MACHINE–BASED INTERPRETATION AND CLASSIFICATION OF IMAGE–DERIVED FEATURES: APPLICATIONS IN DIGITAL PATHOLOGY AND MULTI–PARAMETRIC MRI OF PROSTATE CANCER

by

SHOSHANA B. GINSBURG

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Biomedical Engineering

CASE WESTERN RESERVE UNIVERSITY

May, 2016
CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

We hereby approve the dissertation of

Shoshana Ginsburg

candidate for the degree of Doctor of Philosophy*.

Committee Chair:
Dr. David Wilson

Committee Member:
Dr. Anant Madabhushi

Committee Member:
Dr. Vikas Gulani

Committee Member:
Dr. Michael Kattan

March 10, 2016

*We also certify that written approval has been obtained for any proprietary material contained therein.
For Meir,
who stands by me through thick and thin.
# Contents

Acknowledgements ........................................... xxii

Abstract .................................................... xxiii

1 Introduction .............................................. 1

1.1 Overview ............................................... 1

1.2 Challenges with constructing robust classifiers with high-dimensional features ............................................. 2

1.3 Significance of high-dimensional classifier-building for prostate cancer characterization on multi-parametric magnetic resonance imaging (MRI) ...................................................... 5

1.3.1 Multi-parametric MRI for prostate cancer detection and grading ...................................................... 5

1.3.2 Multi-parametric MRI for predicting risk of prostate cancer recurrence following therapy ...................................................... 7

1.3.3 Significance of prostate zone for characterization of prostate tumors ...................................................... 9

1.4 Significance of high-dimensional classifier-building for predicting patient prognosis on digital pathology ...................................................... 11

1.5 Summary of the major goals of this thesis ...................................................... 11
2 Previous Work and Novel Contributions

2.1 Previous work in constructing interpretable high-dimensional classifiers .................................................. 14

2.2 Previous work in using image-derived features for cancer characterization .............................................. 15
   2.2.1 Image-derived texture features for prostate cancer detection and grading ........................................... 15
   2.2.2 Accounting for differences between transition and peripheral zone tumors ........................................... 17
   2.2.3 Image-derived pharmacokinetic features for prostate cancer detection and grading ........................... 18
   2.2.4 Image-derived digital pathology features for predicting cancer recurrence ........................................... 19

2.3 Novel contributions ........................................................................................................................................... 20
   2.3.1 Developing novel approaches for constructing robust, interpretable classifiers in high-dimensional settings .... 20
   2.3.2 Identifying zone-specific multi-parametric MRI signatures for prostate cancer characterization ................ 21
   2.3.3 Developing a novel approach for estimating pharmacokinetic parameters in the prostate ...................... 21

3 Methods for Classifier-Building in High-Dimensional Spaces .............................................................................. 23

3.1 Overview .......................................................................................................................................................... 23

3.2 Review: Dimensionality Reduction as an Eigenvalue Problem ................................................................. 25
   3.2.1 Eigenvalue Problems for Dimensionality Reduction .............................................................................. 26
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3.2.2 NLDR Methods as Variants of KPCA</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>3.2.2 NLDR Methods as Variants of KPCA</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>3.2.2 NLDR Methods as Variants of KPCA</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>3.3 Calculating Feature Importance</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>3.3.1 Variable Importance in Projections (VIP)</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>3.3.2 Feature Importance in Nonlinear Embeddings (FINE)</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>3.4 Evaluation of FINE on publicly available datasets</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>3.4.1 Datasets</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>3.4.2 Experimental Design</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>3.4.2 Experimental Design</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>3.4.2 Experimental Design</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>3.4.3 Experimental Results and Discussion</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>Overview</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>Materials and methods</td>
<td>42</td>
</tr>
</tbody>
</table>
4.2.1 Data description ........................................... 42
4.2.2 Postprocessing of MRI ................................. 44
4.2.3 Registration of MRI and WMHS slices ......... 44
4.2.4 Feature extraction from multi-parametric MRI ... 46
4.2.5 Evaluation .................................................. 48
4.2.6 Relative importance of feature subsets ........ 49
4.3 Results ........................................................ 50
  4.3.1 Features selected by VIP ......................... 50
  4.3.2 Stability of VIP feature subsets ............... 50
  4.3.3 Classification accuracy provided by selected features .... 52
  4.3.4 Comparison of signal intensity and radiomic features .... 52
  4.3.5 Ranking MRI protocols for prostate cancer localization .... 54
4.4 Discussion .................................................... 56
  4.4.1 Features selected by VIP ......................... 56
  4.4.2 Stability of VIP feature subsets ............... 57
  4.4.3 Classification accuracy provided by selected features .... 57
  4.4.4 Comparison of signal intensity and radiomic features .... 58
  4.4.5 Ranking MRI protocols for prostate cancer localization .... 58
  4.4.6 Limitations ............................................. 59

5 Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer 61
  5.1 Overview .................................................. 61
  5.2 Materials and methods ............................... 62
    5.2.1 Data description .................................. 62
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.2</td>
<td>Annotation of ground truth cancer extent</td>
<td>65</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Feature extraction from T2w MRI</td>
<td>66</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Feature selection via VIP</td>
<td>67</td>
</tr>
<tr>
<td>5.3</td>
<td>Experimental design</td>
<td>67</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Experiment 1: Evaluation of VIP for feature selection</td>
<td>67</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Experiment 2: Comparative evaluation of classifier for biochemical recurrence risk prediction</td>
<td>68</td>
</tr>
<tr>
<td>5.4</td>
<td>Experimental results and discussion</td>
<td>69</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Experiment 1: Evaluation of VIP for feature selection</td>
<td>69</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Experiment 2: Comparative evaluation of classifier for biochemical recurrence risk prediction</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>Applications of FINE in Digital Pathology</td>
<td>75</td>
</tr>
<tr>
<td>6.1</td>
<td>Overview</td>
<td>75</td>
</tr>
<tr>
<td>6.2</td>
<td>Datasets</td>
<td>77</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Predicting breast cancer recurrence based on tissue microarrays</td>
<td>77</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Predicting OncotypeDX risk categories on whole slides</td>
<td>77</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Predicting biochemical recurrence of prostate cancer based on tissue whole mounts</td>
<td>79</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Predicting biochemical recurrence of prostate cancer based on tissue microarrays</td>
<td>80</td>
</tr>
<tr>
<td>6.3</td>
<td>Experimental design</td>
<td>80</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Dimensionality reduction</td>
<td>80</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Classifier training</td>
<td>81</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Performance evaluation measures</td>
<td>82</td>
</tr>
</tbody>
</table>
6.4 Experimental results and discussion ........................................... 82
  6.4.1 Predicting breast cancer recurrence based on tissue microarrays 82
  6.4.2 Predicting OncotypeDX risk categories on whole slides ....... 84
  6.4.3 Predicting biochemical recurrence of prostate cancer based on tissue whole mounts ......................................................... 87
  6.4.4 Predicting biochemical recurrence of prostate cancer based on tissue microarrays ......................................................... 88

7 Classifier–Building in a Multi–Institutional Study: Zone–Specific Prostate Cancer Detection on Multi–Parametric MRI 90
  7.1 Overview ................................................................. 90
  7.2 Materials and Methods .................................................. 91
    7.2.1 Patients ............................................................ 91
    7.2.2 MRI acquisition ...................................................... 93
    7.2.3 Histopathological analysis and cancer annotation on MRI .. 93
    7.2.4 Cancer annotation on pre–biopsy MRI ......................... 94
    7.2.5 MRI post–processing ................................................. 94
    7.2.6 Radiomic features .................................................... 95
    7.2.7 Identifying features for discriminating cancerous from benign voxels ................................................................. 96
    7.2.8 Classifier training and evaluation .................................... 97
  7.3 Results ................................................................. 98
    7.3.1 Selected features for TZ and PZ classifiers .................... 98
    7.3.2 Classifier performance .............................................. 101
    7.3.3 Correlation with Gleason grades ................................. 101
7.3.4 Effect of training cohort on classifier performance . . . . . . . 101
7.3.5 Differences between small and large tumors . . . . . . . . . 104
7.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105

8 Pharmacokinetic Features for Prostate Cancer Detection on MRI 109
8.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109
8.2 Materials and Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
  8.2.1 Patients . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
  8.2.2 MRI Acquisition . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
  8.2.3 Histopathological Analysis . . . . . . . . . . . . . . . . . . . . . 113
  8.2.4 Cancer Annotation on MRI . . . . . . . . . . . . . . . . . . . . . 113
  8.2.5 Pharmacokinetic Modeling on DCE MRI . . . . . . . . . . . . . 114
  8.2.6 Evaluation of Pharmacokinetic Modeling Methods . . . . . . 117
8.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
  8.3.1 Discriminating Cancer from Benign Tissue . . . . . . . . . . . . 119
  8.3.2 Predicting Prostate Cancer Risk Group . . . . . . . . . . . . . . 121
  8.3.3 Reverse Approximation of AIF Based on Pharmacokinetic Pa-
         rameters . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125
8.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

9 Concluding Remarks 129

10 Future Work 134

11 Appendix A: Classification Strategies 136
  11.1 Notation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>List of common mathematical notation in this chapter.</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Description of four publicly available datasets used in this paper.</td>
<td>34</td>
</tr>
<tr>
<td>3.3</td>
<td>Average value of the intrinsic dimensionality parameter $h$, selected via residual variance minimization, for four NIPS datasets and four dimensionality reduction algorithms.</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>Description of four sets of patient studies that were analyzed separately in this study</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>Description of multi-parametric MRI signal intensity features used for TZ and PZ prostate cancer localization and their associated VIP scores. SI, signal intensity</td>
<td>47</td>
</tr>
<tr>
<td>4.3</td>
<td>Description of features computed on a per-voxel basis from T2w MRI and ADC maps and the motivation for using them for prostate cancer localization</td>
<td>48</td>
</tr>
<tr>
<td>4.4</td>
<td>Description of contrast kinetic features computed on a per-voxel basis from DCE MRI</td>
<td>48</td>
</tr>
<tr>
<td>4.5</td>
<td>The top 10 radiomic features</td>
<td>51</td>
</tr>
</tbody>
</table>
5.1 Description of features computed on a per-voxel basis from T2w 
MRI and the motivation for using them for predicting biochemical 
recurrence of prostate cancer. SI = signal intensity. .................... 66

5.2 Top three features selected by the VIP feature selection scheme and 
by mRMR and their associated VIP scores. SD = standard deviation. 70

5.3 Lists of semantic MRI attributes and clinical features assessed via 
the Kattan nomogram. .............................................................. 70

6.1 Description of datasets used in this paper to address four digital 
pathology problems. ER=estrogen receptor. ............................... 77

6.2 Average $h$ and top features selected by FINE in conjunction with 
four DR methods for each of the pathology datasets (obtained by 
voting across 50 bootstrapping rounds). NN=nearest neighbors, SD=standard 
deviation. ................................................................................... 85

7.1 Description of subjects used in this study. DW, diffusion–weighted; 
DCE, dynamic contrast–enhanced; SD, standard deviation ............. 92

7.2 Overview of radiomic features used in this study ....................... 96

7.3 Top ten radiomic features selected based on an intraclass correlation 
coefficient above 0.9 and high AUC values in distinguishing prostate 
cancer from benign tissue. Features extracted from ADC maps are 
shown in bold font; features listed in regular font were extracted 
from T2–weighted MRI. ................................................................. 99
7.4 For all combinations of training and testing datasets, the mean AUC for TZ and PZ cancer detection is shown for the zone–aware ($C^{TZ}$ and $C^{PZ}$) and zone–ignorant ($C^{ALL}$) classifiers. Two–sample t–tests were used to evaluate the statistical significance of differences in AUC between zone–aware and zone–ignorant classifiers. . . . . . . . 102

8.1 For forty subjects and two methods of pharmacokinetic parameter estimation, parameters $K^{trans}$ and $v_e$ are listed as mean ± standard deviation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120

8.2 For both the AIF–and prostate–based methods, the number of patients with significant differences between normal prostate zones and cancer, evaluated via Kruskal–Wallis significance tests, are listed. Additionally, AUC values for cancer detection on the basis of pharmacokinetic parameters are shown as mean (min–max). . . . . . . . . . . . . . . . 121

8.3 Correlation coefficients obtained via Pearson’s correlation between lesion size and 74th and 100th percentiles of $K^{trans}$ and $v_e$ are shown. Correlations that are statistically significant at a level of $\alpha = 0.01$ are shown with an asterisk. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121

8.4 AUC values for discriminating between lesions of different cancer grades based on the 75th and 100th percentiles of $K^{trans}$ and $v_e$ associated with each lesion. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125
List of Figures

1.1 The prostate gland contains two main structures: the central gland, comprised of the transition and central zones, shown in blue and the peripheral zone shown in green. ................................................. 10

3.1 Illustration of modules comprising FINE to identify key features in a high dimensional dataset: First a large number $m$ of features (typically far greater than the number of patient samples $n$) are extracted, DR is performed to reduce the dimensionality of the data to $h$. FINE is then leveraged to identify key features. The predictive power of these key features is then evaluated via a classifier. ......................... 25

3.2 AUC and Jaccard index associated with four dimensionality reduction algorithms, the top five features selected based on FINE scores, and three common feature selection algorithms for NIPS datasets (a) Madelon, (b) Arcene, (c) Dexter, and (d) Gisette. Data points are shown for three classification methods: logistic regression (black outline), support vector machine (green outline), and random forest (red outline) classifiers. ......................................................... 38
4.1 A T2w MR image is shown (a) before and (e) after correction of bias field inhomogeneities. Signal intensity histograms associated with several patients are shown before and after intensity standardization for T2w MRI in (b) and (f), ADC maps in (c) and (g), and DCE MRI in (d) and (h), respectively.

4.2 Checkerboard images illustrating registration of T2w MRI and DCE MRI are shown in (a) and (f) for patients with PZ and TZ cancer, respectively; registration of T2w MRI and ADC maps is shown in (b) and (g). The original 2D whole–mount histological sections, with prostate cancer extent outlined in blue by a pathologist, are shown in (c) and (h), and overlays of whole–mount histological sections and T2w MRI after nonlinear multimodal registration are shown in (d) and (i). Mapped cancer extent on T2w MRI, outlined in red, is shown in (e) and (j).

4.3 The Jaccard index, used to assess the stability of selected feature subsets, is compared when VIP and mRMR are employed to select the top 10 features.

4.4 AUC values, averaged across 30 cross–validation runs, are plotted for each of the top 10 radiomic features for (a) a naive Bayes classifier and (b) a logistic regression classifier. The AUCs that arise when aggregating the top 1–10 radiomic features for classification are shown in (c) for a naive Bayes classifier and in (d) for a logistic regression classifier.
4.5 Ground truth extent of prostate cancer is delineated on T2w MRI in red for a representative slice of (a) a PZ tumor and (b) a TZ tumor. Corresponding DCE MR images at peak contrast enhancement and ADC maps are shown in (b) and (h) and in (c) and (i), respectively. Heatmaps representing the pixel-wise probability of cancer presence, obtained via logistic regression classification, are shown in (d) and (j) using all of the radiomic features, (e) and (k) using signal intensity features, and (f) and (l) using $J_{PZ}$ and $J_{TZ}$, respectively. Red indicates a high probability of cancer presence, yellow indicates a low probability of cancer presence, and blue indicates no cancer presence.

4.6 Relative importance of (a) all features and (b) features in $J_{PZ}$ and $J_{TZ}$ extracted from T2w, DWI, and DCE MRI for each dataset.

5.1 Flowchart illustrating methodology used in this paper for extraction, scoring, ranking, and evaluation of radiomic features for predicting biochemical recurrence risk.
5.2 The top–performing Gabor feature selected by VIP (top) and a first–order statistical feature (bottom) is shown for (a) and (d) a recurrence–free patient and (b) and (e) a patient who develops biochemical recurrence with the ground truth cancer extent bounded in white. The histograms in (c) and (f) show the distributions of these features in cancerous regions of the prostate for the recurrence–free patient (blue) and the patient who experiences biochemical recurrence (red). Note that significant differences in both the skewness and kurtosis of the distributions of the Gabor wavelet feature can be appreciated in (c) but not in (f).

5.3 (a) AUC is compared for various numbers of radiomic features selected by VIP and mRMR. (b) AUC is compared for the top–performing features obtained via the VIP and mRMR feature selection algorithms, as well as for semantic MRI attributes and the Kattan nomogram.

5.4 Scatterplot of the top–performing Gabor features selected by (a) VIP and (b) mRMR.

5.5 Kaplan–Meier survival curves obtained via (a) the three radiomic features selected by VIP and (b) the Kattan nomogram. Differences in survival curves between patients who experienced biochemical failure and those who did not were not statistically significant.
6.1 Digital pathology and feature representations. For dataset \( S_1 \) a representative patch of histology is shown in (a); segmented nuclei are illustrated in (b); and Voronoi and Delaunay graphs constructed based on the nuclei are shown in (c) and (d), respectively. For dataset \( S_2 \) a representative patch of histology is shown in (e); segmented nuclei are illustrated in (f); and texture representations of the histology are shown in (g) and (h). For dataset \( S_3 \) a representative patch of histology is shown in (i); segmented nuclei are illustrated in (j); and Voronoi and Delaunay graphs constructed based on the nuclei are shown in (k) and (l), respectively. For dataset \( S_4 \) a representative patch of histology is shown in (m); segmented glands are illustrated in (n); and Voronoi and Delaunay graphs constructed based on the glands are shown in (o) and (p), respectively.

6.2 AUC and Jaccard index associated with four dimensionality reduction algorithms, the top five features selected based on FINE scores, and three common feature selection algorithms for pathology datasets (a) \( S_1 \), (b) \( S_2 \), (c) \( S_3 \), and (d) \( S_4 \). Data points are shown for three classification methods: logistic regression (black outline), support vector machine (green outline), and random forest (red outline) classifiers.

6.3 Heatmaps of the three top-ranking features selected via FINE for dataset \( S_2 \).

6.4 Kaplan–Meier survival curves \((p = 0.02)\) obtained using the five top-ranking features selected by FINE to predict the likelihood of recurrence–free survival in cohort \( S_4 \).
7.1 Ground truth extent of prostate cancer is delineated on T2w MRI in red for a representative slice of (left) a PZ tumor and (right) a TZ tumor. Feature maps of the top two selected features populating $F_{12}^{PZ}$ and $F_{12}^{TZ}$, respectively, are also shown. Heatmaps representing the pixel-wise probability of cancer presence, obtained via logistic regression classifiers trained on data from $I_1$ and $I_2$, are shown when $C_{12}^{ALL}$ is used and when $C_{12}^{PZ}$ and $C_{12}^{TZ}$ complement each other to detect cancer in both the PZ and TZ. Red indicates a high probability of cancer presence, yellow indicates a low probability of cancer presence, and blue, the absence of cancer.

7.2 Scatterplot of AUC yielded by the spatially-aware classifier on a per-patient basis and (a) Gleason grades for patients from Institution 1 with TZ tumors, (b) Gleason grades for patients from Institution 1 with PZ tumors, and (c) overall tumor sizes for all 80 patients.

8.1 Criteria for inclusion of forty patients with prostate cancer in this study.

8.2 Average concentration of contrast agent in the femoral arteries (AIF), peripheral zone, and transition zone for a sample patient.

8.3 Ground truth extent of prostate cancer is shown for three lesions on the top row. Heatmap representations of $K_{trans}$ are displayed for the AIF–based approach (middle) and PDM (bottom).
8.4 ROC curves illustrate the predictive power of $K^{\text{trans}}$ in discriminating (a) low from high grade cancer and (b) intermediate from high grade cancer and $v_e$ in discriminating (c) low from high grade cancer and (d) intermediate from high grade cancer.
I would like to express my gratitude first and foremost to Dr. Anant Madabhushi, who, as my advisor and mentor, dedicated considerable time and effort to help me achieve success in my studies and research. In particular, I am grateful for his insight that has enabled me to identify and overcome my weaknesses, grow from failure, work hard, and aim for perfection. Thank you also to all members of the Center of Computational Imaging and Personalized Diagnostics for the good times we shared. I would also like to thank my committee members and collaborators, Dr. David Wilson, Dr. Vikas Gulani, Dr. Michael Kattan, Dr. Ivan Jambor, and Dr. Art Rastinehad, for their insight and ideas to make my research strong.

This work was made possible by grants from Rutgers University Presidential Fellowship; the National Science Foundation Graduate Research Fellowship; the National Cancer Institute of the National Institutes of Health under award numbers R01CA136535-01, R01CA140772-01, R21CA167811-01, R21CA179327-01, R21CA195152-01; the National Institute of Diabetes and Digestive and Kidney Diseases under award number R01DK098503-02; the DOD Prostate Cancer Synergistic Idea Development Award (PC120857); the DOD Lung Cancer Idea Development New Investigator Award (LC130463); the DOD Prostate Cancer Idea Development Award; the Ohio Third Frontier Technology Development Grant; the CTSC Coulter Annual Pilot Grant; the Case Comprehensive Cancer Center Pilot Grant; VelaSano Grant from the Cleveland Clinic; and the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University.

