Accurate prostate volume estimation using multifeature active shape models on T2-weighted MRI. *Acad Radiol*, 18(6):745–54, 2011.

[147] G. Weiss and F. Provost. The Effect of Class Distribution on Classifier Learning: An Empirical Study. Technical report, Technical Report Technical Report ML-TR-44, Department of Computer Science, Rutgers University, 2001.

[148] Scott Doyle, James Monaco, John Tomaszewski, Michael Feldman, and Anant Madabhushi. An Active Learning Based Classification Strategy for the Minority Class Problem: Application to Histopathology Annotation. *BMC Bioinform*, pages Accepted, In Press, 2011.

[149] Anant Madabhushi, Jayaram Udupa, and Andre Souza. Generalized scale: theory, algorithms, and application to image inhomogeneity correction. *Comput Vis Image Underst*, 101(2):100–121, 2006.

[150] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, 2006.

[151] Z. Tu. Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering. In *Proc. IEEE ICCV*, pages 1589–1596, 2005.

[152] P. Tiwari, M. Rosen, G. Reed, J. Kurhanewicz, and A. Madabhushi. Spectral embedding based probabilistic boosting tree (ScEPTre): classifying high dimensional heterogeneous biomedical data. In *Proc. MICCAI*, volume 12, pages 844–51, 2009.

[153] RB Potts. Some generalized order-disorder transformations. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 106–109, 1952.

[154] J. Kurhanewicz, D. Vigneron, P. Carroll, and F. Coakley. Multiparametric magnetic resonance imaging in prostate cancer: present and future. *Curr Opin Urol*, 18(1):71–7, 2008.

[155] M. Chen, H. D. Dang, J. Y. Wang, C. Zhou, S. Y. Li, W. C. Wang, W. F. Zhao, Z. H. Yang, C. Y. Zhong, and G. Z. Li. Prostate cancer detection: comparison of T2-weighted imaging, diffusion-weighted imaging, proton magnetic resonance spectroscopic imaging, and the three techniques combined. *Acta Radiol*, 49(5):602–10, 2008.

[156] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinform*, 19(13):1636–1643, 2003.

[157] H. Yoshida and J. Nappi. Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps. *IEEE Trans. Med. Imag.*, 20(12):1261 –1274, 2001.

[158] J.D. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proc 20th ICML*, volume 20, pages 616–623, 2003.

[159] Malik Yousef, Michael Nebozhyn, Hagit Shatkay, Stathis Kanterakis, Louise C. Showe, and Michael K. Showe. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinform*, 22(11):1325–1334, 2006.

[160] J. F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *IEE Proc F*, 140(6):362–370, 1993.

[161] Y. Li, W. Zhang, and C. Lin. Simplify support vector machines by iterative learning. *Neu Inf Proc: Letters and Reviews*, 10:11–17, 2006.

[162] Ludmila Kuncheva, Christopher Whitaker, Fabio Roli, and Josef Kittler. Using Diversity with Three Variants of Boosting: Aggressive, Conservative, and Inverse Multiple Classifier Systems. In *Proc. 3rd Int'l Workshop Mult Class Systems*, pages 717–720, 2002.

Vita

Satish Easwar Viswanath

2012	Ph.D. in Biomedical Engineering, Rutgers University
2005	MSc in Medical Imaging, University of Aberdeen
2004	BE in Information Technology, University of Mumbai

2011-2012	Graduate Fellow, Department of Biomedical Engineering, Rutgers University
2008-2011	Graduate Assistant, Department of Biomedical Engineering, Rutgers University
2007-2008	Coulter Translational Graduate Fellow, Department of Biomedical Engineering, Rutgers University