I am eternally grateful to my husband Meir for his undying support and encouragement, without which this dissertation would not be possible. Thank you...
also to my parents for teaching me to aim for excellence and encouraging me to follow through on my goals and to my in–laws for their love and support. A special thanks to my daughters Chassy and Brochy for the joie de vivre that has brought wonderful balance to my life. Finally, I cannot begin to express my deepest gratitude to Hashem Yisborach for making everything possible.
Abstract

Machine-Based Interpretation and Classification of Image-Derived Features: Applications in Digital Pathology and Multi-Parametric MRI of Prostate Cancer

by Shoshana Ginsburg

The analysis of medical images—from magnetic resonance imaging (MRI) to digital pathology—for disease characterization typically involves extraction of hundreds of features, which may be used to predict disease presence, aggressiveness, or outcome. Unfortunately, the dimensionality of the feature space poses a formidable challenge to the construction of robust classifiers for predicting disease presence and aggressiveness. In this work we present novel strategies to facilitate the construction of robust, interpretable classifiers when the dimensionality of the feature space is high. In the context of prostate cancer, we demonstrate the benefit of our approach for identifying (a) radiomic features that are useful for detecting prostate cancer on multi-parametric MRI, (b) radiomic features that predict the risk of prostate cancer recurrence on T2-weighted MRI, and (c) histomorphometric features describing cellular and glandular architecture on digital pathology images that predict the risk of prostate cancer recurrence post-treatment. In the context
of breast cancer, we identify histomorphometric features describing cancer patterns in estrogen receptor positive (ER+) breast cancer tissue slides that can predict (a) which cancer patients will have recurrence following treatment with tamoxifen and (b) risk category as determined by a 21 gene expression assay called Oncotype DX. Additionally, we also investigate whether radiomic features characterizing prostate tumors that manifest in the peripheral zone of the prostate are different from radiomic features characterizing transition zone tumors, and we develop a novel approach for pharmacokinetic modeling on dynamic contrast–enhanced MRI that relies exclusively on prostate voxels, with no reliance on an arterial input function or reference tissue.
Chapter 1

Introduction

1.1 Overview

The analysis of medical images—from magnetic resonance imaging (MRI) to digital pathology—for disease characterization typically involves extraction of hundreds of features, which may be used to predict disease presence, aggressiveness, or outcome. Unfortunately, the dimensionality of the feature space poses a formidable challenge to the construction of robust classifiers for predicting disease presence and aggressiveness. These challenges have to do with the “curse of dimensionality” [1] and the “curse of data sparsity” [2] since the number of features is significantly larger than the number of training exemplars (typically patient studies). A per-class sample-to-feature ratio of 10:1 is generally recommended for building reliable and generalizable classifiers and predictive models [3], but this sample-to-feature ratio is generally unheard of in medical image analysis problems. Thus, it can be quite challenging to build a robust and generalizable predictive model when hundreds of imaging features are being analyzed.
In this dissertation we present novel strategies that combine dimensionality reduction and feature selection techniques to facilitate the construction of robust classifiers when the dimensionality of the feature space is high. We demonstrate the benefit of our approach for identifying radiomic features that are useful for detecting prostate cancer and predicting the risk of prostate cancer recurrence on multiparametric MRI. Additionally, we demonstrate how our approach can be leveraged to identify histomorphometric features describing cellular and glandular architecture on digital pathology images to predict the risk of breast and prostate cancers recurring post–treatment.

1.2 Challenges with constructing robust classifiers with high–dimensional features

According to the Hughes effect [1], predictive power reduces as dimensionality increases; this effect, termed the “curse of dimensionality,” is accentuated when the sample size is small. Consequently, when only a limited number of training samples is available, a generalizable classifier cannot be constructed with more than a handful of features [3]. In order to construct a generalizable classifier, the dimensionality of the feature space must be reduced.

In order to reduce the dimensionality of the feature space and thereby facilitate classifier construction, dimensionality reduction (DR) is often performed [4–7]. DR involves transforming a high dimensional dataset into a low–dimensional eigenspace; the resulting eigenvectors replace the original features as inputs to the
classifier. The rationale here is that classifiers trained on features in a reduced dimensional space are more robust and reproducible than classifiers constructed in the original high dimensional feature space. Additionally, since the reduced dimensional features capture most of the data variance embedded in the original high dimensional space, the reduced dimensional classifiers are no worse off in terms of class discriminability than classifiers in the original space. However, one major disadvantage of DR is the fact that the new, transformed features tend to be divorced from domain–specific meaning. Consequently, resulting classifiers are opaque, and it is a challenge to identify which features contribute most substantially to class discriminability. In problems involving clinical decision support for medical imaging data, it is often important to not only be able to create an accurate classifier of disease presence and aggressiveness, but also to identify the features that contribute most substantially to class separability. This feature transparency is often a pre–requisite for adoption of clinical decision support classification tools since physicians are typically resistant to opaque "black box" prediction models.

Linear DR methods (e.g., principal components analysis (PCA)) rely on Euclidean distances to estimate similarities between features and do not account for the inherent nonlinear structures underlying most biomedical data, yet it remains challenging to understand the relationships between the original features and the PCA embedding. For high dimensional biomedical datasets, nonlinear DR (NLDR) methods can capture the underlying nonlinear manifold structure of the data. As a result, classifiers modeled on the NLDR representations have been shown to provide higher class discrimination compared to the corresponding classifiers trained
off linear DR representations [8]. For example, since quantitative histomorphometric data is inherently nonlinear, NLDR provides a superior embedding of the data when compared to linear DR [9–11]. However, whereas PCA involves the eigen decomposition of the data itself, NLDR involves the eigen decomposition of a similarity matrix instead. Consequently, compared to PCA, it is even more difficult to reconcile the contributions of the individual features to the low dimensional representations constructed via NLDR schemes.

Apart from DR, feature selection is another approach to reduce the dimensionality of the feature space and thereby enable classifier construction. Feature selection methods include wrappers and embedded methods [12], which typically involve selection of feature subsets that perform well in a particular classifier training framework, and filters [13, 14], which select features based on a variable ranking scheme. Because wrappers and embedded methods utilize a learning algorithm for feature selection, the selected features may not perform well in conjunction with other classifiers. As a result, these methods may not be ideal for relating which computer–extracted features are most strongly associated with disease appearance on digital pathology images. Instead, filter methods are more robust against overfitting, providing more stable sets of selected features [15]. However, one limitation of filters is their inability to account for interactive effects among features. Also, filter methods, as well as wrapper and ensemble methods, may lead to a number of different “optimal” feature subsets that produce equally good classification results [2].

Hence, a new classification method that relies on DR to overcome the curse of
dimensionality but that allows for identifying the relative contributions of the individual features to the reduced dimensional manifold would be highly desirable. While a method has previously been introduced to rank features based on their contributions to an embedding derived via partial least squares (PLS) [16], to the best of our knowledge no methods have been presented to rank features based on the more popular PCA or nonlinear DR techniques.

1.3 Significance of high–dimensional classifier–building for prostate cancer characterization on multi–parametric magnetic resonance imaging (MRI)

1.3.1 Multi–parametric MRI for prostate cancer detection and grading

During the past decade multi–parametric magnetic resonance imaging (MRI) has emerged as the imaging modality of choice for disease staging in patients with biopsy-confirmed prostate cancer [17]. Additionally, MRI is becoming more popular for prostate cancer detection and localization in patients with repeated negative biopsies but elevated prostate specific antigen (PSA) and to facilitate targeted biopsies. Multi–parametric prostate MRI provides excellent contrast of anatomic structures on T2–weighted (T2w) MRI, as well as indications about tissue diffusion characteristics and microvasculature on diffusion–weighted (DW) MRI and dynamic contrast–enhanced (DCE) MRI, respectively. On multi–parametric MRI, prostate cancer detection rates for radiologists in the peripheral zone (PZ) of up to
an area under the curve (AUC) of 0.88 and in the transition zone (TZ) of up to an AUC of 0.73 have been reported [18], although the detection rates tend to be higher for high grade prostate cancer. The fact that one quarter of prostate cancers in the TZ might not be detected on MRI may be due to the presence of benign tumor confounding pathologies such as benign prostatic hyperplasia, which occurs primarily in the TZ and looks similar to carcinoma on T2w MRI [19]. Furthermore, areas of benign stroma or hyperplasia may manifest as restricted diffusion on DW MRI and heterogeneous enhancement on DCE MRI [20]. As a result, functional imaging is not necessarily useful for detecting TZ tumors.

Previous studies have shown that computer-assisted diagnosis (CAD) tools provide increased sensitivity and specificity in detecting prostate cancer on multiparametric MRI, to complement a radiologist’s assessment. Recently there has been substantial interest in the role of computer-extracted (or radiomic) texture features to quantitatively describe tissue microarchitecture and morphology. A number of recent CAD approaches have attempted to use texture features including first and second order co-occurring statistical measurements from T2w and diffusion weighted MRI and kinetic features from DCE MRI for prostate cancer detection [21–34]. It is thought that radiomic features are capable of quantitatively describing tissue microarchitecture and morphology, which provide clues about cancer presence.

Dynamic contrast-enhanced (DCE) MRI, which is useful for analyzing tissue microvasculature and angiogenesis, has been shown to facilitate improved prostate
cancer detection compared to T2w MRI alone [35–38]. On DCE MRI, tumors manifest rapid and increased enhancement and early washout compared to surrounding normal prostate tissue [39]. Pharmacokinetic analysis of DCE MRI enables the determination of parameters, such as $K^{\text{trans}}$ (transfer constant) and $v_e f$ (extravascular-extracellular volume fraction), describing tumor vasculature perfusion and permeability [40]. These constants are known to be elevated in prostate tumors and have been shown to be useful not only for detecting prostate cancer, but also for distinguishing prostate cancer grades [41–43].

1.3.2 Multi-parametric MRI for predicting risk of prostate cancer recurrence following therapy

Primary localized treatment options include radical prostatectomy and radiation therapy, which can be administered as external beam radiation therapy (EBRT) or brachytherapy with or without concomitant hormonal therapy. Unfortunately, as many as one third of patients who undergo localized treatment for prostate cancer develop biochemical recurrence [44], which, according to the ASTRO definition [45], is characterized by a rise in serum prostate-specific antigen (PSA) of at least 2 ng/mL above the nadir PSA. Approximately one third of prostate cancers associated with biochemical recurrence will eventually metastasize [44]. Furthermore, biochemical recurrence is associated with a marginally but significantly increased risk of clinical progression and prostate cancer-related mortality [46]. Consequently, pre-treatment prediction of a patient’s propensity towards biochemical
recurrence following radiation therapy would be beneficial for choosing an optimal treatment strategy. Thus, patients who are likely to experience biochemical recurrence after undergoing radiation therapy may be advised to choose other treatments.

Several nomograms have been developed for predicting patient response to radiation therapy; the most widely accepted one is the Kattan nomogram [47]. The Kattan nomogram incorporates baseline PSA level, tumor stage, Gleason score, radiation dose, and the use of hormonal therapy to predict the probability of remaining recurrence–free for five years. Although these factors are good predictors of biochemical recurrence, the Kattan nomogram is limited by the presence of benign prostatic hyperplasia, which impacts pretreatment PSA levels, and inaccuracies in the determination of the Gleason score that result from biopsy sampling errors.

During the past decade, magnetic resonance imaging (MRI) has emerged as an accurate method for evaluating prostate cancer stage and grade [48–51]. Due to the ability to discern anatomical structures on T2w MRI, semantic attributes of prostate cancer, such as the presence of extracapsular extension and seminal vesicle invasion and surrogate measurements of tumor size, can be assessed on T2w MRI. These semantic attributes have all been shown to be powerful independent predictors of five–year biochemical recurrence–free survival [52–54]. Moreover, when incorporated in a predictive model together with the clinical features assessed by the Kattan nomogram, these semantic or qualitative MRI features obtained via a radiologist’s interpretation can substantially augment the accuracy of the Kattan nomogram in predicting biochemical recurrence risk, leading to an area under the receiver operating characteristic (ROC) curve (AUC) of 0.78 compared to 0.61 for
Nevertheless, semantic features represent only a small fraction of possible MRI-derived attributes that may be useful for predicting biochemical recurrence. A number of recent studies have shown that radiomic features, such as texture features derived from T2w MRI, are useful for detecting prostate cancer [6, 22, 25, 55, 56] and differentiating between high and low Gleason grade [57] prostate cancer in vivo [58]. It is thought that radiomic features are capable of quantitatively describing tumor microarchitecture and morphology, which provide clues about tumor aggressiveness. Because high grade tumors tend to be more aggressive, they are more likely to trigger biochemical recurrence after treatment. Consequently, we expect that microarchitectures in the tumor may be correlated with biochemical recurrence risk, as well as Gleason grade. Since radiomic features were previously shown to effectively discriminate between prostate cancers of high and low Gleason grades [58], we hypothesize that radiomic features may capture morphometric clues for predicting biochemical recurrence as well.

1.3.3 Significance of prostate zone for characterization of prostate tumors

The prostate gland can be divided into three primary anatomical regions: the peripheral zone (PZ), the transition zone (TZ), and the central zone (see Figure 1.1). Prostate cancer in the central zone is rare [59]. While some computer assisted diagnosis tools are designed to look for prostate cancer only in the PZ [23–25, 60, 61] or only in the TZ [26, 27], many approaches tend to be zone–agnostic or zone–ignorant. However, the appearance of prostate cancer on multi–parametric MRI
tends to depend on the tumor’s location in the prostate gland [21, 22]. Whereas PZ tumors usually manifest as round or ill-defined hypointense lesions, TZ tumors are usually moderately hypointense, lenticular-shaped lesions, often with speculated margins [39]. Additionally, radiomic texture features characterizing TZ tumors tend to be different from those characterizing PZ tumors on multi-parametric MRI [21, 22, 30].

Although there has been some recent work on identifying zone specific radiomic features associated with prostate cancer in the TZ and PZ [22], this limited study involved only radiomic features extracted from T2w MRI and did not make use of DW imaging or DCE MRI. Furthermore, studies evaluating differences in radiomic features between TZ and PZ tumors have been specific to a single institution, and the resilience of these features was not evaluated in a cross-institutional setting. This is a particularly important consideration since the variance (or drift) in MRI parameters (T1w, T2w, Diffusion) across vendor platforms and scanners is well-known and documented [62]. It would be especially valuable to evaluate in a
multi-institutional study whether radiomic features for prostate cancer detection from multi-parametric 3 Tesla (T) MRI in the TZ are similar to the features that are useful for prostate cancer detection in the PZ.

1.4 Significance of high-dimensional classifier-building for predicting patient prognosis on digital pathology

Quantitative histomorphometry (QH) is the process of computationally modeling the appearance of disease on digital pathology images via image-based features. QH approaches typically involve the extraction of a large number of features describing the texture, color, and spatial arrangement of nuclei and glands on digitized images of tissue slides [9, 63–67]. These features have been shown to be useful in determining cancer aggressiveness [9, 66, 67] and in predicting the likelihood of a patient’s cancer recurring following treatment [65, 68]. For breast and prostate cancers (as well as other cancers), knowledge about which QH features are most predictive of risk of recurrence could potentially lead to improved disease characterization upon biopsy and better planning of therapy in the adjuvant and neoadjuvant settings.

1.5 Summary of the major goals of this thesis

In this work we develop novel strategies for building robust and generalizable classifiers when the feature dimensionality is high without compromising on the
interpretability of the classifier. We demonstrate the benefit of our approach for constructing classifiers and identifying radiomic features that are useful for (a) diagnosing prostate cancer on multi-parametric MRI, (b) predicting risk of post-radiotherapy prostate cancer biochemical recurrence T2w MRI, and (c) predicting prognosis of patients who underwent therapy for breast and prostate cancers based on digital pathology images. Additionally, we also investigate whether radiomic features characterizing prostate tumors that manifest in the PZ are different from radiomic features characterizing TZ tumors, and we develop a novel approach for pharmacokinetic modeling on dynamic contrast-enhanced MRI of the prostate.

In Chapter 3 we introduce methods to identify useful features for classifier construction in high-dimensional spaces. In Chapters 4-6 we demonstrate the benefit of these methods to overcome the “curse of dimensionality” when building classifiers with high-dimensional features for diagnosing prostate cancer on MRI in both the TZ and PZ, predicting the risk of biochemical recurrence of prostate cancer following radiation therapy based on pre-treatment MRI, and predicting the risk of breast and prostate cancer recurrence based on digital pathology images. Chapter 7 describes a multi-institutional study that identified distinct sets of radiomic features extracted from multi-parametric MRI for diagnosing TZ and PZ tumors, and in Chapter 8 we present a novel method for estimating pharmacokinetic parameters in the prostate since these parameters are beneficial for distinguishing between low, intermediate, and high grade prostate lesions in both the TZ and PZ. The rest of this dissertation discusses the following topics:

- Developing novel strategies for leveraging high-dimensional medical image datasets to construct models for characterizing disease.
• Demonstrating applications of these strategies for (a) identifying radiomic features useful for detecting TZ and PZ prostate cancers on multi-parametric MRI, (b) identifying T2w MRI texture features that are useful for predicting biochemical recurrence risk following radiotherapy for prostate cancer, and (c) identifying QH features useful for predicting treatment outcome and risk assessment of estrogen receptor positive (ER+) breast cancers and risk of biochemical recurrence of prostate cancers following radical prostatectomy.

• Development and multi-institutional evaluation of radiomic feature sets for cancer detection in the TZ and PZ.

• Developing a novel approach for estimating pharmacokinetic parameters in the prostate by leveraging inherent differences between pharmacokinetic characteristics of the TZ and PZ without relying on knowledge of the AIF or any population-averaged values.
Chapter 2

Previous Work and Novel Contributions

2.1 Previous work in constructing interpretable high-dimensional classifiers

A number of groups have applied DR to reduce the dimensionality of high dimensional biomedical data and thereby construct more robust classifiers [4–6, 69–72]. These groups first extract a large number of features from biomedical images or signals, apply DR to obtain a set of low dimensional features, and use the newly derived features as input to a classifier or predictive model. However, to the best of our knowledge, these approaches have not attempted to identify which of the original, high dimensional features contribute most substantially to class discriminability in the transformed, low dimensional feature space.

Several groups have combined feature selection and DR in a single optimization routine by achieving a sparse reconstruction of the high dimensional data [73–75]. Other groups have employed more traditional feature selection routines (e.g.
sequential feature selection) in conjunction with NLDR to preserve the structure of the embedding while eliminating redundant features [76–78]. In both cases the number of high dimensional features contributing to the embedding is significantly decreased. Consequently, this makes the problem of model interpretation in the low dimensional embedding space a little more tractable. However, these approaches do not permit interpretation or quantification of the contributions of the selected features to class discriminability in the reduced dimensional embedding space.

With regard to linear DR variants, a method exists for quantifying the extent that individual, high dimensional features contribute to classification on embeddings obtained via partial least squares (PLS). Chong and Jun introduced the concept of variable importance in projections (VIP), a measure of each feature’s contribution to classification on a low dimensional embedding obtained via PLS [16]. However, similar methods have not been established when the more popular PCA algorithm is used for linear DR, nor for any NLDR scheme.

2.2 Previous work in using image–derived features for cancer characterization

2.2.1 Image–derived texture features for prostate cancer detection and grading

A number of studies have explored methods for computer–aided detection (CAD) and localization of prostate cancer on MRI [6, 21–34, 50, 55, 56, 79–84]. Most
of these studies exploit multiparametric MRI, including T2w, dynamic contrast-enhanced (DCE), and diffusion-weighted (DW) MRI protocols [29, 50]. For example, several groups [82, 83] have reported that combining apparent diffusion coefficient (ADC) maps obtained from DW MRI with T2w MRI improves prostate cancer detection when compared to T2w MRI alone, and Ampeliotis et al. [84] reported a statistically significant improvement in cancer detection accuracy when signal intensities from DCE and T2w MRI were combined, compared to the use of individual MRI protocols. Furthermore, a number of recent studies have shown that radiomic features describing textures on MRI are useful for detecting prostate cancer [6, 22, 25, 55, 56]. These studies use a combination of texture features extracted from T2w MRI and quantitative features extracted from MR spectroscopy, DW MRI, or DCE MRI, and they report improved prostate cancer detection and localization accuracy compared to MRI signal intensities. While most of these studies exploit multiparametric MRI, which is more informative of prostate cancer presence than T2w MRI alone, some have shown that it is possible to construct effective CAD classifiers for prostate cancer localization using T2w MRI alone [22].

More recently a handful of studies have explored using multiparametric MRI for predicting prostate cancer grade, in addition to location and extent [48–51]. Prostate cancers associated with higher Gleason grades have been shown to be correlated with lower T2w MRI signal intensities [48] and lower ADC values on DW MRI [49], and quantitative features extracted from ADC maps and DCE MRI are moderately correlated with Gleason scores [50, 51].

Furthermore, several clinical observational studies have explored the use of pre–treatment MRI to predict biochemical recurrence following treatment [52, 54,
For example, predictive models that exploit attributes on T2w MRI, such as tumor size, presence and extent of extracapsular spread, and presence or absence of seminal vesicle invasion, were shown to outperform the Kattan nomogram in predicting progression to biochemical recurrence following radiation therapy [52]. However, none of these studies used radiomic MRI features to predict risk of biochemical recurrence. Although radiomic texture features have previously been shown to be useful for prostate cancer detection, localization, and grading [6, 22, 25, 29, 48–51, 55, 56, 82–84], to the best of our knowledge no one has considered looking at the association between radiomic texture features on pre–treatment T2w MRI and subsequent biochemical recurrence.

### 2.2.2 Accounting for differences between transition and peripheral zone tumors

Although there has been some recent work on identifying zone specific radiomic features associated with prostate cancer in the TZ and PZ [22], this limited study was specific to a single institution, and the resilience of radiomic features was not evaluated in a cross–institutional setting. This is a particularly important consideration since the variance (or drift) in MRI parameters (T1w, T2w, Diffusion) across vendor platforms and scanners is well-known and documented [62]. Consequently, there is a real need to evaluate in a multi–institutional study whether radiomic features for prostate cancer detection are robust across institutions and whether statistically significant differences in radiomic features between TZ and PZ tumors persist when inter–institutional differences are considered.
2.2.3 Image–derived pharmacokinetic features for prostate cancer detection and grading

There are a number of different techniques for estimating pharmacokinetic parameters on MRI, many of which rely on the Kety-Tofts pharmacokinetic model [86]. Some of these approaches leverage the arterial input function (AIF), which can be measured by introducing a catheter into the patient’s artery and sampling blood at intervals throughout the image acquisition process or estimated in an artery in the image field-of-view [87]. Unfortunately, the former is not desirable due to its invasiveness, and the latter requires that the DCE MRI be acquired at a high temporal resolution, attained at the expense of lowering the spatial resolution. Alternative techniques involve “lifting” a population–averaged AIF from published literature in order to estimate $K_{\text{trans}}$ and $v_e$ in the tumor [88, 89]. However, AIFs vary between subjects, and ignoring the inherent inter–patient variation leads to errors in estimating patient-specific pharmacokinetic parameters [90, 91].

Several methods have been developed to estimate patient–specific pharmacokinetic parameters without relying on a population-averaged AIF [92–94]. Some of these approaches avoid relying on a population–averaged AIF but rely instead on population–averaged pharmacokinetic parameter values for a reference tissue (such as skeletal muscle) [92]. Although this approach leads to more accurate pharmacokinetic parameter estimation in the tumor compared to approaches that rely on a population–averaged AIF, pharmacokinetic parameters vary between subjects even for well–characterized reference tissues [95]. To overcome the limitations of relying on population-averaged values, several groups have developed methods for estimating relative pharmacokinetic parameter values of two contrast
agents that are co–injected [96] or of two tissues within the image field–of–view [97, 98]. Without making any population-based assumptions, these methods provide patient–specific relative pharmacokinetic parameter values [96, 97]. However, the physiological interpretation of these relative pharmacokinetic parameters is not straightforward, and the actual pharmacokinetic parameter values can only be estimated up to a scaling factor via these approaches [96, 97]. Furthermore, these studies used simulations and mouse models for evaluation; they did not evaluate their pharmacokinetic models on human DCE MRI data. To the best of our knowledge, there is no published method for estimating pharmacokinetic parameter values without relying on any population–averaged values that has been evaluated in humans.

### 2.2.4 Image–derived digital pathology features for predicting cancer recurrence

A handful of studies have investigated QH features for computationally modeling the appearance of disease on digital pathology images [9, 63–67, 99, 100]. These groups quantify the texture, color, and spatial arrangement of nuclei and glands on digitized images of tissue slides. These features have been shown to be useful in determining cancer aggressiveness [9, 66, 67] and in predicting the likelihood of a patient’s cancer recurring following treatment [65, 68]. In order to reduce the dimensionality of the feature space and thereby facilitate classifier construction, dimensionality reduction (DR) is often performed [4–7]. Nevertheless, no one has identified which QH features are most predictive of risk of recurrence, knowledge
that could potentially lead to improved disease characterization upon biopsy and better planning of therapy in the adjuvant and neoadjuvant settings.

2.3 Novel contributions

2.3.1 Developing novel approaches for constructing robust, interpretable classifiers in high-dimensional settings

The first goal of this work is to develop and evaluate novel approaches for constructing robust and interpretable classifiers for disease characterization using high-dimensional image features. We introduce Variable Importance in Projection (VIP) (see Chapter 3), a variable ranking scheme to quantify the contributions of individual features to classification on an embedding derived via PCA [21]. This variable ranking scheme exploits the mapping between the original feature space and the data in the PCA embedding space to quantify the contributions of individual features. We then extend VIP to NLDR methods via feature ranking in nonlinear embeddings (FINE). Since VIP exploits the mapping between the original feature space and the data in the embedding space, a mapping that is not defined in the context of KPCA, FINE involves approximating this mapping so that feature importance can be measured in the same manner as VIP. Then, by substituting a similarity matrix in place of the kernel matrix in KPCA, FINE is made applicable to NLDR algorithms such as Isomap [101], Laplacian eigenmaps [102], and locally linear embeddings [103]. In general, the VIP and FINE methods involve performing DR to reduce the dimensionality of high dimensional data, identifying key
features, and using the selected features identified by VIP and FINE to construct a robust classifier.

2.3.2 Identifying zone–specific multi–parametric MRI signatures for prostate cancer characterization

The second goal of this work is to identify radiomic features associated with TZ and PZ tumors on multi–parametric MRI and to evaluate them in a multi–institutional setting. Toward this end, VIP was leveraged to identify stable sets of radiomic features extracted from MRI that are useful for diagnosing prostate cancer in both the TZ and PZ, as well as to identify radiomic features that predict the risk of biochemical recurrence of prostate cancer following radiation therapy with high accuracy. Furthermore, we evaluate the benefit of using distinct sets of radiomic features for detecting TZ and PZ tumors. To the best of our knowledge, this is the first study to evaluate the performance of radiomic features for prostate cancer detection on MRI in a multi–institutional setting.

2.3.3 Developing a novel approach for estimating pharmacokinetic parameters in the prostate

A third goal of this work is to develop an approach to estimate pharmacokinetic parameters associated with prostate cancer without relying on any population–averaged values and to evaluate its ability to distinguish cancer grades on DCE MRI. Our strategy takes advantage of the fact that the prostate is comprised of two primary anatomical regions—the PZ and the TZ—and that quantitative T1 values
and pharmacokinetic parameter values differ between the PZ and TZ [104–106].

Our two-step approach first leverages these inherent differences between the PZ and TZ to estimate relative pharmacokinetic parameter values between these two prostate zones and subsequently estimates voxel-wise pharmacokinetic parameters subject to the constraints induced during the first step of this approach. To the best of our knowledge, there is no published method for AIF–free estimation of pharmacokinetic parameter values without relying on any population–averaged values.
Chapter 3

Methods for Classifier–Building in High–Dimensional Spaces

3.1 Overview

In this chapter we present a general approach for ranking features based on their contributions to accurate classification in a low dimensional embedding space. First we introduce Variable Importance in Projection (VIP) (see Chapter 3), a variable ranking scheme to quantify the contributions of individual features to classification on an embedding derived via PCA [21]. This variable ranking scheme exploits the mapping between the original feature space and the data in the PCA embedding space to quantify the contributions of individual features.

Nevertheless, many high dimensional biomedical datasets have a nonlinear manifold structure. Since PCA relies on Euclidean distances to estimate similarities between features, PCA does not account for the inherent nonlinear structures underlying most biomedical data. Nonlinear DR (NLDR) methods can capture the underlying nonlinear manifold structure of the data, and classifiers modeled on
the NLDR representations have been shown to provide higher class discrimination compared to the corresponding classifiers trained off of linear DR representations [8]. However, whereas PCA involves the eigen decomposition of the data itself, NLDR involves the eigen decomposition of a kernelized form of the data. Furthermore, this kernel need not be explicitly defined. Since there is no closed-form expression for the mapping between the original feature space and the embedding space, the reverse mapping is also not defined. Consequently, compared to PCA, it is even more difficult to reconcile the contributions of the individual features to the low dimensional representations constructed via NLDR schemes.

Therefore, we also present a more general approach than VIP, which we call feature importance in nonlinear embeddings (FINE), for ranking features based on their contributions to accurate classification in a low dimensional embedding space obtained via NLDR. This is accomplished by approximating the mapping between the data in the original feature space and the low dimensional representation of the data obtained by kernel PCA. Once this mapping has been estimated, the contributions of individual features to classification can be computed in the same manner as VIP (see Chapter 3). Furthermore, several NLDR schemes, including Isomap [101] and Laplacian eigenmaps [102], have been shown to be analagous to kernel PCA by simply modifying the kernel function [107]. An illustration of the FINE method is shown in Figure 3.1. In general, the FINE method involves performing DR to reduce the dimensionality of high dimensional data, identifying key features, and using the selected features identified by FINE to construct a robust classifier.
Chapter 3. Methods for Classifier–Building in High–Dimensional Spaces

3.2 Review: Dimensionality Reduction as an Eigenvalue Problem

In this section we review the existing theory showing how several linear and non-linear DR algorithms can be formulated in terms of eigenvalue problems. A list of common mathematical notation used in the coming section is in Table 3.1.
Chapter 3. Methods for Classifier–Building in High-DimensionalSpaces

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$n \times m$</td>
<td>Data matrix</td>
</tr>
<tr>
<td>$X$</td>
<td>$n \times m$</td>
<td>Centered data matrix</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>$n \times m$</td>
<td>Diagonal matrix of singular values</td>
</tr>
<tr>
<td>$U$</td>
<td>$n \times n$</td>
<td>Matrix of eigenvectors</td>
</tr>
<tr>
<td>$V$</td>
<td>$m \times m$</td>
<td>Matrix of right-singular vectors, or loadings</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>$n \times m$</td>
<td>Diagonal matrix of eigenvalues</td>
</tr>
<tr>
<td>$T$</td>
<td>$n \times m$</td>
<td>Principal components matrix</td>
</tr>
<tr>
<td>$Y$</td>
<td>$n \times 1$</td>
<td>Class labels</td>
</tr>
<tr>
<td>$K$</td>
<td>$n \times n$</td>
<td>Kernel matrix</td>
</tr>
<tr>
<td>$Z$</td>
<td>$n \times n$</td>
<td>Eigenvectors of kernel matrix</td>
</tr>
<tr>
<td>$L$</td>
<td>$n \times m$</td>
<td>Graph Laplacian</td>
</tr>
</tbody>
</table>

TABLE 3.1: List of common mathematical notation in this chapter.

3.2.1 Eigenvalue Problems for Dimensionality Reduction

Given a matrix $A \in \mathbb{R}^{n \times m}$, the eigenvectors associated with $A$ can be found by solving an ordinary eigenvalue problem:

$$AU = \lambda U,$$  \hspace{1cm} (3.1)

where the columns of $U$ are the eigenvectors associated with $A$. This eigenvalue problem can be solved by using an iterative technique or the singular value decomposition. According to the singular value decomposition,

$$A = U\Sigma V^T,$$ \hspace{1cm} (3.2)

where the columns of $U \in \mathbb{R}^{n \times n}$ are the left-singular vectors, or eigenvectors of the covariance matrix $A^T A$; the columns of $V \in \mathbb{R}^{m \times m}$ are the right-singular vectors, or eigenvectors of the Gram matrix $AA^T$; and $\Sigma \in \mathbb{R}^{n \times m}$ is a diagonal matrix whose diagonal entries are the singular values of $A$, or the square roots of the eigenvalues.
shared by both the covariance and Gram matrices.

Many DR algorithms, including PCA, PLS, Isomap, and their kernelized versions, involve solving an eigenvalue problem to find an optimal subspace to embed the data. Below we review how several popular DR methods can be formulated as eigenvalue problems [107].

**Principal Components Analysis**

Consider a centered data matrix $X \in \mathbb{R}^{n \times m}$, whose rows represent $n$ samples and whose columns represent $m$ features. PCA involves solving eq. (3.1) in the case that $A = X$. This problem can be solved by the singular value decomposition as

$$X = U\Sigma V^\top$$  \hspace{1cm} (3.3)

or, alternatively, the eigen decomposition of $X^\top X$:

$$X^\top X = U\Lambda U^\top.$$  \hspace{1cm} (3.4)

Here $\Lambda$ is a diagonal matrix whose diagonal entries are the eigenvalues of $X^\top X$. Letting $T = U\Sigma$, the principal components of $X$ are the columns of $T$, the scores matrix containing the projection of the original data into the PCA embedding space. $V$ is the loadings matrix, whose elements describe the correlation between the scores and the original features. Thus,

$$X = TV^\top,$$  \hspace{1cm} (3.5)

and $T = XV$ since $V^\top V = I$. 

When DR is desired, only a subset of the eigenvectors are used to reconstruct the data. Thus,

\[ X \approx T_h V_h^T, \]  

(3.6)

where \( h \in \mathbb{N}, h < m, T_h \in \mathbb{R}^{n \times h}, \) and \( V_h \in \mathbb{R}^{m \times h}. \)

**Kernel PCA**

Whereas both PCA and PLS only take into account linear relationships among features, kernel PCA (KPCA) involves a nonlinear mapping of the data. KPCA involves the eigen decomposition of a centered kernel matrix \( K, \) as follows:

\[ K = Z \Lambda Z^T. \]  

(3.7)

Thus, the principal components of \( K \) are contained in

\[ T = KZ^{\frac{1}{2}}. \]  

(3.8)

In KPCA the kernel matrix \( K \) does not need to be explicitly defined. Rather, \( K \) can be defined in the dual form:

\[ K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle. \]  

(3.9)

Consequently, \( K \) can represent any kernel, such as a Gaussian or radial basis function kernel, or a distance or similarity matrix.
3.2.2 NLDR Methods as Variants of KPCA

Many NLDR algorithms, including Isomap, Laplacian eigenmaps, and locally linear embeddings (LLE), involve the eigen decomposition of a similarity matrix. Here we review how these three NLDR algorithms can be formulated as variants of KPCA \[107\].

**Isomap**

Isomap \[101\] involves creating a neighborhood graph by using $K$ nearest neighbors or $\epsilon$-neighborhoods to determine a set of neighboring points associated with each datapoint. Then, Isomap estimates the geodesic distances $d(i, j)$ between points $i$ and $j$ on the graph by computing the shortest path between points. Thus, we obtain the kernel matrix $K$ whose elements are

$$K_{ij} = d(i, j). \quad (3.10)$$

Since all elements of $K$ are positive, $K$ must be centered before Isomap can be solved by eq. (3.7) as a variant of KPCA.

**Laplacian Eigenmaps**

Given a neighborhood graph constructed using $K$ nearest neighbors or $\epsilon$-neighborhoods, Laplacian eigenmaps (LE) \[102\] determines an optimal manifold representation of the data by computing the graph Laplacian $L$:
Laplacian eigenmaps is equivalent to KPCA such that $K = L^\dagger$, where $\dagger$ denotes the pseudo-inverse.

**Locally Linear Embeddings**

LLE [103] constructs a neighborhood-preserving mapping by computing the Euclidean distances $D_{ij}$ between points and their nearest neighbors, creating a local covariance matrix $C$:

$$C_{ij} = \frac{1}{2}(D_i + D_j - D_{ij} - \sum_{ij} D_{ij}). \quad (3.11)$$

LLE is equivalent to KPCA such that

$$K = \lambda_{max} I - M. \quad (3.12)$$

Here $I$ is an identity matrix and $M = (I - W)(I - W^T)$, where the $w_i = \frac{\sum_j C^{-1}_{ij}}{\sum_{ij} C^{-1}_{ij}}$ are the linear coefficients that optimally reconstruct $x_i$ from its nearest neighbors.
3.3 Calculating Feature Importance

In this section we demonstrate how the variable importance in projection (VIP) score allows for feature weighting and ranking based on their contributions to classification within a linearly-derived embedding. Then we discuss our extension to VIP that enables the computation of FINE scores when the embedding is derived using an NLDR scheme.

3.3.1 Variable Importance in Projections (VIP)

The importance of an individual feature to classification on an embedding depends on two factors: how much each eigenvector contributes to the embedding, and how much each feature contributes to each eigenvector. All of the DR algorithms discussed in Section 3.2 provide a projection matrix $\mathbf{T}$ and a loadings matrix $\mathbf{V}$. The features that contribute most to the $i$th dimension of the embedding are those with the largest weights in the $i$th loading vector. Thus, the fraction $\left(\frac{v_{ji}}{||v||}\right)^2$ reveals how much the $j$th feature contributes to the $i$th principal component in the low-dimensional embedding. The overall importance of the $j$th feature depends also on (a) the regression coefficients $b_i$, which relate the transformed data back to the class labels, and (b) the transformed data $t_i$. The variable importance in projections (VIP) score is computed for each feature as follows:

$$\pi_j = \sqrt{\frac{\sum_{i=1}^{h} b_i^2 t_i^T t_i \left(\frac{v_{ji}}{||v||}\right)^2}{\sum_{i=1}^{h} b_i^2 t_i^T t_i}},$$

(3.13)
where \( m \) is the number of features in the original, high-dimensional feature space, \( t_i \) is the \( i \)th principal component vector, and \( v_i \) is the \( i \)th loading vector. The \( b_i \) are the coefficients that solve the regression equation

\[
y = Tb^T,
\]

which correlates the scores with the outcome vector \( y \).

For DR methods that are intrinsically unsupervised, the exploitation of class labels in computing the VIP score leads to the identification of features that provide good class discrimination in the embedding space. The degree to which a feature contributes to classification in the transformed space is directly proportional to its associated VIP score. Thus, features with VIP scores near zero have little predictive power, and the features with the highest VIP scores contribute the most to class discrimination on the embedding.

Alternatively, VIP scores can be normalized to the interval \([0, 1]\) as follows:

\[
\hat{\pi}_j = \frac{\pi_j^2}{m}.
\]

When the VIP scores are normalized in this way, \( \sum_{j=1}^{m} \hat{\pi}_j = 1 \). Consequently, \( \hat{\pi}_j \) is the fraction that feature \( j \) contributes to classification in the embedding space compared to the entire feature set. Furthermore, the aggregate VIP score associated with a feature subset \( \mathcal{J} \subset \{1, ..., m\} \) can be calculated as

\[
\hat{\pi}_{\mathcal{J}} = \frac{1}{m} \sum_{j \in \mathcal{J}} \pi_j^2.
\]
3.3.2 Feature Importance in Nonlinear Embeddings (FINE)

The expression for computing the importance of features according to VIP (3.13) relies on the loadings matrix $V$. For KPCA and its variants (e.g., Isomap, Laplacian eigenmaps, and locally linear embeddings) $V$ is not defined because the transformed data is not directly related to the original data, but only to the kernel matrix (see eq. (3.8)). Furthermore, the mapping $\Phi : X \rightarrow K$ that relates the kernel matrix to the original data is not necessarily computed, as $\Phi$ is only implicitly defined.

Nevertheless, combining equations (3.5) and (3.8) yields

$$X = KZ^{\frac{1}{2}}V^T,$$  \hspace{1cm} (3.17)

Thus, without explicit knowledge of $\Phi$, we can estimate $V$ as follows:

$$V^T \approx (KZ^{\frac{1}{2}})^\dagger X.$$  \hspace{1cm} (3.18)

Approximating $V$ as $V' = X^T((KZ^{\frac{1}{2}})^\dagger)^T$ facilitates the computation of the feature importance in nonlinear embeddings (FINE) using eq. (3.13).

3.4 Evaluation of FINE on publicly available datasets

3.4.1 Datasets

In order to evaluate FINE in terms of its ability to identify a feature subset that (a) is stable and (b) provides good classification accuracy, we chose the NIPS 2003
Table 3.2: Description of four publicly available datasets used in this paper.

Feature Selection Challenge datasets because they all suffer from the “curse of dimensionality”. All of these datasets have been made publicly available as benchmarking datasets for feature selection algorithms. The NIPS 2003 Feature Selection Challenge included five datasets, all involving binary classification problems; we used the four datasets that contain non-binary features. More details regarding these four datasets can be found in Table 3.2.

### 3.4.2 Experimental Design

#### Dimensionality Reduction

FINE was implemented to identify key contributors to classification on embeddings obtained via four DR approaches: PCA, Isomap, LE, and LLE. Some of the publicly available datasets evaluated in this paper contained hundreds or thousands of instances. As a result, it was necessary to randomly sample a small number of instances to construct the embeddings, which can be computationally intractable when the number of instances is too high. Consequently, 50 rounds of bootstrapping were performed. For each of the NIPS 2003 Feature Selection Challenge datasets 75% of the instances, up to a maximum of 100 instances, were randomly selected during each round of bootstrapping to construct the embedding.
and tune parameters. Parameters associated with these DR methods—the intrinsic dimensionality parameter $h$, which is used in all three embeddings; $K$, which is used to create a neighborhood graph for Isomap, LE, and LLE; and $\sigma$, which is needed to compute the graph Laplacian for LE—were chosen to reduce residual variance [101]:

$$1 - \rho(D^E, D^G).$$

(3.19)

Here $D^E$ is a matrix of Euclidean distances between points in the low-dimensional embedding, $D^G$ is the DR algorithm’s estimate of point-wise distances (e.g., for Isomap $D^G = K$), and $\rho$ denotes the linear correlation coefficient. Finally, PCA, Isomap, LE, and LLE embeddings were constructed using the parameter values chosen by minimizing eq. (3.19).

**Classifier Training**

Once embeddings were constructed, FINE scores were computed, and the features associated with the highest FINE scores were selected. Because a samples-to-features ratio of 10:1 is generally recommended to build a robust classifier [3], the number of selected features was limited to one tenth of the number of samples in the training dataset. Then, within the same boostrapping rounds, classifiers were trained in conjunction with the features selected by FINE. It is important to note that the objective of FINE is not merely to select features that maximize classification accuracy, but rather to identify a stable set of features that also provide good class discriminability. Consequently, our focus was less on constructing the most
accurate classifier and more on identifying features that could yield stable and reproducible classifiers. The logistic regression classifier employs linear regression when the outcome variable is binomially distributed; thus, it allows the features themselves to drive classification. Due to its simplicity, the logistic regression classifier was chosen to evaluate classifier accuracy associated with features selected by FINE. Additionally, we evaluated the classification accuracy associated with the selected features in conjunction with two other classifiers: a support vector machine (SVM) with a radial basis function kernel (scale factor = 1) and a random forest (RF) with 50 trees. Please refer to Appendix A for more details regarding these classification strategies.

**Performance Evaluation Measures**

The feature selection performance of FINE was evaluated in comparison to three filter methods commonly used for feature selection: the \( t \)-test, Fisher score, and Gini index [108]. Evaluation was performed to assess (a) stability and (b) classification accuracy associated with selected feature subsets. Stability was evaluated by the Jaccard index [109], which measures the degree of overlap between feature sets:

\[
J(J_1, J_2) = \frac{|J_1 \cap J_2|}{|J_1 \cup J_2|},
\]

where \( J_1 \subset \{1, \ldots, m\} \) and \( J_2 \subset \{1, \ldots, m\} \) are two feature subsets. If a feature selection algorithm is stable, repeated implementations will lead to similar feature subsets and \( J \) close to one.
Table 3.3: Average value of the intrinsic dimensionality parameter $h$, selected via residual variance minimization, for four NIPS datasets and four dimensionality reduction algorithms.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$h_{\text{PCA}}$</th>
<th>$h_{\text{ISO}}$</th>
<th>$h_{\text{LE}}$</th>
<th>$h_{\text{LLE}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madelon</td>
<td>36</td>
<td>4</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>Arcene</td>
<td>15</td>
<td>5</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Dexter</td>
<td>6</td>
<td>2</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>Gisette</td>
<td>6</td>
<td>2</td>
<td>14</td>
<td>47</td>
</tr>
</tbody>
</table>

Additionally, classification accuracy associated with feature subsets selected by the FINE method was evaluated in conjunction with LR, SVM, and RF classifiers. Based on the prediction results obtained from these classifiers, receiver operating characteristic (ROC) curves were generated, and the area under the ROC curve (AUC) was used to evaluate classifier accuracy in conjunction with different feature subsets. The classifiers were evaluated on a dedicated validation set provided by the NIPS 2003 Feature Selection Challenge.

### 3.4.3 Experimental Results and Discussion

**Madelon**

Figure 3.2 displays the AUC and $J$ values associated with feature subsets selected by the $t$-test, Fisher score, Gini index, and FINE. Feature subsets obtained via the $t$-test, Fisher score, and Gini index were highly unstable ($J = 0.04$ in all cases), varying considerably between rounds of bootstrapping. In contrast, $J$ associated with FINE$_{\text{PCA}}$, FINE$_{\text{Isomap}}$, and FINE$_{\text{LE}}$ was slightly higher (0.06–0.07), and the features obtained via FINE$_{\text{LLE}}$ were moderately stable, yielding $J = 0.22$. Regardless of classification method (LR, SVM or RF), the top ranking features selected
Figure 3.2: AUC and Jaccard index associated with four dimensionality reduction algorithms, the top five features selected based on FINE scores, and three common feature selection algorithms for NIPS datasets (a) Madelon, (b) Arcene, (c) Dexter, and (d) Gisette. Data points are shown for three classification methods: logistic regression (black outline), support vector machine (green outline), and random forest (red outline) classifiers.
by FINE\textsubscript{PCA}, FINE\textsubscript{Isomap}, FINE\textsubscript{LE} and FINE\textsubscript{LLE} outperform the embedding vectors themselves, providing AUC values between 0.55 and 0.65 while leveraging only 13–16 features. In contrast, the AUC values yielded by feature sets selected via the Gini index, Fisher score, or \textit{t}–test ranged from 0.54 to 0.56, with the highest AUC values relying on 97 features.

**Arcene**

In spite of the high dimensionality of this protein expression dataset, all feature selection methods provided the same perfect stability ($J = 1$), although the top-ranking features were different for each feature selection algorithm. When the embedding vectors were used in conjunction with classifiers, PCA provided AUC values as high as 0.77–0.79, higher than any other method. However, features selected based on FINE\textsubscript{PCA} were no better than random guessing. Whereas the Gini index and \textit{t}–test provided AUC values between 0.69 and 0.77, the Fisher score, FINE\textsubscript{Isomap}, and FINE\textsubscript{LE} yielded AUC values between 0.6 and 0.7.

**Dexter**

Feature subsets selected via all four FINE methods and the Fisher score were highly unstable, probably due to the fact that this is a sparse dataset with 20,000 features. Surprisingly, both the Gini index and the \textit{t}–test yielded fairly stable datasets ($J = 0.39$ and $J = 0.13$, respectively). Due to the constraint of a 10:1 samples–to–features ratio, a maximum of 30 features were selected. As a result, AUC values ranged between 0.5 and 0.55 for almost all feature selection algorithms. The only exception
was the Gini index, which led to feature sets associated with AUC values as high as 0.87–0.92, depending on the classifier used.

**Gisette**

The feature subset selected based on FINE\(_{LE}\) was associated with \(J = 0.99\), whereas feature subsets selected via FINE\(_{PCA}\), FINE\(_{Isomap}\), and FINE\(_{LLE}\), as well as the Fisher score, were highly unstable. When up to 300 features were selected, AUC values ranging from 0.91 to 0.95 were achieved for all four FINE methods, as well as for the Gini index, Fisher score, and \(t\)-test. Regardless of the DR algorithm, FINE-based feature subset selection led to higher AUC values than the embedding vectors themselves.
Chapter 4

Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI

4.1 Overview

In this chapter we investigate the ability of VIP to identify radiomic features describing prostate cancer morphology on multi–parametric MRI that are useful for detecting prostate cancer in both the TZ and PZ. VIP is leveraged to identify sets of radiomic features that collectively characterize prostate cancer appearance on multi–parametric MRI, and we evaluate the performance of these features within a decision support classifier for cancer detection and localization. Prostate cancer localization accuracy is assessed by comparing voxel–wise classification results with “ground truth” cancer extent, delineated by a pathologist on ex vivo histology and mapped to in vivo MRI via nonlinear registration. A secondary goal of this work is to investigate zonal differences in prostate cancer characteristics on
Chapter 4. Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI

4.2 Materials and methods

4.2.1 Data description

This retrospective study was approved by the Institutional Review Board at Beth Israel Deaconess Medical Center. This study was an arm of a prospective study that included 108 subjects (mean age, 58.5 years; age range, 47–72 years) with biopsy-confirmed prostate cancer (median Gleason score, 7; range, 6–9) who were scheduled for radical prostatectomies. Forty-five subjects were excluded for lack of DCE MRI or corresponding digitized whole mount histological sections (WMHS), and another 40 were excluded because the pathologist did not annotate cancer on the WMHS. A total of 23 cases were included in the current study, of which 15 cases included T2w, DWI, and DCE MRI, while the remaining eight cases included only T2w and DCE MRI.

Before surgery, the patients were imaged using a combined torso–phased array and endorectal coil (MedRad, Pittsburgh, PA) using a 3T whole-body MRI scanner (Genesis Signa LX Excite; GE Medical Systems, Milwaukee, WI). The parameters for axial T2w MRI were repetition time / echo time (TR/TE) = 6375/ 165 msec, slice thickness of 1.5–2 mm (no gap between slices), and matrix size of 320 x 224–192 voxels with a field of view (FOV) of 12 x 12 cm. The DCE MRI protocol included two precontrast T1–weighted gradient echo images, acquired at 95–second intervals before a bolus injection of 0.1 mmol/kg of gadolinium-DTPA, and five
postcontrast images acquired at the same temporal resolution. The DCE MRI parameters were TR/TE = 9.3/4.2 msec, flip angle = 18° , FOV = 14 x 14 cm and matrix size 256 x 224 (interpolated to 256 x 256 matrix), with no phase wrap. Transverse DWI parameters were TR/ TE = 6500/80.6 msec, FOV = 24 x 24 cm, matrix size 256 x 192, B-value = 0,1000 s/mm², two averages and 25 directions. Instead of using more averages, we used 25 directions to improve results in diffusion tensor imaging and anisotropic maps and enhance contrast in ADC maps.

A pathologist and radiologist working in unison visually identified 96 corresponding 2D WMHS and axial MRI slices from these 23 studies. These correspondences were established by means of anatomical fiducials such as the urethra, veromontanum, and prominent nodules of benign prostatic hyperplasia that were visually discernible on both histology and MRI. Based on the recommendations of McNeal [110], each patient was classified as having TZ or PZ cancer if more than 70% of the cancer volume was present in a particular zone. To ensure that the sets of TZ and PZ cancer were distinct from each other, only sections displaying an explicit tumor focus in either the TZ or the PZ were included in this analysis. Based on the zonal locations of the prostate tumors and the MRI protocols available for each patient, four sets of patient studies were composed and analyzed separately in this study (see Table 4.1).
4.2.2 Postprocessing of MRI

T2w and DCE MRI were corrected for acquisition-based MRI intensity artifacts that affect image analysis algorithms [111]. The most significant artifact is the bias field on T2w and DCE MRI, which occurs due to usage of an endorectal probe [112]. Bias field artifacts were corrected by the N3 algorithm [113], which incrementally deconvolves smooth bias field estimates from acquired image data (see Fig. 4.1; see Appendix B for more details regarding correction of the bias field artifact). A second artifact, intensity nonstandardness [62], refers to the issue of inter- and intra-patient MRI “intensity drift,” which causes MRI intensities to lack tissue-specific numeric meaning [62]. This artifact was corrected by interactive implementation of the generalized scale algorithm [62], which aligns image intensity histograms across different MRI studies, thereby enabling MRI intensities to have a consistent tissue-specific numeric meaning (see Fig. 4.1; see Appendix C for more details regarding the intensity standardization procedure).

4.2.3 Registration of MRI and WMHS slices

In order to obtain “ground truth” annotation of prostate cancer extent on in vivo MRI, nonlinear registration of multi-parametric MRI and WMHS was performed [114, 115]. First, T2w MRI and ADC maps obtained from DWI were spatially aligned with DCE MRI via volumetric affine registration, which corrected inter-acquisition movement and interprotocol resolution differences. Slice correspondences between T2w, DCE, and ADC images, as well as relative voxel locations and sizes, were determined using DICOM image header information. After inter-protocol alignment, all MRI data were analyzed at the DCE MRI resolution. Once
Figure 4.1: A T2w MR image is shown (a) before and (e) after correction of bias field inhomogeneities. Signal intensity histograms associated with several patients are shown before and after intensity standardization for T2w MRI in (b) and (f), ADC maps in (c) and (g), and DCE MRI in (d) and (h), respectively.
T2w, DCE, and ADC images were spatially aligned, \textit{in vivo} MRI was registered with \textit{ex vivo} WMHS. Registration of WMHS and MRI is complicated by differences in image intensities and nonlinear changes in the shape of the prostate due to both the endorectal coil and deformations to the histological sections upon fixation and sectioning. We therefore used a nonrigid registration scheme [114] driven by a higher-order variant of mutual information that handles images with very different intensities (e.g., MRI and WMHS data) and deformation characteristics (e.g., \textit{in vivo} to \textit{ex vivo}). First, affine alignment of WMHS to the corresponding T2w MRI slice was performed to correct large translation, rotation, and scale differences. Then the rigid alignment of WMHS and T2w MRI was improved by means of a fully automated nonlinear hierarchical B-spline mesh grid image warping scheme [114].

After corresponding 2D WMHS slices were aligned with MRI, the spatial extent of the cancer was mapped from WMHS slices onto corresponding MRI slices. The spatial extent of the cancer mapped onto MRI was examined and manually corrected (as required) by a radiologist using Photoshop (Adobe Systems, San Jose, CA). The final result was a labeling of each MRI voxel within the prostate as corresponding to cancer or benign prostate tissue (see Fig. 4.2).

### 4.2.4 Feature extraction from multi-parametric MRI

Signal intensity (SI) features considered in this study included T2w MRI intensities, time-resolved DCE MRI signal intensities and tissue concentration curves, estimated using the method of Medved et al [116], and ADC values for the studies with available ADC maps (see Table 4.2). Additionally, we computed a large number of radiomic features that have previously been used for prostate cancer
Chapter 4. Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI

Figure 4.2: Checkerboard images illustrating registration of T2w MRI and DCE MRI are shown in (a) and (f) for patients with PZ and TZ cancer, respectively; registration of T2w MRI and ADC maps is shown in (b) and (g). The original 2D whole–mount histological sections, with prostate cancer extent outlined in blue by a pathologist, are shown in (c) and (h), and overlays of whole–mount histological sections and T2w MRI after nonlinear multimodal registration are shown in (d) and (i). Mapped cancer extent on T2w MRI, outlined in red, is shown in (e) and (j).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Amount</th>
<th>Significance</th>
<th>VIP (PZ)</th>
<th>VIP (TZ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2w MRI SI</td>
<td>1</td>
<td>T2w MRI SI is low in tumors relative to surrounding normal tissue</td>
<td>0.68</td>
<td>0.60</td>
</tr>
<tr>
<td>ADC</td>
<td>1</td>
<td>ADC values are low in tumors due to increased cell density and tissue disorganization in tumors</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td>DCE MRI SI</td>
<td>6</td>
<td>High microvessel density in tumor regions leads to increased contrast enhancement compared to normal peripheral tissue</td>
<td>Total: 2.39</td>
<td>Total: 2.28</td>
</tr>
<tr>
<td>Contrast agent concentration</td>
<td>6</td>
<td>Due to increased angiogenesis, contrast agent uptake is faster in tumors than surrounding normal tissue</td>
<td>Total: 0.82</td>
<td>Total: 0.76</td>
</tr>
</tbody>
</table>

Table 4.2: Description of multi–parametric MRI signal intensity features used for TZ and PZ prostate cancer localization and their associated VIP scores. SI, signal intensity
Chapter 4. Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI

<table>
<thead>
<tr>
<th>Feature</th>
<th>Amount</th>
<th>Description</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>First–order statistics</td>
<td>14 T2w + 14 ADC</td>
<td>Mean, standard deviation, range of intensities; non–steerable gradient features obtained by convolution with Sobel and Kirsch operators</td>
<td>Localize image regions with significant SI changes, accurately detect region boundaries</td>
</tr>
<tr>
<td>Co–occurrence</td>
<td>36 T2w + 36 ADC</td>
<td>Statistical features computed from the joint probability distribution of intensity value co–occurrences</td>
<td>Differentiate between homogeneous regions of low SI in prostate cancer and hyper–intense normal prostate</td>
</tr>
<tr>
<td>Gabor wavelet</td>
<td>48 T2w + 48 ADC</td>
<td>Multi–orientation features computed from a Gaussian function convolved with a sinusoid</td>
<td>Quantify features qualitatively assessed by radiologists examining cancer appearance</td>
</tr>
<tr>
<td>Haar wavelet coefficients</td>
<td>12 T2w + 12 ADC</td>
<td>Multi–level coefficients from Haar wavelet decomposition</td>
<td>Accentuate amorphous nature of non–cancer regions with foci of low SI</td>
</tr>
</tbody>
</table>

**Table 4.3:** Description of features computed on a per–voxel basis from T2w MRI and ADC maps and the motivation for using them for prostate cancer localization

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to peak</td>
<td>The interval between the pre–contrast time point and post–contrast time point at which the lesion achieved maximum signal intensity</td>
</tr>
<tr>
<td>Maximum uptake</td>
<td>The signal intensity associated with the time to peak</td>
</tr>
<tr>
<td>Uptake rate</td>
<td>The rate of change in signal intensity over the interval between the first time point and the time to peak</td>
</tr>
<tr>
<td>Washout rate</td>
<td>The rate of change in signal intensity over the interval between the time to peak and the time point at which lowest signal enhancement is achieved</td>
</tr>
<tr>
<td>Enhancement</td>
<td>The signal intensity at the first post–contrast time point</td>
</tr>
<tr>
<td>Enhancement ratio</td>
<td>The ratio of enhancement to maximum uptake</td>
</tr>
</tbody>
</table>

**Table 4.4:** Description of contrast kinetic features computed on a per–voxel basis from DCE MRI

detection on T2w MRI [22]. These 112 features, which we extracted on a per–voxel basis from both T2w MRI and ADC maps, are described briefly in Table 4.3. Additionally, six kinetic features [117] were obtained for each DCE MRI voxel (Table 4.4). All feature extraction was implemented using Matlab (MathWorks, Natick, MA).

### 4.2.5 Evaluation

VIP was evaluated in terms of its ability to find a subset of features that 1) remains stable (i.e., robust to perturbations in the data) and 2) provides high accuracy in
cancer localization. The stability of VIP was evaluated by the Jaccard index $J$ (3.20). If a feature selection algorithm is stable, repeated implementations will lead to similar feature subsets and a Jaccard index close to one. In this study we compared the stability of VIP with minimum-redundancy-maximum relevance (mRMR) [14], a popular feature selection scheme. Classification accuracy was evaluated by using two Bayesian classifiers [15]: the parametric logistic regression classifier and the nonparametric naive Bayes classifier (see Appendix A for more details). Based on the voxel-wise prediction results obtained from these classifiers, receiver operating characteristic (ROC) curves representing the tradeoff between cancer detection sensitivity and specificity were generated. The area under the ROC curve (AUC) was used to evaluate classifier accuracy in conjunction with different feature subsets. In order to ensure robustness of AUC estimates, a 3-fold randomized cross-validation procedure, using approximately 2/3 of the patient studies for training and the remaining 1/3 for testing, was repeated 30 times, and the average AUC and $J$ values were obtained.

4.2.6 Relative importance of feature subsets

The importance of a feature subset $J$ is quantified as:

$$
\phi_J = \sqrt{\sum_{j \in J} (\phi_j^3)}.
$$

(4.1)
It follows that the relative importance of subset $J_1$ compared to subset $J_2$ can be expressed as:

$$\bar{\phi}_{J_1/J_2} = 100 \times \left( \frac{\pi_{J_1}}{\pi_{J_1 \cup J_2}} \right). \quad (4.2)$$

### 4.3 Results

#### 4.3.1 Features selected by VIP

Table 4.5 lists the 10 features with the highest VIP scores for PZ and TZ cancer localization; these features constituted $J_{PZ}$ and $J_{TZ}$, respectively. Note that two features are shared by $J_{PZ}$ and $J_{TZ}$: the enhancement ratio from DCE MRI and a Haar wavelet coefficient derived from DWI. Whereas $J_{TZ}$ included eight co-occurrence features extracted from T2w MRI, $J_{PZ}$ contained mainly Gabor and Haar wavelet features and no co-occurrence features. Half of the features in $J_{PZ}$ were derived from DWI, three from T2w MRI and two from DCE MRI: enhancement ratio and time–to–peak.

#### 4.3.2 Stability of VIP feature subsets

Figure 4.3 displays the Jaccard index for feature subsets selected by VIP and mRMR. For all four datasets, the Jaccard index was higher for VIP than mRMR. For datasets $S_2$, $S_3$, and $S_4$, this difference was statistically significant. In particular, the Jaccard index associated with VIP for TZ cancer localization (dataset $S_3$) was as high as 0.68, signifying that the feature subset selected by VIP for TZ cancer localization was highly stable and relatively unaffected by perturbations in the data.
Chapter 4. Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MRI Protocol</th>
<th>Feature</th>
<th>VIP Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PZ tumors</td>
<td>T2w MRI</td>
<td>Haar: Horizontal coefficient (level 4)</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gabor ($\theta = 90^\circ, \lambda = 45.3$)</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haar: Horizontal coefficient (level 3)</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td>DWI MRI</td>
<td>Haar: Horizontal coefficient (level 4)</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haar: Horizontal coefficient (level 3)</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gabor ($\theta = 22.5^\circ, \lambda = 45.3$)</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gabor ($\theta = 90^\circ, \lambda = 22.6$)</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gabor ($\theta = 90^\circ, \lambda = 45.3$)</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enhancement ratio</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time to peak</td>
<td>1.67</td>
</tr>
<tr>
<td>TZ tumors</td>
<td>T2w MRI</td>
<td>Co–occurrence: Inertia ($w = 7$)</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Difference average ($w = 7$)</td>
<td>2.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Difference variance ($w = 7$)</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Difference average ($w = 5$)</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Difference variance ($w = 5$)</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Inverse difference moment ($w = 5$)</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Inverse difference moment ($w = 7$)</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Co–occurrence: Inertia ($w = 5$)</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>DWI MRI</td>
<td>Haar: Horizontal coefficient (level 4)</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>DCE MRI</td>
<td>Enhancement ratio</td>
<td>2.83</td>
</tr>
</tbody>
</table>

Table 4.5: The top 10 radiomic features

Figure 4.3: The Jaccard index, used to assess the stability of selected feature subsets, is compared when VIP and mRMR are employed to select the top 10 features.
4.3.3 Classification accuracy provided by selected features

AUC values obtained by using $J_{TZ}$ and $J_{PZ}$ in conjunction with logistic regression and naive Bayes classifiers followed similar trends (see Fig. 4.4(c,d)), although the logistic regression classifiers generally provided higher AUC values than the naive Bayes classifier. Figure 4.5(f,l), respectively, illustrate the accuracy with which logistic regression classifiers constructed in conjunction with $J_{PZ}$ and $J_{TZ}$ localize cancer on MRI. In contrast to classifiers using all of the radiomic features (see Fig. 4.5(d,j)), the classifiers that used only the features selected by VIP accurately identified both the tumor location and the approximate size and shape of the lesion. When $J_{TZ}$ was used in conjunction with logistic regression to classify voxels as cancer or benign, an AUC value of 0.85 was achieved (Fig. 4.4(d)); when a logistic regression classifier was used in conjunction with $J_{PZ}$, an AUC of 0.79 was obtained (Fig. 4.4(d)). In contrast, when features extracted from DWI were not available (dataset $S_1$), the average AUC using the entire $J_{PZ}$ dropped to 0.75. When VIP was used to select features for cancer detection independent of zonal prostate location, AUC values were substantially reduced (Fig. 4.4(c,d)).

4.3.4 Comparison of signal intensity and radiomic features

The VIP scores associated with each of the SI features are listed in Table 4.2; none of the SI features was associated with VIP scores greater than one, indicating that they contributed less to prostate cancer localization than the feature with the mean VIP score. In contrast, VIP scores associated with the top–performing radiomic features, listed in Table 4.5, ranged from 1.5 to 2.83. Figures 4.5(b) and 4.3(f) illustrate that logistic regression classifiers constructed in conjunction with SI features...
Figure 4.4: AUC values, averaged across 30 cross-validation runs, are plotted for each of the top 10 radiomic features for (a) a naive Bayes classifier and (b) a logistic regression classifier. The AUCs that arise when aggregating the top 1–10 radiomic features for classification are shown in (c) for a naive Bayes classifier and in (d) for a logistic regression classifier.
Chapter 4. Applications of VIP for Zone–Specific Prostate Cancer Detection on Multi-Parametric MRI

Figure 4.5: Ground truth extent of prostate cancer is delineated on T2w MRI in red for a representative slice of (a) a PZ tumor and (b) a TZ tumor. Corresponding DCE MR images at peak contrast enhancement and ADC maps are shown in (b) and (h) and in (c) and (i), respectively. Heatmaps representing the pixel–wise probability of cancer presence, obtained via logistic regression classification, are shown in (d) and (j) using all of the radiomic features, (e) and (k) using signal intensity features, and (f) and (l) using $J_{PZ}$ and $J_{TZ}$, respectively. Red indicates a high probability of cancer presence, yellow indicates a low probability of cancer presence, and blue indicates no cancer presence.

allowed for localization of PZ and TZ cancer, respectively, with high sensitivity but poor specificity. In contrast, logistic regression classifiers constructed in conjunction with $J_{TZ}$ and $J_{PZ}$ localized cancer with high sensitivity and specificity (Fig. 4.5(f,l)).

4.3.5 Ranking MRI protocols for prostate cancer localization

The relative contributions of T2w, DWI, and DCE MRI features to cancer localization were assessed using eq. (4.2). For TZ cancer localization T2w MRI features contributed nearly twice as much to classification accuracy in the PCA embedding
space as DWI features (Fig. 4.6(a)). When only features comprising $J_{TZ}$ were considered (Fig. 4.6(b)), T2w MRI features contributed almost 70% of the classification performance, with DCE MRI features contributing approximately 20% and DWI features contributing approximately 10% (i.e., each of the 10 features contributes approximately 10%). For PZ cancer localization, however, features extracted from both T2w MRI and ADC maps had nearly identical contributions to classification accuracy in the PCA embedding space (dataset $S_2$, Fig. 4.6(a)). When only features comprising $J_{PZ}$ were considered (Fig. 4.6(b)), DWI MRI features contributed almost 50% of the classification performance, with DCE MRI features contributing approximately 20% and T2w features contributing nearly 30%.
4.4 Discussion

4.4.1 Features selected by VIP

The types of features identified by VIP as useful for cancer localization were found to depend on the spatial location of the tumor in the prostate. For example, while $J_{TZ}$ was comprised primarily of co-occurrence texture features and included only one wavelet feature, $J_{PZ}$ contained many Gabor and Haar wavelet features but no co-occurrence features. The importance of these steerable filters suggests that perhaps PZ tumors manifest a textural orientedness that is not manifested by TZ tumors. Among the features extracted from DCE MRI, time-to-peak was found to be useful for cancer localization in the PZ but not in the TZ. In contrast, the signal enhancement ratio, which was previously shown to be informative for breast cancer localization [118], was one of only two features identified by VIP as useful for tumor localization in both the TZ and the PZ. It is important to note that although time-to-peak was not included in $J_{TZ}$ (because co-occurrence features extracted from T2w MRI had higher VIP scores), the VIP score associated with time-to-peak for TZ tumor localization (1.86) is higher than the VIP score associated with time-to-peak for PZ tumor localization (1.67). This finding is supported by Sung et al [119], who found that time-to-peak was a better independent predictor of cancer presence in the TZ than in the PZ. Our finding that nearly distinct subsets of radiomic features are useful for prostate cancer localization in the TZ and PZ, respectively, appears to corroborate previous studies that have suggested that PZ and TZ tumors have distinct appearances on MRI [22, 39, 40, 120].
4.4.2 Stability of VIP feature subsets

In order to ensure that features identified by VIP are truly predictive of cancer presence, VIP was evaluated in terms of its ability to identify a stable set of features that provides high accuracy in localizing prostate cancer on MRI. The stability of the feature subsets selected by VIP was assessed using the Jaccard index and compared with mRMR, a popular feature selection method. The Jaccard index was higher for VIP than mRMR for all four datasets considered in this work, and this difference was statistically significant for datasets $S_2$, $S_3$, and $S_4$. Nevertheless, while the Jaccard index associated with $J_{TZ}$ (dataset $S_3$) was especially high (0.68), the Jaccard index associated with $J_{PZ}$ (datasets $S_1$ and $S_2$) was only mediocre (0.38–0.43). This may be attributed to the fact that the TZ tumors considered in this study shared similar characteristics, while the PZ tumors considered in our study were more diverse in terms of shape, size, location, and tumor microarchitecture and morphology.

4.4.3 Classification accuracy provided by selected features

The predictive power of the features selected by VIP was evaluated by AUC values associated with naive Bayes and logistic regression classifiers that were constructed in conjunction with $J_{TZ}$ and $J_{PZ}$ for voxel–wise cancer localization. Regardless whether a naive Bayes or a logistic regression classifier was used, AUC values obtained using $J_{TZ}$ and $J_{PZ}$ were high (0.73–0.85) and comparable with other studies, which reported AUC values ranging from 0.73 to 0.86 (10,11). In comparison, studies that used MRI features to discriminate between cancerous lesions and normal regions–of–interest reported AUC values ranging from 0.71 to 0.94 [25, 50, 121].
Nevertheless, we found that AUC values were higher for TZ cancer localization than for PZ cancer localization. This trend, which was previously reported by Viswanath et al [22], may be attributed to the fact that the TZ tumors analyzed in our study were similar in size and located in the same region of the TZ, whereas the PZ tumors varied considerably.

### 4.4.4 Comparison of signal intensity and radiomic features

Previous studies have reported improved prostate cancer detection and localization accuracy when radiomic features are used in addition to MRI signal intensities [25, 55]. Using the VIP scheme we quantitatively compared the relative contributions of MRI signal intensities and radiomic features. Radiomic features from T2w MRI and ADC maps were found to be more predictive of cancer localization than T2w MRI signal intensities and ADC values, which were associated with low VIP scores for both TZ and PZ cancer localization. Among DCE MRI features, several radiomic kinetic features were associated with high VIP scores, while the original time–resolved intensity values and tissue concentration curves were not. The relatively low importance of many of the DCE MRI features may be attributed to the fact that our DCE MRI data were of very low temporal resolution (95 sec).

### 4.4.5 Ranking MRI protocols for prostate cancer localization

A number of studies [82, 83, 119, 122, 123] have explored the added benefit of multi–parametric MRI, in comparison to T2w MRI alone for prostate cancer detection but have reported contradictory results. For example, while some have reported that combining ADC maps and T2w MRI improves cancer detection when
compared to T2w MRI alone [82, 123], others found that the addition of ADC maps does not significantly improve cancer detection [122]. In our study the VIP scheme was employed to quantify how much each MRI protocol contributed to accurate cancer localization in both the TZ and the PZ. For TZ cancer localization, T2w MRI features contributed almost twice as much as DWI features; when only the features in $J_{TZ}$ are considered, features extracted from T2w MRI contributed more than five times as much as DWI features. Thus, although it is well established that ADC values are more predictive of cancer presence than T2w signal intensities [50, 121], texture features extracted from T2w MRI appear to be more predictive of TZ cancer presence than both ADC values and ADC texture features. In contrast, for PZ tumor localization radiomic features from both T2w MRI and ADC maps contributed substantially to classification performance. Our finding that ADC maps are more beneficial for localizing PZ tumors than TZ tumors may be attributed to a higher DWI signal-to-noise ratio near the endorectal coil. We note that although DCE MRI features comprised only 3% of the features explored in this study, they played a large role in localization of both TZ and PZ cancer.

4.4.6 Limitations

There were a few limitations to our study. First, our patient cohort was small, consisting of only 10 PZ tumors and 5 TZ tumors with T2w, DWI, and DCE MRI and an additional 8 PZ tumors with only T2w and DCE MRI. Although this small cohort size is similar to other studies [22, 25, 55, 121], the small sample size does have ramifications with regard to classifier generalizability. Applying PCA is helpful for building a classifier when the sample size is small and the data dimensionality
is high; however, even PCA can be subject to the Hughes effect, causing instability of the PCA embedding when the sample size is too small [124]. In our study, the PCA embedding was somewhat unstable because of the small cohort size; this led to some instability in VIP scores and hence feature sets selected based on VIP scores. A second limitation was the fact that we did not quantitatively evaluate the coregistration of MRI and histology. Some error may have been introduced in the determination of the slice correspondences between histology sections and MRI slices and may have affected the determination of “ground truth” cancer extent. Nevertheless, we believe that our coregistration was more rigorous than previous studies that manually delineated cancer regions on MRI by visual registration with pathology information [25, 55, 121]. Third, VIP is limited by the fact that it is an unsupervised, linear DR method. Because PCA is unsupervised, the PCA embedding may be driven by differences between normal TZ and PZ, as well as differences between cancer and normal prostate regions. Consequently, it is possible that some features that are highly predictive of cancer presence may not be identified by VIP. Additionally, because PCA considers only linear correlations between features, it is possible that a different set of features may be selected if the VIP scheme were applied to an embedding obtained from a nonlinear DR scheme. A final limitation is our choice of radiomic features. Although we endeavored to extract a comprehensive collection of radiomic features, we could not include all possible image features, such as local binary pattern and histogram of gradient features. Nevertheless, the VIP scheme is a general framework to identify the relative contributions of any available MRI protocols and MRI features to accurate cancer localization.
Chapter 5

Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

5.1 Overview

In this chapter we explore the ability of VIP to identify radiomic features describing prostate cancer morphology on pre-treatment MRI that are useful for predicting whether a patient will develop biochemical recurrence within ten years of radiation therapy. Radiomic texture features, which quantitatively describe tumor micro-architecture and morphology on MRI, have been shown to provide clues about a tumor’s aggressiveness, which is thought to be associated with biochemical recurrence risk. However, while radiomic features have been employed for predicting cancer presence and grade, they have not been evaluated in the context of predicting risk of biochemical recurrence. This work seeks to evaluate the role of radiomic texture features in predicting risk of biochemical recurrence on a cohort
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

of sixteen patients who underwent pre–treatment 1.5 Tesla (T) T2w MRI.

The primary goal of this chapter is to construct a predictive model that uses a small number of radiomic T2w MRI features to predict biochemical recurrence risk following radiation therapy. Toward this end, we extract a combination of first–order statistical features, which highlight regions with large changes in signal intensity; non–steerable gradient features, which highlight tumor boundaries; co–occurrence features, which highlight homogeneous regions of low signal intensity in aggressive cancer; and Gabor wavelet features, which consider multi–scale and multi–orientation architectures within the tumor. Then, partial least squares (PLS) is implemented to reduce the dimensionality of the high–dimensional MRI features. We leverage VIP [16] to rank each radiomic feature according to the extent of its contribution to accurate prediction of biochemical recurrence in the PLS embedding subspace. Finally, the few features associated with the highest VIP scores are used to construct a predictive model that is compared to the Kattan nomogram in terms of its ability to accurately distinguish between patients who will develop biochemical recurrence and those who will remain recurrence–free. A summary of our approach is illustrated in Figure 5.1.

5.2 Materials and methods

5.2.1 Data description

Sixteen patients with biopsy–confirmed prostate cancer who underwent external beam radiation therapy or brachytherapy and subsequently participated in at least
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

Figure 5.1: Flowchart illustrating methodology used in this paper for extraction, scoring, ranking, and evaluation of radiomic features for predicting biochemical recurrence risk.
five years of follow–up were retrospectively included in this study. Biochemical recurrence after radiation therapy was defined, according to the ASTRO definition, as a rise of 2 ng/mL or more above the nadir PSA [45]. Using this criterion, seven patients experienced biochemical recurrence within 10 years of the end of treatment, while the remaining nine patients remained recurrence–free for at least five years. The average time until biochemical recurrence was 6.8 years, and the mean length of follow–up time in the non–failure patients was 7.0 years. Among these sixteen patients two patients who developed biochemical recurrence and three who remained recurrence–free underwent hormonal therapy in addition to radiation therapy.

Prior to treatment each patient had been clinically referred for a prostate cancer MR staging exam for improved therapeutic selection; this exam included acquisition of T2w MRI and MR spectroscopy. MRI was performed by using a 1.5 T whole–body MRI unit (Signa; GE Medical Systems, Milwaukee, Wisconsin). The patients were imaged while in the supine position by using a body coil for signal excitation and a pelvic phased–array coil (GE Medical Systems) combined with a balloon–covered expandable endorectal coil (Medrad, Pittsburgh, PA) for signal reception. Data sets were acquired as $16 \times 8 \times 8$ phase–encoded spectral arrays (1024 voxels) by using a nominal spectral resolution of $0.24–0.34 \text{ cm}^3$, 1000/130, and a 17–min acquisition time. MR spectroscopy was not used in our study for biochemical recurrence risk prediction; however, MR spectra were used to determine cancer presence and extent on MRI. Consequently, three–dimensional MR spectroscopic imaging data were processed and aligned with the corresponding T2w imaging data using a combination of in–house software and Interactive Display Language
(Research Systems, Boulder, Colorado) software tools [125]. The raw spectral data were apodized with a 1–Hz Gaussian function and Fourier transformed in the time domain and in three spatial domains.

### 5.2.2 Annotation of ground truth cancer extent

MR spectral voxels were annotated by an expert radiologist with more than 25 years of experience on a 5-point scale adapted from the standardized 5-point scale developed by Jung et al. [126], where each spectrum is defined as either (1) definitely benign, (2) probably benign, (3) equivocal, (4) probably malignant, or (5) definitely malignant. While the scale described by Jung et al. was based on metabolic ratios of MR spectra alone, the spectral annotations in this work were performed by incorporating the presence and strength of hypointensities on T2w MRI as well. MRS voxels annotated as “likely benign” or “probably benign” were considered to be “benign” for our analyses, while voxels annotated as “probably malignant” or “likely malignant” were considered cancerous. MR spectroscopy voxels annotated as “equivocal” were excluded from further analysis. Semantic MRI attributes, including tumor size, presence of extracapsular spread, and presence of seminal vesicle invasion were also recorded. Following annotation of ground truth cancer extent on MR spectroscopy, T2w MRI was brought into spatial alignment with MR spectroscopy using information in the image descriptor files. Thus, texture features were extracted only from T2w MRI voxels labeled as cancerous, and non–cancerous regions were ignored in subsequent analyses.
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

<table>
<thead>
<tr>
<th>Feature</th>
<th>No.</th>
<th>Parameters</th>
<th>Description</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-order statistical</td>
<td>8</td>
<td>window size</td>
<td>Mean, standard deviation, and range of intensities</td>
<td>Localize regions with significant changes in signal intensity</td>
</tr>
<tr>
<td>Non-steerable gradient</td>
<td>13</td>
<td>window size</td>
<td>Image convolutions with Sobel and Kirsch operators</td>
<td>Accurately detect region boundaries</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>13</td>
<td>distance, window size</td>
<td>Statistical features computed from joint pdf of intensity value co-occurrences</td>
<td>Differentiate homogeneous regions of low SI in aggressive cancer from high SI in normal prostate</td>
</tr>
<tr>
<td>Gabor wavelet</td>
<td>54</td>
<td>orientation, wavelength</td>
<td>Multi-orientation features computed from Gaussian function convolved with sinusoid</td>
<td>Quantify local and global features qualitatively assessed by radiologists analyzing cancer</td>
</tr>
</tbody>
</table>

Table 5.1: Description of features computed on a per-voxel basis from T2w MRI and the motivation for using them for predicting biochemical recurrence of prostate cancer. SI = signal intensity.

5.2.3 Feature extraction from T2w MRI

Recall that a voxel in the T2w MR image is defined as \( z \in Z \), where \( Z \) is a 3-dimensional grid of MRI voxels. Each \( z \in Z \) is associated with a label \( y(z) \in \{0, 1\} \), where \( y(z) = 1 \) if voxel \( z \) is cancerous and \( y(z) = 0 \) otherwise. Let \( Z^1 = \{ z : y(z) = 1 \} \). Each voxel in \( Z^1 \) is also associated with the feature vector \( F(z) \), which is comprised of radiomic features extracted from T2w MRI. A brief description of the image texture features extracted in this work, as well as our motivation for using them to predict biochemical recurrence, is provided in Table 5.1.

Following voxel-level extraction of texture features from cancerous prostate voxels (see Figure 5.2), patient-level features were subsequently computed by calculating the mean, median, standard deviation, skewness, and kurtosis of the distributions of each of these features. Thus, for \( j \in \{1, \ldots, 88\} \) the mean value of the \( j \)th feature was calculated as \( \frac{1}{|Z|} \sum_{z \in Z^1} F_j(z) \), where \(|*|\) denotes cardinality. In additional to the mean, the median, standard deviation, skewness and kurtosis were similarly computed for each of the 88 features, together providing a single, patient-level feature vector \( x \) containing 440 features.
5.2.4 Feature selection via VIP

In order to reduce the dimensionality of the data, we use PLS [15], a popular linear dimensionality reduction method. VIP (see eq. 3.13) was employed in order to identify a handful of features that have the greatest influence in the PLS model. In order to ensure generalizability of the selected feature set, leave-one-out cross-validation was performed. At each step of this cross-validation procedure, the data from all but one subject was used to construct the PLS embedding and subsequently to compute feature-wise VIP scores. After the cross-validation procedure was completed, the top three radiomic features were identified by voting across all cross-validation iterations, and the most commonly selected features were thus identified.

5.3 Experimental design

5.3.1 Experiment 1: Evaluation of VIP for feature selection

In order to evaluate whether the VIP scheme is indeed selecting features with high predictive power, the top 1–10 features with the highest VIP scores were employed to construct logistic regression models that predict a patient’s risk of developing biochemical recurrence. Based on the results obtained from these predictive models, receiver operating characteristic (ROC) curves representing the tradeoff between classifier sensitivity and specificity were generated. The area under the ROC curve (AUC) was used to comparatively evaluate classifier accuracy in conjunction
with different feature subsets. The AUC associated with each of the logistic regression classifiers was evaluated using a leave-one-out cross-validation procedure to ensure generalization of the classifier and to obtain a robust estimate of AUC.

Additionally, in order to compare VIP with state-of-the-art feature selection algorithms in terms of its ability to identify features with high predictive power, we compared VIP with minimum-redundancy-maximum-relevance (mRMR) [14], a popular information theoretic feature selection scheme. AUC values associated with logistic regression classifiers constructed both in conjunction with the top 1–10 features selected by mRMR and with the top 1–10 features selected by VIP were obtained using leave-one-out cross-validation.

5.3.2 Experiment 2: Comparative evaluation of classifier for biochemical recurrence risk prediction

The top three radiomic features obtained by the VIP scheme were further compared to (a) semantic MRI attributes (see Table 5.3) and (b) the Kattan nomogram in terms of their ability to accurately predict biochemical recurrence following radiation therapy. The Kattan nomogram incorporates clinical features (see Table 5.3) to predict the probability of remaining recurrence-free for five years. To evaluate the predictive power of semantic attributes on MRI, a logistic regression classifier was trained in a leave-one-out cross-validation procedure to predict biochemical recurrence based on the following semantic attributes: number of cancerous MRS voxels, a surrogate measurement of tumor size; presence or absence of extracapsular spread; and presence or absence of seminal vesicle invasion. The AUC associated with this predictive model, as well as the area under the ROC curve drawn
based on the survival probabilities obtained using the Kattan nomogram, was thus obtained.

In addition to comparing AUC values associated with these methods, Kaplan–Meier survival analysis was also performed. Kaplan–Meier survival probabilities, calculated as the ratio of the number of subjects who remain event–free to the total number of study subjects, are useful for measuring the fraction of subjects who remain recurrence–free at any given time after treatment. Because Kaplan–Meier survival analysis inherently accounts for censored data, it is able to account for subjects in our study who were lost to follow–up prior to ten years following treatment. Kaplan–Meier survival curves that stratify patients based upon their risk of developing biochemical recurrence were estimated and compared for the top three VIP–selected radiomic features and the Kattan nomogram.

5.4 Experimental results and discussion

5.4.1 Experiment 1: Evaluation of VIP for feature selection

The three radiomic features that contribute most to accurate prediction of biochemical recurrence following radiation therapy in the PLS embedding space are the skewness and kurtosis of the distributions of three Gabor wavelet features (see Table 5.2 and Figure 5.2). When these three features were used in conjunction with a logistic regression classifier to predict biochemical recurrence risk, an AUC of 0.74 was obtained; when the top four features were used, an AUC of 0.83 was achieved
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

<table>
<thead>
<tr>
<th>VIP: Top 4 Features</th>
<th>VIP Scores</th>
<th>mRMR: Top 4 Features</th>
<th>VIP Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor feature 9 skewness</td>
<td>1.30 ± 0.31</td>
<td>Gabor feature 14 SD</td>
<td>1.04 ± 0.09</td>
</tr>
<tr>
<td>Gabor feature 8 skewness</td>
<td>1.28 ± 0.29</td>
<td>Gabor feature 51 kurtosis</td>
<td>0.77 ± 0.28</td>
</tr>
<tr>
<td>Gabor feature 44 kurtosis</td>
<td>1.18 ± 0.22</td>
<td>Gabor feature 52 skewness</td>
<td>0.65 ± 0.19</td>
</tr>
</tbody>
</table>

Table 5.2: Top three features selected by the VIP feature selection scheme and by mRMR and their associated VIP scores. SD = standard deviation.

<table>
<thead>
<tr>
<th>Semantic MRI Attributes</th>
<th>Kattan Nomogram Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor size</td>
<td>Pre-treatment PSA</td>
</tr>
<tr>
<td>Extracapsular spread</td>
<td>Tumor stage</td>
</tr>
<tr>
<td>Seminal vesicle invasion</td>
<td>Gleason score</td>
</tr>
<tr>
<td></td>
<td>Radiation dose</td>
</tr>
<tr>
<td></td>
<td>Hormonal treatment</td>
</tr>
</tbody>
</table>

Table 5.3: Lists of semantic MRI attributes and clinical features assessed via the Kattan nomogram.

(see Figure 5.3(a)). In Figure 5.3 gray-scale representations of the Gabor filter associated with the highest VIP score illustrate over-expression in a patient who develops biochemical recurrence within four years compared to a patient who remains recurrence-free for ten years. While the distributions of this Gabor feature in these two patients have very different shapes, and hence different values for skewness and kurtosis, the shapes of the distributions of an arbitrarily-selected first-order statistical feature are very similar for both patients (see Figure 5.2).

A scatterplot of the three radiomic features with the highest VIP scores shows excellent separation between subjects who will develop biochemical recurrence and those who remain recurrence-free, as a plane drawn through the scatterplot would lead to misclassifications of only one subject in each class (see Figure 5.4(a)). This suggests that PLS, a linear dimensionality reduction method, is sufficient for
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

Figure 5.2: The top-performing Gabor feature selected by VIP (top) and a first-order statistical feature (bottom) is shown for (a) and (d) a recurrence-free patient and (b) and (e) a patient who develops biochemical recurrence with the ground truth cancer extent bounded in white. The histograms in (c) and (f) show the distributions of these feature in cancerous regions of the prostate for the recurrence-free patient (blue) and the patient who experiences biochemical recurrence (red). Note that significant differences in both the skewness and kurtosis of the distributions of the Gabor wavelet feature can be appreciated in (c) but not in (f).
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

Figure 5.3: (a) AUC is compared for various numbers of radiomic features selected by VIP and mRMR. (b) AUC is compared for the top-performing features obtained via the VIP and mRMR feature selection algorithms, as well as for semantic MRI attributes and the Kattan nomogram.

This data and that nonlinear methods for dimensionality reduction are not necessary in this context. Interestingly, patient 2410, the only subject who developed biochemical recurrence but whose datapoint is present among the cluster of recurrence-free subjects, did not experience biochemical recurrence until 7.6 years after radiation therapy. This is the longest span of time until biochemical recurrence seen in our patient cohort.

It is notable that the radiomic features identified by the VIP scheme were all Gabor wavelet features. In fact, the features selected via the mRMR feature selection routine were also exclusively Gabor wavelet features. These findings suggest the importance of Gabor wavelet features in predicting progression to biochemical
Chapter 5. Applications of VIP for Predicting Biochemical Recurrence Following Radiation Therapy for Prostate Cancer

5.4.2 Experiment 2: Comparative evaluation of classifier for biochemical recurrence risk prediction

In comparison to the top–performing radiomic MRI features, semantic attributes on MRI and the Kattan nomogram provided significantly lower AUC values of recurrence following radiation therapy for prostate cancer. Gabor filters [127] provide multi–scale, multi–orientation features that reflect localized frequency characteristics within the tumors and that may hint to microarchitectural patterns of more aggressive tumors that are more likely to recur. The ability of Gabor wavelet features to capture microarchitectural patterns that characterize cancer has been previously documented, as Gabor wavelet features were identified as important for detection and localization of prostate tumors on T2w MRI [21, 22] and ADC maps [21].

Figure 5.4: Scatterplot of the top–performing Gabor features selected by (a) VIP and (b) mRMR.
0.63 and 0.58, respectively (see Figure 5.3(b)). Kaplan–Meier survival curves that stratify the patients based upon their risk of developing biochemical recurrence are shown in Figure 5 for the three radiomic features selected by VIP and for the Kattan nomogram. It is clear from Figure 5.5 that VIP–selected features provide better separation between patients who will develop biochemical recurrence and those who will remain recurrence–free than the Kattan nomogram, although the differences between the two survival curves are not statistically significant. These results suggest that a few radiomic texture features provide more accurate prediction of biochemical recurrence risk than both the Kattan nomogram and semantic attributes on MRI.
Chapter 6

Applications of FINE in Digital Pathology

6.1 Overview

Quantitative histomorphometry (QH) is the process of computationally modeling the appearance of disease on digital pathology images via image-based features. QH approaches typically involve the extraction of a large number of features describing the texture, color, and spatial arrangement of nuclei and glands on digitized images of tissue slides [9, 63–67]. These features have been shown to be useful in determining cancer aggressiveness [9, 66, 67] and in predicting the likelihood of a patient’s cancer recurring following treatment [65, 68]. Nevertheless, since it is typical to extract hundreds or even thousands of features from digital pathology images, the dimensionality of the feature space poses a formidable challenge to the construction of robust classifiers for predicting disease presence and aggressiveness.

In this chapter we demonstrate that FINE (see Chapter 3) can be employed
to identify QH features that are associated with disease aggressiveness and outcome in the context of four different scenarios involving breast and prostate cancer digitized tissue microarray and whole slide images. In the context of breast cancer, we focus on two problems that relate to predicting treatment outcome and risk assessment of estrogen receptor positive (ER+) breast cancers. The working hypothesis for the two breast cancer problems we address in this paper is that computer–extracted features describing cancer patterns in ER+ breast cancer tissue slides can predict (a) which cancer patients will have recurrence following treatment with tamoxifen and (b) risk category as determined by a 21 gene expression assay called Oncotype DX. Oncotype DX is a commercial assay that allows for prediction of which patients would not benefit from adjuvant chemotherapy (identified as low risk patients) and which patients would benefit from adjuvant chemotherapy (identified as high risk patients). In the context of prostate cancer, we employ FINE to identify computer–extracted image features of cancer patterns on tissue images that are associated with risk of biochemical recurrence. Toward this end, we look at QH features extracted from tissue microarray images obtained from needle core biopsies and whole mount radical prostatectomy specimens. For both breast and prostate cancers, knowledge about which QH features are most predictive of risk of recurrence could potentially lead to improved disease characterization upon biopsy and better planning of therapy in the adjuvant and neoadjuvant settings. FINE provides the ability to identify key QH features that contribute most substantially to class discriminability.
### 6.2 Datasets

#### 6.2.1 Predicting breast cancer recurrence based on tissue microarrays

Tissue microarrays were obtained from a cohort of 48 patients with ER+ breast cancer who underwent chemotherapy, some of whom experienced recurrence and some of whom remained recurrence–free (see Table 6.1). A total of 53 QH features were extracted from tissue microarrays stained with hematoxylin and eosin. These features included quantitative descriptors of Voronoi, Delaunay, and minimum spanning tree graphs connecting nuclei within each of the individual tissue cylinders (see Figure 6.1). FINE is leveraged to identify which of the QH features are most useful for predicting recurrence risk.

#### 6.2.2 Predicting OncotypeDX risk categories on whole slides

QH features were extracted from whole slide tissue biopsy images obtained from a cohort of 140 subjects (see Table 6.1) for whom OncotypeDX recurrence risk scores
Chapter 6. Applications of FINE in Digital Pathology

**Figure 6.1:** Digital pathology and feature representations. For dataset $S_1$, a representative patch of histology is shown in (a); segmented nuclei are illustrated in (b); and Voronoi and Delaunay graphs constructed based on the nuclei are shown in (c) and (d), respectively. For dataset $S_2$, a representative patch of histology is shown in (e); segmented nuclei are illustrated in (f); and texture representations of the histology are shown in (g) and (h). For dataset $S_3$, a representative patch of histology is shown in (i); segmented nuclei are illustrated in (j); and Voronoi and Delaunay graphs constructed based on the nuclei are shown in (k) and (l), respectively. For dataset $S_4$, a representative patch of histology is shown in (m); segmented glands are illustrated in (n); and Voronoi and Delaunay graphs constructed based on the glands are shown in (o) and (p), respectively.
79

were known (OncotypeDX risk scores of 0–18 indicate low risk, 19–30 indicate intermediate risk, and 30–99 indicate high risk of recurrence). Biopsy samples were stained with hematoxylin and eosin and digitized. 2343 QH features describing the texture and shape of nuclei were computed on digital pathology (see Figure 6.1). FINE is leveraged to identify which of the QH features are most useful for classifying low versus high OncotypeDX scores, a surrogate predictor of recurrence risk.

6.2.3 Predicting biochemical recurrence of prostate cancer based on tissue whole mounts

Tissue whole mounts were acquired from 40 subjects with biopsy–confirmed prostate cancer who underwent radical prostatectomy and were followed for at least five years afterwards. Tissue whole mounts were stained with hematoxylin and eosin, and a total of 56 cancer regions were delineated on digitized whole mounts. Finally, 242 graph–based features describing nuclear arrangements on pathology that were previously shown to be useful for prostate cancer characterization [65] were extracted from each cancerous region. These features included quantitative descriptors of Voronoi, Delaunay, and minimum spanning tree graphs connecting nuclei on digital pathology (see Figure 6.1); architectural features describing cell clusteredness; Fourier descriptors quantifying nuclear morphology; and texture features computed based on co–occurrence matrices. FINE is leveraged to identify which of the QH features are most useful for predicting risk of biochemical recurrence within five years of radical prostatectomy since biochemical recurrence is a major risk factor for poor outcome in prostate cancer patients [128].
6.2.4 Predicting biochemical recurrence of prostate cancer based on tissue microarrays

Tissue microarrays were acquired from 44 subjects with biopsy–confirmed prostate cancer participating in an active surveillance protocol. Tissue was stained with hematoxylin and eosin, and a total of 242 graph–based features describing glandular arrangements on pathology were extracted from the microarrays (see Table 6.1, Figure 6.1). These features included quantitative descriptors of Voronoi, Delaunay, and minimum spanning tree graphs connecting glands on digital pathology; architectural features describing gland arrangements; Fourier descriptors quantifying gland morphology; and texture features computed based on co–occurrence matrices. FINE is leveraged to identify which of the QH features are most useful for predicting risk of biochemical recurrence.

6.3 Experimental design

6.3.1 Dimensionality reduction

FINE was implemented to identify key contributors to classification on embeddings obtained via four DR approaches: PCA, Isomap, LE, and LLE. For each digital pathology dataset, a class–balanced set of approximately 75% of the instances were sampled during each of fifty rounds of bootstrapping to train the classifier, and the remaining instances were used to evaluate the classifier during each bootstrapping round. Parameters associated with these DR methods—the intrinsic dimensionality parameter $h$, which is used in all three embeddings; $K$, which is used
to create a neighborhood graph for Isomap, LE, and LLE; and $\sigma$, which is needed to compute the graph Laplacian for LE—were chosen during each bootstrapping round to reduce residual variance (see eq. (3.19)).

### 6.3.2 Classifier training

Once embeddings were constructed, FINE scores were computed, and the features associated with the highest FINE scores were selected. Because a sample-to-feature ratio of 10:1 is generally recommended to build a robust classifier [3], the number of selected features was limited to one tenth of the number of samples in the training dataset. Then, within the same bootstrapping rounds, classifiers were trained in conjunction with the features selected by FINE. It is important to note that the objective of FINE is not merely to select features that maximize classification accuracy, but rather to identify a stable set of features that also provide good class discriminability. Consequently, our focus was less on constructing the most accurate classifier and more on identifying features that could yield stable and reproducible classifiers. The logistic regression classifier employs linear regression when the outcome variable is binomially distributed; thus, it allows the features themselves to drive classification. Due to its simplicity, the logistic regression classifier was chosen to evaluate classifier accuracy associated with features selected by FINE. Additionally, we evaluated the classification accuracy associated with the selected features in conjunction with two other classifiers: a support vector machine (SVM) with a radial basis function kernel (scale factor = 1) and an RF classifier with fifty trees.
6.3.3 Performance evaluation measures

The feature selection performance of FINE was evaluated in comparison to three filter methods commonly used for FS: the $t$–test, Fisher score, and Gini index [108]. Evaluation was performed to assess (a) stability and (b) classification accuracy associated with selected feature subsets. Stability was evaluated by the Jaccard index (see eq. (3.20)). Additionally, classification accuracy associated with feature subsets selected by the FINE method was evaluated in conjunction with LR, SVM, and RF classifiers. Based on the prediction results obtained from these classifiers, receiver operating characteristic (ROC) curves were generated, and the area under the ROC curve (AUC) was used to evaluate classifier accuracy in conjunction with different feature subsets. For each of the digital pathology datasets, a class–balanced set of up to 75% of the instances were sampled during each round of bootstrapping to train the classifier, and the remaining instances were used as independent data to evaluate the classifier.

6.4 Experimental results and discussion

6.4.1 Predicting breast cancer recurrence based on tissue microarrays

Figure 6.2 displays the values of $J$ associated with feature subsets selected by the $t$–test, Fisher score, Gini index, and FINE. For dataset $S_1$, the $t$–test, Fisher score, and Gini index provided moderately stable feature sets. Whereas $\text{FINE}_{\text{LLE}}$, $\text{FINE}_{\text{Isomap}}$,
Figure 6.2: AUC and Jaccard index associated with four dimensionality reduction algorithms, the top five features selected based on FINE scores, and three common feature selection algorithms for pathology datasets (a) $S_1$, (b) $S_2$, (c) $S_3$, and (d) $S_4$. Data points are shown for three classification methods: logistic regression (black outline), support vector machine (green outline), and random forest (red outline) classifiers.
and FINE\textsubscript{LE} provided low to moderately stable feature subsets, FINE\textsubscript{PCA} provided higher stability ($J = 0.5$).

AUC values obtained by using the selected features in conjunction with LR, SVM, and RF classifiers are shown in Figure 6.2. The top ranking features selected by FINE\textsubscript{PCA} and FINE\textsubscript{LLE} perform on par with features selected by the Gini index, $t$–test and Fisher score, yielding AUC values between 0.8 and 0.83. For all four DR methods, the top–ranking features selected by FINE provided higher AUC values than the embedding vectors themselves. The top–ranking features selected by FINE\textsubscript{PCA}, FINE\textsubscript{Isomap}, FINE\textsubscript{LE}, and FINE\textsubscript{LLE} were similar, including the number of subgraphs with low, intermediate, and high Hosoya indices and local measures of cell clusteredness (see Table 6.2). Hosoya features are closely associated with cell clusteredness since the Hosoya index measures connectedness of subgraphs representing clusters of cells. Since highly proliferative tumors tend to manifest more closely clustered cells, cell clusteredness is closely related with proliferation, which is highly predictive of recurrence risk for breast cancer [129].

### 6.4.2 Predicting OncotypeDX risk categories on whole slides

For dataset $S_2$, $J$ associated with feature subsets selected by the $t$–test, Fisher score, Gini index, and FINE remains consistently below 0.27 (see Figure 6.2). Nevertheless, FINE\textsubscript{PCA}, FINE\textsubscript{Isomap}, FINE\textsubscript{LE}, and FINE\textsubscript{LLE} all provide higher stability than the comparative strategies, with $J$ ranging from 0.64–0.91. The stability of feature rankings obtained via FINE can be attributed to the robustness of the data embedding in the low dimensional space. FINE\textsubscript{PCA} provides the most stable feature subset, followed by FINE\textsubscript{ISO}, FINE\textsubscript{LE}, and FINE\textsubscript{LLE}.
<table>
<thead>
<tr>
<th>S1</th>
<th>PCA</th>
<th>5</th>
<th>Top 5 Features</th>
<th>S2</th>
<th>Isomap</th>
<th>5</th>
<th>Top 5 Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Filled area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intermediate Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Convex area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean NN in 50–pixel radius</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 10 SD RGB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Disorder of distance to 7 NN</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 55 SD RGB</td>
</tr>
<tr>
<td></td>
<td>Isomap</td>
<td>5</td>
<td>Disorder of distance to 7 NN</td>
<td></td>
<td></td>
<td></td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean NN in 50–pixel radius</td>
<td></td>
<td></td>
<td></td>
<td>Filled area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD of NN in 50–pixel radius</td>
<td></td>
<td></td>
<td></td>
<td>Convex area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 10 SD RGB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 55 SD RGB</td>
</tr>
<tr>
<td></td>
<td>LE</td>
<td>5</td>
<td>Disorder of distance to 7 NN</td>
<td></td>
<td>LE</td>
<td>7</td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean NN in 50–pixel radius</td>
<td></td>
<td></td>
<td></td>
<td>Filled area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD of NN in 50–pixel radius</td>
<td></td>
<td></td>
<td></td>
<td>Convex area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 10 SD RGB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intermediate Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 55 SD RGB</td>
</tr>
<tr>
<td></td>
<td>LLE</td>
<td>23</td>
<td>Low Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Filled area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intermediate Hosoya index</td>
<td></td>
<td></td>
<td></td>
<td>Convex area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Delaunay $\triangle$ area min/max</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 10 SD RGB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Delaunay $\triangle$ side length min/max</td>
<td></td>
<td></td>
<td></td>
<td>Gray level 55 SD RGB</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>5</td>
<td>Fourier descriptor 5 SD</td>
<td></td>
<td>PCA</td>
<td>6</td>
<td>Invariant moment 3 min/max</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Delaunay side length min/max</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 4 min/max</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean tensor correlation</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 7 mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 1 min/max</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 5 min/max</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD of distance to 3 NN</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 5 min/max</td>
</tr>
<tr>
<td></td>
<td>Isomap</td>
<td>5</td>
<td>Mean information measure 1</td>
<td></td>
<td>Isomap</td>
<td>5</td>
<td>Invariant moment 5 SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD of contrast variance</td>
<td></td>
<td></td>
<td></td>
<td>Fourier descriptor 8 mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD of intensity variance</td>
<td></td>
<td></td>
<td></td>
<td>Mean connected component size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean NN in 40–pixel radius</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 5 mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Perimeter ratio SD</td>
<td></td>
<td></td>
<td></td>
<td>Mean connected component size</td>
</tr>
<tr>
<td></td>
<td>LE</td>
<td>5</td>
<td>Fourier descriptor 1 min/max</td>
<td></td>
<td>LE</td>
<td>5</td>
<td>Invariant moment 3 min/max</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean tensor correlation</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 4 min/max</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 6 SD</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 6 SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tensor contrast energy SD</td>
<td></td>
<td></td>
<td></td>
<td>Fourier descriptor 10 mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD of entropy</td>
<td></td>
<td></td>
<td></td>
<td>Fourier descriptor 8 SD</td>
</tr>
<tr>
<td></td>
<td>LLE</td>
<td>21</td>
<td>Fourier descriptor 9 min/SD</td>
<td></td>
<td>LLE</td>
<td>21</td>
<td>Fourier descriptor 8 mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fourier descriptor 9 min/SD</td>
<td></td>
<td></td>
<td></td>
<td>Fourier descriptor 7 median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fourier descriptor 7 SD</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 7 median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 6 min/max</td>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 2 min/max</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Invariant moment 5 min/max</td>
<td></td>
<td></td>
<td></td>
<td>Delaunay $\triangle$ area min/max</td>
</tr>
</tbody>
</table>

**Table 6.2**: Average $h$ and top features selected by FINE in conjunction with four DR methods for each of the pathology datasets (obtained by voting across 50 bootstrapping rounds). NN=nearest neighbors, SD=standard deviation.
AUC values obtained by using the selected features in conjunction with three types of classifiers are shown in Figure 6.2. By selecting feature subsets comprised of three features or less, FINE provided AUCs ranging between 0.75–0.76, higher than the $t$–test, Fisher score, and Gini index. As was the case with dataset $S_1$, the histomorphometric features ranked highest by FINE were relatively independent of DR scheme. In fact, the top three features were always nuclear area, convex area, and filled area, as nuclear area tends to be higher in patients with high OncotypeDX risk scores (see Figure 6.3). Two Gabor features were found useful in conjunction with Isomap, while two attributes of gray level histograms were useful in conjunction with PCA, LE, and LLE (see Table 6.2). Interestingly, these features are all part of the Bloom Richardson grading lexicon (nuclear pleomorphism and chromatin patterns) [130], and grade has been shown to be one of the strongest predictors of disease outcome in ER+ breast cancers [131]. Hence, the results from FINE are biologically intuitive.
6.4.3 Predicting biochemical recurrence of prostate cancer based on tissue whole mounts

For dataset $S_3$, feature subsets selected based on FINE scores provided high stability, with $J$ values ranging from 0.64–0.74. In contrast, feature subsets selected based on the $t$–test, Fisher score, and Gini index were substantially lower (see Figure 6.2). The top ranking features selected by $\text{FINE}_{\text{PCA}}, \text{FINE}_{\text{Isomap}}, \text{FINE}_{\text{LE}},$ and $\text{FINE}_{\text{LLE}}$ perform on par with features selected by the $t$–test, Fisher score, and Gini index and the embedding vectors themselves. For this dataset the AUC values ranged from 0.52–0.64 depending on the classifier used. These AUC values are similar to AUC values previously reported for this problem. For example, Lee et al. [65] reported AUC values of 0.56–0.67 for predicting biochemical recurrence risk after radical prostatectomy for treating prostate cancer. Only when histomorphometric features were combined with proteomic features obtained via mass spectrometry to predict biochemical recurrence risk following radical prostatectomy were higher AUC values of 0.74–0.93 achieved [4, 132].

The five top–ranking histomorphometric features were similar for $\text{FINE}_{\text{PCA}}, \text{FINE}_{\text{LE}},$ and $\text{FINE}_{\text{LLE}}$ and included primarily morphological features: Fourier descriptors, invariant moments, and tensor features (see Table 6.2). $\text{FINE}_{\text{Isomap}}$, however, returned a very different feature set that included three texture features and two features describing glandular architecture. These features were previously found to be useful for predicting prostate cancer aggressiveness on pathology [133]. Prostate cancer aggressiveness is a part of the Kattan nomogram for predicting the risk of biochemical recurrence of prostate cancer [134], so it is reasonable that these features would be useful for predicting the risk of biochemical recurrence of
6.4.4 Predicting biochemical recurrence of prostate cancer based on tissue microarrays

For dataset $S_4$, FINE$_{PCA}$ and FINE$_{LE}$ are associated with $J$ ranging between 0.22–0.26 (see Figure 6.2), although $J$ associated with feature subsets selected based on the Fisher score were more stable. As was the case with datasets $S_1$ and $S_2$, FINE$_{PCA}$ and FINE$_{LLE}$ provided more stable feature subsets than FINE$_{Isomap}$ and FINE$_{LE}$.

The top ranking features selected by FINE$_{PCA}$, FINE$_{Isomap}$, FINE$_{LE}$ and FINE$_{LLE}$ provided AUC values ranging from 0.54-0.64, on par with the $t$–test, Fisher score, and Gini index (AUC range: 0.57–0.67). The top–ranking histomorphometric features were relatively similar for all DR schemes; they included a combination of Fourier descriptors, invariant moments, and several morphological features computed via a cell graph (see Table 6.2). These features were previously shown to be
useful for predicting biochemical recurrence of prostate cancer [lee13].

In addition to assessing AUC values, Kaplan–Meier survival analysis [135] was also performed for this dataset (see Figure 6.4) since the time until recurrence was available for all subjects who experienced biochemical recurrence. Kaplan–Meier survival probabilities, calculated as the ratio of the number of subjects who remain event–free to the total number of study subjects, are useful for measuring the fraction of subjects who remain recurrence–free at any given time after treatment. Kaplan–Meier survival curves that stratify patients based upon their risk of developing biochemical recurrence are shown in Figure 6. These survival curves were computed using a logistic regression classifier trained on the five top–ranking histomorphometric features. The difference between the two survival curves is statistically significant ($p = 0.02$), according to the log–rank test.
Chapter 7

Classifier–Building in a Multi–Institutional Study: Zone–Specific Prostate Cancer Detection on Multi–Parametric MRI

7.1 Overview

In this chapter we evaluate in a multi–institutional study whether radiomic features for prostate cancer detection from multi–parametric 3 Tesla (T) MRI in the transition zone (TZ) are similar to the features that are useful for prostate cancer detection in the peripheral zone (PZ). For this study multi–parametric MRI was obtained from eighty patients at three institutions. Due to differences in MRI parameters (T1w, T2w, Diffusion) across vendor platforms and scanner [62], VIP and FINE are insufficient for identifying features that are resilient in a cross–institutional setting. First–order statistical, co–occurrence, and wavelet features were extracted
from both T2w MRI and ADC maps, while contrast kinetic features were extracted from DCE MRI. Feature selection was performed to identify the ten most discriminating and consistent features on a per-voxel basis. The top identified radiomic features within the TZ and PZ were then separately combined via two logistic regression classifiers to detect prostate cancer in the TZ and PZ, respectively.

7.2 Materials and Methods

7.2.1 Patients

This retrospective study included 87 patients from three institutions (University of Turku, Finland; St. Vincent’s Prostate Cancer Centre, Sydney; and Northshore LIJ, New York) and was approved by the institutional review board of each of the three institutions. All patients underwent multi-parametric MRI due to suspicion for prostate cancer either prior to prostate biopsy (52 patients) [136] or prior to radical prostatectomy (35 patients). Five patients from the former group were excluded due to poor quality of MRI, and two patients were excluded because complete multi-parametric MRI was not performed. Whole-mount prostatectomy specimens were also available for the 35 patients who underwent radical prostatectomy following MRI, as well as 18 patients who underwent pre-biopsy MRI but subsequently underwent radical prostatectomy. Thus, a total of 80 patients were included (age range, 40-79 years; median, 64 years); radical prostatectomy specimens were available for 51 of these patients. Further details regarding the patients included in this study can be found in Table 7.1.
Table 7.1: Description of subjects used in this study. DW, diffusion-weighted; DCE, dynamic contrast-enhanced; SD, standard deviation
7.2.2 MRI acquisition

MRI was performed with a 3-Tesla MR scanner either with a body coil (67 patients) or an endorectal coil (13 patients). The multi-parametric MRI protocol consisted of T2-weighted imaging, diffusion-weighted imaging, and dynamic contrast-enhanced imaging. MRI acquisition details are listed in Table 1.

7.2.3 Histopathological analysis and cancer annotation on MRI

For subjects from University of Turku who underwent radical prostatectomy following MRI acquisition, whole mount prostatectomy sections were obtained, processed as described in [137], and stained with hematoxylin and eosin. All of the histopathologic material was analyzed by one experienced genitourinary pathologist (8 years of experiences in genitourinary pathology) in consensus with another pathologist (6 years of experiences in genitourinary pathology). The Gleason score was assigned as a combination of primary, secondary, and tertiary Gleason grades according to the 2005 International Society of Urological Pathology Modified Gleason Grading System [57]. A tertiary Gleason grade was assigned when a Gleason grade pattern higher than the primary and secondary Gleason grade patterns was present but accounted for less than 5% of the tumor [138]. Following histopathological analysis, the histologic slides (50 × 75 mm²) were digitized in 2400 dpi resolution using a high resolution scanner.

In order to obtain “ground truth” annotation of prostate cancer extent on MRI, deformable co-registration of MRI and whole mount histological sections was performed. Correspondences between histological sections and T2w MRI slices were
determined by a genitourinary pathologist and radiologist working in unison. Subsequently, corresponding histological sections and MRI slices were co-registered using an interactive B–spline elastic registration scheme [114]. The final result was a labeling of each MRI voxel within the prostate as corresponding to cancer or benign prostate tissue.

### 7.2.4 Cancer annotation on pre–biopsy MRI

For the 29 subjects who did not undergo radical prostatectomy following MRI acquisition, ground truth for prostate cancer extent from excised surgical histopathology was not available. These included 16 subjects from St. Vincent’s Prostate Cancer Centre and 13 subjects from Northshore LIJ. For the subjects from St. Vincent’s Prostate Cancer Centre, a genitourinary radiologist assessed the multi–parametric MRI for cancer presence and subsequently annotated prostate cancer extent, if present, on T2w MRI. For the subjects from Northshore LIJ, a genitourinary radiologist annotated the extent of visible cancer on T2w MRI within sextants associated with positive biopsy results. All lesions were correlated with fusion targeted biopsy results.

### 7.2.5 MRI post–processing

DCE MRI and ADC maps [137] obtained from DWI were spatially aligned with T2w MRI via volumetric affine registration. This step allowed for correction of inter–acquisition movement and inter–protocol resolution differences. After inter–protocol alignment, all MRI data were computationally analyzed at the T2w MRI resolution of \(0.625 \times 0.625 \times 3 \text{ mm}^2\). The prostate capsule and TZ were manually
annotated on T2w MRI by a radiologist with 7 years of experience in prostate MRI. Finally, T2w and DCE MRI were corrected for acquisition-based MRI intensity artifacts [111]. We first corrected for intensity non-standardness [62]—the issue of inter- and intra-patient T2w MRI “intensity drift”—which causes T2w MRI intensities to lack tissue-specific numeric meaning [62] (see Appendix C for more details regarding this algorithm). This effect was corrected by interactive implementation of the generalized scale algorithm [62], which aligns image intensity histograms across different MRI studies, thereby enabling MRI intensities to have a consistent tissue-specific numeric meaning. Additionally, for patients who were imaged using an endorectal probe, the bias field artifact occurring on T2w and DCE MRI was corrected by the N3 algorithm [113] (see Appendix B).

7.2.6 Radiomic features

Our feature set included signal intensities on T2w MRI, ADC values, and six kinetic features describing the uptake and washout of contrast on DCE MRI (see Table 2). In addition, 224 radiomic features (see Table 7.2), which were previously shown to be useful for prostate cancer detection on multi-parametric MRI [21, 30] were extracted. These included first-order statistical, co-occurrence, and wavelet features computed from both T2w MRI and ADC maps. These features are designed to accentuate smooth and spiculated margins and to differentiate between homogeneous regions of low signal intensity associated with prostate cancer and surrounding normal prostate tissue.
Chapter 7. Classifier–Building in a Multi–Institutional Study: Zone–Specific Prostate Cancer Detection on Multi–Parametric MRI

Table 7.2: Overview of radiomic features used in this study

<table>
<thead>
<tr>
<th>Features</th>
<th>Pulse Sequence</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal intensities</td>
<td>Axial T2–weighted</td>
<td>–</td>
</tr>
<tr>
<td>T2–weighted</td>
<td>Diffusion–weighted</td>
<td>–</td>
</tr>
<tr>
<td>ADC</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Texture</td>
<td>T2–weighted, ADC</td>
<td>–</td>
</tr>
<tr>
<td>Co–occurrence features [31]</td>
<td>T2–weighted, ADC</td>
<td>θ = 0 – 2.75, λ = 2.8 – 45.3</td>
</tr>
<tr>
<td>Co–occurrence features [31]</td>
<td>T2–weighted, ADC</td>
<td>–</td>
</tr>
<tr>
<td>Haar wavelet [33]</td>
<td>T2–weighted, ADC</td>
<td>–</td>
</tr>
<tr>
<td>Kinetic</td>
<td>DCE</td>
<td>–</td>
</tr>
<tr>
<td>Time–to–peak</td>
<td>DCE</td>
<td>–</td>
</tr>
<tr>
<td>Initial enhancement</td>
<td>DCE</td>
<td>–</td>
</tr>
<tr>
<td>Maximum enhancement</td>
<td>DCE</td>
<td>–</td>
</tr>
<tr>
<td>Enhancement ratio</td>
<td>DCE</td>
<td>–</td>
</tr>
<tr>
<td>Uptake ratio</td>
<td>DCE</td>
<td>–</td>
</tr>
<tr>
<td>Washout rate</td>
<td>DCE</td>
<td>–</td>
</tr>
</tbody>
</table>

7.2.7 Identifying features for discriminating cancerous from benign voxels

For TZ and PZ tumors, respectively, radiomic features were selected based on (a) resilience and lack of variability between patients and institutions and (b) ability to discriminate between cancerous and benign prostate voxels. Feature resilience across patients and institutions was determined based on Cronbach’s intraclass correlation coefficient (ICC) [139], which measures the level of concordance in feature values. Features associated with an ICC > 0.9 were considered resilient to inter-patient and inter-institutional differences. For each feature identified as being resilient, binary logistic regression was implemented to classify individual voxels as cancerous or benign, and the area under the receiver-operating characteristic (ROC) curve (AUC) was calculated based on the posterior probabilities of a voxel being classified as cancerous. Finally, features associated with both the highest AUC values and ICC > 0.9 were identified as being useful for characterizing prostate cancer in a multi–institutional setting. Two separate sets of features
were identified: one set of features that characterize TZ cancers \((F^{TZ})\) and another set of features to characterize PZ cancers \((F^{PZ})\). Additionally, a third set of features was identified that did not specifically consider account zonal anatomy \((F^{ALL})\). Separate radiomic feature sets were obtained based on patients from University of Turku \((I_1)\), patients from St. Vincent’s Prostate Cancer Centre \((I_2)\), and the combination of patients who underwent radical prostatectomy from \(I_1\) and \(I_2\) \((I_{12} = I_1 \cup I_2)\). Thus, separate feature sets \(F^{TZ}_1, F^{TZ}_2, F^{TZ}_{12}, F^{PZ}_1, F^{PZ}_2, \) and \(F^{PZ}_{12}\) were obtained. Similarly, separate feature sets \(F^{ALL}_1, F^{ALL}_2, \) and \(F^{ALL}_{12}\) were obtained. Note that only feature sets \(F^{TZ}_{12}, F^{PZ}_{12}, \) and \(F^{ALL}_{12}\) were selected based on resilience to inter-patient and inter-institutional differences, while \(F^{TZ}_1, F^{TZ}_2, F^{PZ}_1, F^{PZ}_2, F^{ALL}_1, \) and \(F^{ALL}_2\) were selected based on resilience to inter-patient differences only.

### 7.2.8 Classifier training and evaluation

A leave-one-patient-out cross-validation scheme was employed to train two separate logistic regression classifiers [15]: one used \(F^{TZ}\) to detect cancer in the TZ \((C^{TZ})\) and the other used \(F^{PZ}\) to detect cancer in the PZ \((C^{PZ})\). For comparison, a zone-ignorant classifier \((C^{ALL})\) that leveraged \(F^{ALL}\) to detect cancer across the entire prostate was also trained and evaluated. The performance of \(C^{TZ}, C^{PZ}, \) and \(C^{ALL}\) was evaluated using the area under the receiver operating characteristic curve (AUC). Separate classifiers were trained using each feature set: \(C^{TZ}_1, C^{TZ}_2, C^{TZ}_{12}, C^{PZ}_1, C^{PZ}_2, C^{PZ}_{12}, C^{ALL}_1, C^{ALL}_2, \) and \(C^{ALL}_{12}\) based on \(F^{TZ}_1, F^{TZ}_2, F^{TZ}_{12}, F^{PZ}_1, F^{PZ}_2, F^{PZ}_{12}, F^{ALL}_1, F^{ALL}_2, \) and \(F^{ALL}_{12}\).
7.3 Results

7.3.1 Selected features for TZ and PZ classifiers

The top ten radiomic features selected for detecting TZ and PZ tumors are listed in Table 7.3. The features in each of $F^{PZ}_1$, $F^{PZ}_{12}$, and $F^{PZ}_{2}$ included a combination of Gabor wavelet features, co-occurrence features, and edge descriptors extracted from T2w MRI. In contrast, $F^{PZ}_2$ included only co-occurrence features. Only two of the features in $F^{PZ}_1$, and none in $F^{PZ}_2$ or $F^{PZ}_{12}$ were based on ADC maps; no features extracted from DCE MRI were included in $F^{PZ}$. Gabor wavelet features, co-occurrence features, and edge descriptors extracted from T2w MRI were useful for detecting TZ cancer, too. Unlike for $F^{PZ}$, however, each of $F^{TZ}_1$, $F^{TZ}_2$, and $F^{TZ}_{12}$ included between one and six features computed from ADC maps. The top ranked Gabor wavelet and co-occurrence features in $F^{TZ}_{12}$ and $F^{PZ}_{12}$ are shown in Figure 7.1.

The features in $F^{ALL}$ corresponded more closely with the features in $F^{PZ}$ than $F^{TZ}$. In particular, 3/10 features were common to both $F^{ALL}_1$ and $F^{PZ}_1$, 2 co-occurrence features (sum variance and information measure 1) were common to both $F^{ALL}_2$ and $F^{PZ}_2$, and 9/10 features were identical between $F^{ALL}_{12}$ and $F^{PZ}_{12}$. By contrast, two Gabor features were common to both $F^{ALL}_1$ and $F^{TZ}_1$, two Kirsch edge descriptors were identical between $F^{ALL}_{12}$ and $F^{TZ}_{12}$, and $F^{ALL}$ and $F^{PZ}_2$ did not intersect at all.
### Chapter 7. Classifier–Building in a Multi–Institutional Study: Zone–Specific

Prostate Cancer Detection on Multi–Parametric MRI

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference entropy ($w = 5$)</td>
<td>Sum variance ($w = 3$)</td>
<td>Sum variance ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.12, \lambda = 22.6$)</td>
<td>Gabor ($\theta = 0.79, \lambda = 2.8$)</td>
<td>Sum entropy ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 0.39, \lambda = 22.6$)</td>
<td>Gabor ($\theta = 0, \lambda = 5.7$)</td>
<td>Sum variance ($w = 7$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 2.36, \lambda = 22.6$)</td>
<td>Gabor ($\theta = 1.96, \lambda = 11.3$)</td>
<td>Sum entropy ($w = 7$)</td>
</tr>
<tr>
<td>Info. measure 1 ($w = 7$)</td>
<td>Kirsch edge descriptor 3</td>
<td>Sum average ($w = 5$)</td>
</tr>
<tr>
<td>Energy ($w = 3$)</td>
<td>Kirsch edge descriptor 1</td>
<td>Info. measure 1 ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 2.36, \lambda = 8.2$)</td>
<td>Energy ($w = 7$)</td>
<td>Diff. variance ($w = 5$)</td>
</tr>
<tr>
<td>T2w MRI intensity</td>
<td>Gabor ($\theta = 1.86, \lambda = 45.3$)</td>
<td>Info. measure 1 ($w = 3$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor ($\theta = 3.85, \lambda = 45.3$)</td>
<td>Gabor ($\theta = 2.75, \lambda = 45.3$)</td>
<td>Gabor ($\theta = 1.18, \lambda = 22.6$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.94, \lambda = 5.7$)</td>
<td>Gabor ($\theta = 0, \lambda = 2.8$)</td>
<td>T2w MRI intensity</td>
</tr>
<tr>
<td>Haar diagonal coefficient</td>
<td>Gabor ($\theta = 0.79, \lambda = 11.3$)</td>
<td>Gabor ($\theta = 2.75, \lambda = 45.3$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 0, \lambda = 11.3$)</td>
<td>Gabor ($\theta = 1.57, \lambda = 11.3$)</td>
<td>Info. measure 2 ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 0, \lambda = 5.7$)</td>
<td>Gabor ($\theta = 2.36, \lambda = 8.2$)</td>
<td>Gabor ($\theta = 1.57, \lambda = 45.3$)</td>
</tr>
<tr>
<td>Energy ($w = 5$)</td>
<td>Intensity range ($w = 5$)</td>
<td>Gabor ($\theta = 2.75, \lambda = 45.3$)</td>
</tr>
<tr>
<td>Inverse diff. moment ($w = 5$)</td>
<td>Gabor ($\theta = 1.57, \lambda = 2.8$)</td>
<td>Gabor ($\theta = 0, \lambda = 45.3$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.86, \lambda = 8.2$)</td>
<td>Gabor ($\theta = 1.57, \lambda = 5.7$)</td>
<td>Intensity mean ($w = 3$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.57, \lambda = 2.8$)</td>
<td>Sobel horizontal edge</td>
<td>Gabor ($\theta = 1.57, \lambda = 22.6$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy ($w = 5$)</td>
<td>Sum entropy ($w = 3$)</td>
<td>Sum variance ($w = 7$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.57, \lambda = 11.3$)</td>
<td>Gabor ($\theta = 2.75, \lambda = 2.8$)</td>
<td>Sum entropy ($w = 7$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.57, \lambda = 22.6$)</td>
<td>Gabor ($\theta = 2.36, \lambda = 11.3$)</td>
<td>Sum variance ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 1.57, \lambda = 11.3$)</td>
<td>Gabor ($\theta = 1.18, \lambda = 22.6$)</td>
<td>Sum entropy ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 2.75, \lambda = 11.3$)</td>
<td>Gabor ($\theta = 0.39, \lambda = 11.3$)</td>
<td>Sum average ($w = 5$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 2.75, \lambda = 22.6$)</td>
<td>Gabor ($\theta = 2.36, \lambda = 8.2$)</td>
<td>Info. measure 1 ($w = 5$)</td>
</tr>
<tr>
<td>Kirsch edge descriptor 3</td>
<td>Kirsch edge descriptor 1</td>
<td>Diff. variance ($w = 5$)</td>
</tr>
<tr>
<td>Kirsch edge descriptor 2</td>
<td>Gabor ($\theta = 0.39, \lambda = 22.6$)</td>
<td>Sum average ($w = 7$)</td>
</tr>
<tr>
<td>Gabor ($\theta = 3.86, \lambda = 8.2$)</td>
<td>Gabor ($\theta = 1.57, \lambda = 11.3$)</td>
<td>Info. measure 1 ($w = 7$)</td>
</tr>
<tr>
<td>Diagonal edge descriptor</td>
<td>Kirsch edge descriptor 3</td>
<td>Sum average ($w = 3$)</td>
</tr>
</tbody>
</table>

**Table 7.3:** Top ten radiomic features selected based on an intraclass correlation coefficient above 0.9 and high AUC values in distinguishing prostate cancer from benign tissue. Features extracted from ADC maps are shown in bold font; features listed in regular font were extracted from T2–weighted MRI.
Figure 7.1: Ground truth extent of prostate cancer is delineated on T2w MRI in red for a representative slice of (left) a PZ tumor and (right) a TZ tumor. Feature maps of the top two selected features populating $F_{12}^{PZ}$ and $F_{12}^{TZ}$, respectively, are also shown. Heatmaps representing the pixel-wise probability of cancer presence, obtained via logistic regression classifiers trained on data from $I_1$ and $I_2$, are shown when $C_{12}^{ALL}$ is used and when $C_{12}^{PZ}$ and $C_{12}^{TZ}$ complement each other to detect cancer in both the PZ and TZ. Red indicates a high probability of cancer presence, yellow indicates a low probability of cancer presence, and blue, the absence of cancer.
7.3.2 Classifier performance

When applied to detect PZ cancer on the studies in $I_1$ and $I_2$, $C_{PZ1}$, $C_{PZ2}$, and $C_{PZ12}$ yielded AUC values ranging between 0.61 and 0.71, whereas these classifiers yielded lower AUC values when applied to studies in $I_3$ (0.54-0.58; see Table 7.3). $C_{PZ1}$, $C_{PZ2}$, and $C_{PZ12}$ detected PZ cancer with higher AUC than $C_{ALL1}$, $C_{ALL2}$, and $C_{ALL12}$, respectively. The difference in AUC between $C_{PZ1}$ and $C_{ALL1}$, as well as between $C_{PZ2}$ and $C_{ALL2}$, was statistically significant ($p < 0.05$). $C_{TZ1}$, $C_{TZ2}$, and $C_{TZ12}$ detected prostate cancer in the TZ on a per-voxel level with AUC values ranging between 0.54 and 0.68, and $C_{ALL1}$, $C_{ALL2}$, and $C_{ALL12}$ yielded statistically similar AUC values (0.53-0.66; see Table 7.4). There were no statistically significant differences in AUC values between $C_{ALL}$ and $C_{TZ}$.

7.3.3 Correlation with Gleason grades

Spearman’s correlation coefficient was used to evaluate the strength of correlations between AUC values representing cancer detection accuracy and Gleason grades associated with prostate lesions. When $C_{TZ12}$ was applied for cancer detection on MRI from $I_1$, AUC values correlated poorly with Gleason grades ($\rho = 0.05$, see Figure 7.2(a)); when $C_{PZ12}$ was applied for cancer detection on MRI from $I_1$, AUC values correlated somewhat with Gleason grades ($\rho = 0.16$, see Figure 7.2(b)).

7.3.4 Effect of training cohort on classifier performance

A two-sample Student’s t-test was used to evaluate whether statistically significant differences existed between AUC values yielded by $C_{TZ1}$, $C_{TZ2}$, and $C_{TZ12}$, as well as
Table 7.4: For all combinations of training and testing datasets, the mean AUC for TZ and PZ cancer detection is shown for the zone–aware ($C^{TZ}$ and $C^{PZ}$) and zone–ignorant ($C^{ALL}$) classifiers. Two-sample t–tests were used to evaluate the statistical significance of differences in AUC between zone–aware and zone–ignorant classifiers.
Figure 7.2: Scatterplot of AUC yielded by the spatially-aware classifier on a per-patient basis and (a) Gleason grades for patients from Institution 1 with TZ tumors, (b) Gleason grades for patients from Institution 1 with PZ tumors, and (c) overall tumor sizes for all 80 patients.
between $C_{1}^{PZ}$, $C_{2}^{PZ}$, and $C_{12}^{PZ}$. When applied to detect cancer on the studies in $I_{1}$, $C_{1}^{PZ}$, $C_{2}^{PZ}$, and $C_{12}^{PZ}$ yielded statistically similar AUC values ($p > 0.14$ for TZ cancer and $p > 0.56$ for PZ cancer). Similarly, in detecting cancer on studies in $I_{2}$ and $I_{3}$, $C_{1}^{PZ}$, $C_{2}^{PZ}$, and $C_{12}^{PZ}$ yielded statistically similar AUC values ($p > 0.55$ for TZ cancer and $p > 0.62$ for PZ cancer).

Pearson’s correlation coefficient was used to evaluate how well AUC values for $C_{1}^{PZ}$, $C_{2}^{PZ}$, and $C_{12}^{PZ}$, as well as $C_{1}^{TZ}$, $C_{2}^{TZ}$, and $C_{12}^{TZ}$, correlated with respect to each other. AUC values for $C_{1}^{PZ}$, $C_{2}^{PZ}$, and $C_{12}^{PZ}$ correlated well with each other on MRI from $I_{1}$ ($\rho = 0.56 - 0.82$), $I_{2}$ ($\rho = 0.69 - 0.87$), and $I_{3}$ ($\rho = 0.85 - 0.98$). Similarly, AUC values for $C_{1}^{TZ}$, $C_{2}^{TZ}$, and $C_{12}^{TZ}$ correlated well with each other on MRI from $I_{1}$ ($\rho = 0.57 - 0.75$), $I_{2}$ ($\rho = 0.64 - 0.81$), and $I_{3}$ ($\rho = 0.82 - 0.99$).

### 7.3.5 Differences between small and large tumors

A two–sample Student’s t–test was used to evaluate whether both small (< 0.5 cm$^{3}$ or < 1 cm$^{3}$ on T2w MRI) and large tumors (> 1 cm$^{3}$ on T2w MRI) share the same characteristics on multi–parametric MRI. After Bonferroni correction for the effect of multiple testing, there were no significant differences between small and large tumors. Cancer detection improved with tumor size, as AUC values were somewhat correlated with tumor size ($\rho = 0.14$, see Figure 7.2(c)). Five small tumors from $I_{1}$ that were <1 cm$^{3}$ were associated with high Gleason grades (>7) and were therefore clinically significant. These tumors ranged in size from 0.35–0.5 cm$^{3}$. Four of these tumors were associated with Gleason scores ranging between 0.58–0.79, while one was associated with an AUC value below 0.5. Of nine tumors smaller than 0.2 cm$^{3}$ (considered the limit for tumor visibility on MRI), four were
detected with an AUC ranging between 0.65–0.83, while the remaining five were associated with AUC values below 0.5.

7.4 Discussion

In this study we identified and evaluated radiomic features associated with TZ and PZ tumors on multi-parametric MRI. We found that distinct sets of radiomic features were useful for cancer detection in the TZ and PZ, respectively. Furthermore, the cancer detection accuracy associated with these features was not significantly different across the three institutions considered in this study ($p > 0.14$). To the best of our knowledge, this was the first study to evaluate the performance of radiomic features for prostate cancer detection on MRI in a multi-institutional setting.

Our findings indicate the importance of Gabor wavelet features, co-occurrence texture features, and edge descriptors for distinguishing prostate cancer from benign prostate tissue in the TZ and the PZ. Combinations of Gabor wavelets and co-occurrence features were previously shown to be useful for both TZ and PZ cancer detection [21, 22]. The multiscale, steerable Gabor wavelets that were dominant in $F_{TZ}$ regardless of training cohort appear to model localized frequency characteristics, thereby distinguishing between the hypo-intense, homogeneous texture of TZ tumors and the more heterogeneous surrounding normal TZ tissue. The co-occurrence features, which dominate $F_{2}^{PZ}$ and play roles in $F_{1}^{PZ}$ and $F_{12}^{PZ}$ have previously been found to be particularly useful for distinguishing between hypo-intense PZ cancer and hyper-intense normal PZ tissue [22]. The dearth of ADC and DCE features in $F^{PZ}$ and $F^{TZ}$ may be related to DWI and DCE MRI data quality.
Regardless of training cohort, $F^{TZ}$ and $F^{PZ}$ did not overlap at all. This finding suggests that multi-parametric MRI radiomic features identified as useful for cancer detection in the PZ were distinct from radiomic features that are useful for cancer detection in the TZ. In fact, it has previously been noted that TZ and PZ tumors tend to differ in appearance and that radiomic features that are useful for detecting TZ tumors are different from those that are useful for detecting PZ tumors [21, 22]. However, past studies involved MRI acquired with an endorectal coil [21, 22]; hence, differences in appearance between TZ and PZ tumors could have been affected by biases induced by the endorectal coil in the PZ. Since our study considered MRI cases obtained with a surface coil, as well as some acquired with an endorectal coil, our findings may be less susceptible to potential biases induced by an endorectal coil.

Accounting for differences between TZ and PZ tumors by identifying unique feature sets $F^{TZ}$ and $F^{PZ}$ and subsequently developing distinct classifiers $C^{TZ}$ and $C^{PZ}$ led to significantly improved cancer detection in the PZ ($p \leq 0.05$). $C^{PZ}$ yielded voxel-wise AUC values ranging between 0.54 and 0.71, whereas $C^{ALL}$ performed no better than random guessing (AUC $\leq 0.51$). In contrast to the PZ, in the TZ $C^{TZ}$ and $C^{ALL}$ performed similarly to each other, providing AUC values ranging between 0.54–0.68. This result may possibly be due to the inherent difficulty in distinguishing between tumors in the TZ and confounding disease, such as benign prostatic hyperplasia, that manifests predominantly in the TZ. As a result, $C^{TZ}$ was not able to discriminate between tumors and benign confounders in the TZ much better than $C^{ALL}$.

The AUC values associated with $C^{TZ}$ and $C^{PZ}$ were lower than those obtained
in other studies [21, 22], which reported voxel-wise AUC values as high as 0.73–0.86 for cancer detection in the PZ and/or TZ. However, the results reported in [21, 22] were all based on cross-validation within a single institution. Our lower AUC could possibly be attributed to the fact that 75% of the tumors in our cohort were < 1 cm\(^3\) in size. Nevertheless, 80% of small tumors that were clinically significant (Gleason grade > 7) were detected on MRI.

The AUC values correlated somewhat with tumors size, as larger tumors were associated with higher AUC values (see Figure 7.2(e)). Also, due to the multi-institutional nature of our study, the MRI acquisition parameters (e.g., b-values, temporal resolution of DCE MRI) differed substantially between institutions. As a result, the classifiers were trained on MRI studies obtained with one set of parameters and independently evaluated on MRI studies obtained at another institution with a very different set of acquisition parameters.

The correlation between Gleason grades and AUC values for all tumors from I\(_1\) was poor in both the TZ and PZ (see Figures 7.2(a,b)). This is not surprising since the classifiers were trained to discriminate between prostate cancer and benign tissue; they were not trained to distinguish between cancer grades. Additionally, since this study was not designed at the outset to separate tumors by Gleason scores, an overwhelming majority of the tumors were associated with intermediate Gleason scores (3+4 or 4+3). However, in our study all high grade tumors were correctly detected, and all but one intermediate grade tumor was detected.

There were several limitations to our study. Firstly, this was a multi-institutional study. While this was a unique aspect of our study, it was also a limitation because
the MRI acquisition parameters differed between institutions. For example, the b-values used to acquire diffusion-weighted MRI, as well as the temporal resolution of the DCE MRI, differed between institutions. Perhaps this lack of consistency may have contributed to why primarily T2-weighted MRI features were chosen during feature selection, whereas most ADC- and DCE-based features were not highly ranked by our feature selection scheme. However, we argue that validity of radiomics-based classification methods really need to be evaluated in a multi-institutional setting, and, as far as we can tell, this study represents the first attempt at just that. Secondly, whereas the MRI data from $I_1$ and $I_2$ was acquired using a body coil, the MRI data from $I_3$ was acquired using an endorectal coil. This may be have contributed to the lower cancer detection accuracy associated with $I_3$, particularly in the PZ, where the effect of the endorectal coil would be most seen (AUC ranged between 0.53–0.64 for TZ and 0.54–0.58 for PZ, compared to 0.54–0.68 for TZ and 0.60–0.71 for PZ). Thirdly, to keep the data unbiased we included all lesions annotated on prostatectomy specimens, regardless of size or MR visibility. Some lesions were only a few voxels large on MRI and suffered from the partial volume effect, making it difficult to register and characterize them correctly.
Chapter 8

Pharmacokinetic Features for Prostate Cancer Detection on MRI

8.1 Overview

In this chapter we develop a novel approach to estimate pharmacokinetic parameters on prostate DCE MRI. Pharmacokinetic modeling on MRI typically relies on knowledge of the arterial input function (AIF), which can be estimated in an artery in the image field–of–view [87]. However, AIF estimation requires that the DCE MRI be acquired at a high temporal resolution, attained at the expense of lowering the spatial resolution. Since high spatial resolution is crucial for detecting small tumors, a variety of techniques have been published that circumvent the AIF while estimating pharmacokinetic parameters [90–94, 98]. However, these methods rely on population–averaged pharmacokinetic parameter values; as a result, pharmacokinetic parameters are biased.
In this chapter we develop an approach to estimate pharmacokinetic parameters in the prostate without relying on any population-averaged values and evaluate its ability to detect and grade prostate cancer on DCE MRI. Our strategy takes advantage of the fact that the prostate is comprised of two primary anatomical regions—the PZ and the TZ—and that quantitative T1 values and pharmacokinetic parameter values differ between the PZ and TZ [104–106]. Our two-step prostate-based method (PBM) is unique in that it is entirely prostate–based because it does not consider any non–prostate voxels. The PBM was compared with an AIF–based method in terms of ability of estimated pharmacokinetic parameters associated with prostate cancer, as determined on DCE MRI following fusion with pathology, to discriminate (a) between prostate cancer and benign tissue and (b) between low, intermediate, and high grade lesions. Additionally, pharmacokinetic parameters obtained via the PBM were further validated based on their ability to approximate the AIF.

8.2 Materials and Methods

8.2.1 Patients

This retrospective study, which was approved by the institutional review board at the University of Turku in Finland, included forty patients who underwent research MRI examination either (a) before their first biopsy, as a part of a two–institutional clinical study [136], or (b) before robotic assisted laparoscopic prostatectomy (RALP), as a part of an ongoing single institutional clinical trial. Patient inclusion criteria are shown in Figure 8.1.
8.2.2 MRI Acquisition

In order to suppress bowel peristalsis, 1.0 mg of glucagon (GlucaGen, Novo Nordisk, Bagsvaerd, Denmark) was given subcutaneously immediately before the start of the examination. Furthermore, patients were instructed to take laxatives in the evening before the MRI examination in order to limit rectal air (Laxoberon, Boehringer Ingelheim, Ingelheim, Germany). The MR imaging was performed using a 3T MR scanner (Magnetom Verio 3T, Siemens Healthcare, Erlangen, Germany). Body array and spine array coils were used as receiver coils, while whole body single channel RF coils was used as transmit coils. No endorectal coil was used. The imaging protocol consisted of triplanar T2-weighted turbo spin-echo imaging, single shot spin-echo based DWI, three-dimensional 1H–MRS, and DCE–MRI. The three–dimensional 1H–MRS data sets and DWI were not analyzed in the current study. The total duration of the MRI examination was about 60 minutes.
Anatomical triplanar T2-weighted images were acquired using a turbo spin-echo sequence with repetition time/echo time (TR/TE) 6400-8640/101 ms, field of view (FOV) 200 × 200 mm², matrix size 320 × 320, slice thickness 3 mm, number of signal averages (NSA) 2 and acquisition time 2.27 min (transverse images), 2.47 min (sagittal images), 2.78 min (coronal images). Generalized autocalibrating partially parallel acquisitions (GRAPPA) [140] was used with the acceleration factor of 2 and 32 reference lines for autocalibration. Axial dynamic contrast enhanced images were acquired before, during and after injection of contrast agent (0.1 mmol/kg Dotarem, Guerbet, France) which was injected 30 s after the beginning of the sequence through a peripheral vein at a rate of 2 ml/s via a mechanical injector (Spectris, Medrad, Indianola, USA). In total, 60 time points at a temporal resolution of 6.9 seconds were acquired using a three-dimensional VIBE sequence [141] with the following parameters: TR/TE 5.43/1.87 seconds, 15-degree flip angle, FOV 240 × 240 mm², matrix size 192 × 192, slice thickness 3 mm, acquisition voxel size 1.25 × 1.25 × 3.0 mm³, a bandwidth of 260Hz/pixel acquiring 6/8 (75%) of k-space in phase-encoding direction, GRAPPA with the acceleration factor of 2 and 24 reference lines, acquisition time 6.9 min. Before the contrast injection, 3D VIBE images with five different flip angles of 2, 5, 8, 10, 15 degrees were obtained for calculation of pre-contrast longitudinal relaxation time (T10). Dynamic contrast enhanced imaging data sets for T10 maps were calculated using an Osirix DCE plugin (http://kyungs.bol.ucla.edu/software/DCE_tool/DCE_tool.html).
8.2.3 Histopathological Analysis

All of the histopathologic material was analyzed by one experienced genitourinary pathologist (8 years of experiences in genitourinary pathology) in consensus with another pathologist (at least 4 years of experiences in genitourinary pathology). Whole mount prostatectomy sections were processed as described in [137] and stained with hematoxylin and eosin. The Gleason score was assigned as a combination of primary, secondary, and tertiary Gleason grade according to the 2005 International Society of Urological Pathology Modified Gleason Grading System [57]. Tertiary Gleason grade was assigned in the presence of Gleason grade pattern higher than the primary and secondary, where the tertiary component was estimated visual to account for less than 5% of the tumor [138]. Lesions with a Gleason score ≤ 6 were considered low grade, lesions with a Gleason score of 3+4 were considered intermediate grade, and lesions with a Gleason score of 4+3 or ≥8 were considered high grade. Among the forty patients in our study cohort, 26 had tumors in the TZ, and 38 patients had tumors in the PZ. A total of 103 individual lesions were annotated on histology, including 35 in the TZ (19 low grade, 8 intermediate grade, and 8 high grade) and 68 in the PZ (40 low grade, 14 intermediate grade, and 14 high grade).

8.2.4 Cancer Annotation on MRI

The first step of the PBM involves estimating pharmacokinetic constants associated with normal TZ and PZ tissue; thus, it was important to have precise maps of cancer and normal prostate tissue on DCE MRI. In order to obtain “ground truth” annotation of prostate cancer extent on MRI, co-registration of MRI and whole
mount histological sections was performed. Correspondences between histological sections and T2w MRI slices were determined by a genitourinary pathologist and radiologist working in unison. Subsequently, corresponding histological sections and MRI slices were co-registered using an interactive B-spline elastic registration scheme [114]. The final result was a labeling of each MRI voxel within the prostate as corresponding to cancer or benign prostate tissue. DCE MRI was spatially aligned with T2w MRI via volumetric affine registration, which corrected inter-acquisition movement and inter-protocol resolution differences. After inter-protocol alignment, all MRI data were analyzed at the T2w MRI resolution.

8.2.5 Pharmacokinetic Modeling on DCE MRI

The right and left femoral arteries, as well as the prostate capsule and TZ, were manually annotated on T2w MRI by a urologic radiologist. The AIF was then obtained by averaging the concentration of contrast agent in the blood plasma across all femoral arterial pixels. Subsequently, AIF-based pharmacokinetic parameters were calculated by leveraging the AIF to solve the Kety-Tofts model [86] for each prostate voxel. In order to assess the effect of temporal down-sampling of the DCE MRI data, every second DCE MRI instance was removed, thereby reducing the effective temporal resolution from 6.9 seconds to 13.8 seconds.

Our novel, prostate-based method (PBM) estimates pharmacokinetic parameters in the prostate without utilizing any non-prostate voxels. The PBM leverages
a reference region model [92] to estimate the pharmacokinetic constants associated with each prostate voxel. Assuming that two regions in the image field–of–view do not share the same pharmacokinetic characteristics, the reference region model provides for the estimation of (a) relative pharmacokinetic parameter ratios between the two regions or (b) pharmacokinetic parameters associated with one region provided that pharmacokinetic parameters associated with the other region are known. The reference region model [92] assumes a simple two–compartment model in which the contrast agent diffuses from the blood plasma into the extravascular–extracellular spaces of two tissues. This system can be described by two differential equations:

\[
\frac{d}{dt}C_{PZ}(t) = K_{PZ} C_p(t) - \frac{K_{PZ}}{v_{PZ}} C_{PZ}(t) \tag{8.1}
\]

\[
\frac{d}{dt}C_{TZ}(t) = K_{TZ} C_p(t) - \frac{K_{TZ}}{v_{TZ}} C_{TZ}(t) \tag{8.2}
\]

where \(C_p(t)\) is the concentration of contrast agent in the blood plasma, also known as the AIF; \(C_{PZ}(t)\) and \(C_{TZ}(t)\) are the concentrations of contrast agent in the PZ and TZ, respectively; \(K_{PZ}\) and \(K_{TZ}\) are the transfer constants representing diffusion of contrast agent from capillaries into the PZ and TZ; and \(v_{PZ}\) and \(v_{TZ}\) are the extravascular–extracellular volume fractions for the PZ and TZ. Combining these two equations and solving the resulting differential equation leads to the following
expressions [92]:

\[
C_{PZ}(t) = \frac{K_{TZ}}{K_{PZ}} C_{TZ}(t) + \frac{K_{TZ}}{K_{PZ}} \left( \frac{K_{TZ}}{v_{TZ}} - \frac{K_{PZ}}{v_{PZ}} \right) \int_0^T C_{TZ}(t) \exp \left( -\frac{K_{PZ}}{v_{PZ}} (T - t) \right) dt \tag{8.3}
\]

\[
C_{TZ}(t) = \frac{K_{PZ}}{K_{TZ}} C_{PZ}(t) + \frac{K_{PZ}}{K_{TZ}} \left( \frac{K_{PZ}}{v_{PZ}} - \frac{K_{TZ}}{v_{TZ}} \right) \int_0^T C_{PZ}(t) \exp \left( -\frac{K_{TZ}}{v_{TZ}} (T - t) \right) dt \tag{8.4}
\]

Our two–step PBM exploits the inherent differences in pharmacokinetic characteristics associated with normal TZ and PZ (see Figure 8.2). In the first step of the PBM, eq. (8.3) was implemented in an unconstrained curve–fitting routine to estimate the ratios \( K_{TZ}/K_{PZ} \), \( K_{TZ}/v_{TZ} \), and \( K_{PZ}/v_{PZ} \). During this step, \( C_{PZ}(t) \) and \( C_{TZ}(t) \) were taken as the contrast agent concentration-versus-time curve averaged over all voxels in the PZ and TZ, respectively. The second step of the PBM used these ratios as constraints in a constrained curve–fitting routine to re–implement eq. (8.3) to estimate \( K_{PZ} \) and \( K_{TZ} \) and get a more accurate estimate of \( K_{PZ}/v_{PZ} \) and \( K_{TZ}/v_{TZ} \). These estimates of the pharmacokinetic constants for the PZ and TZ are region–based and not voxel–specific, but they are patient–specific. Finally, once \( K_{TZ} \) and \( v_{TZ} \) became known, these values were input as constants in eq. (8.3), which was implemented on a per–voxel basis to estimate \( K_{PZ} \) and \( v_{PZ} \) for every voxel in the PZ. Similarly, once \( K_{PZ} \) and \( v_{PZ} \) are known, these values were input as constants in eq. (8.4), which was implemented on a per–voxel basis to estimate \( K_{TZ} \) and \( v_{TZ} \) for every voxel in the TZ. All pharmacokinetic modeling was performed using Matlab R2015b (The Mathworks, Inc., Natick, MA).
8.2.6 Evaluation of Pharmacokinetic Modeling Methods

Pharmacokinetic parameters were evaluated based on (a) their usefulness in discriminating prostate cancer from benign tissue, (b) their usefulness in discriminating cancer grades, and (c) how well they can approximate the true AIF. Differences in pharmacokinetic parameter values associated with normal and cancerous prostate voxels were assessed on a per–patient level using the non–parametric Kruskal–Wallis significance test. The sensitivity and specificity of pharmacokinetic parameters for discriminating between cancerous and benign prostate voxels were assessed for each patient by the area under the receiver operating characteristic (ROC) curve (AUC).
In order to evaluate the usefulness of pharmacokinetic parameters for discriminating cancer grades, ROC analysis was implemented on both a per-voxel and a per-lesion basis. For voxel-based ROC analysis the sensitivity and specificity of voxel-specific $K^{\text{trans}}$ and $v_e$ values for discriminating between low and high grade cancer and between intermediate and high grade cancer was assessed via AUC. For lesion-based ROC analysis, each lesion needed to be associated with a single value of $K^{\text{trans}}$ and $v_e$, although these parameters were computed on a per-voxel level. Therefore, each lesion was assigned a single value for $K^{\text{trans}}$ and $v_e$ based on the 75th or 100th percentile of the values in each lesion. Then, AUC was leveraged to evaluate the ability of $K^{\text{trans}}(75)$, $K^{\text{trans}}(100)$, $v_e(75)$, and $v_e(100)$ to discriminate between (a) low versus high grade lesions, (b) intermediate versus high grade lesions, and (c) low versus intermediate and high grade lesions. Additionally, Pearson’s correlation between $K^{\text{trans}}(75)$, $K^{\text{trans}}(100)$, $v_e(75)$, and $v_e(100)$ and lesion size was used to assess the impact of lesion size on detectability.

Furthermore, pharmacokinetic parameters obtained via the PBM were evaluated based on their ability to approximate the AIF calculated in the femoral arteries. The mean $K^{\text{trans}}$ and $v_e$ values in the prostate, as well as the mean tissue contrast concentration in the prostate, were used as inputs to the Kety-Tofts model, which was solved for the AIF. This approximated AIF was compared to the true AIF calculated in the femoral arteries based on sum-of-squared differences. All statistical analyses were performed using Matlab R2015b (The Mathworks, Inc., Natick, MA).
8.3 Results

8.3.1 Discriminating Cancer from Benign Tissue

Pharmacokinetic parameters associated with cancer and benign TZ and PZ tissue are listed in Table 8.1 for both the PBM and AIF–based method. For the AIF–based method, \( K^{\text{trans}} \) averages 0.40 and 0.32 in normal TZ and PZ, respectively, and decreases to 0.36 and 0.29 in TZ and PZ cancer. Likewise, the average \( v_e \) values are higher in normal TZ than cancerous TZ (0.29 versus 0.24) but are higher in cancerous PZ than normal PZ (0.30 versus 0.23). For the PBM, however, the average \( K^{\text{trans}} \) increases from 0.22 to 0.24 in cancerous TZ, compared to normal TZ, but decreases from 0.87 to 0.81 in cancerous PZ compared to normal PZ; \( v_e \) is the same in cancer and benign regions of both the TZ and PZ. When temporal downsampling decreases the temporal resolution by 50\%, the average \( v_e \) increases while the average \( K^{\text{trans}} \) remains largely unaffected for the AIF–based method; this trend is reversed for the PBM. Regardless of pharmacokinetic parameter estimation method, estimated values for \( K^{\text{trans}} \) and \( v_e \) fell within the range of reported values for normal and cancerous prostate [104, 105, 142].

In order to evaluate the ability of \( K^{\text{trans}} \) and \( v_e \) to discriminate between cancer and benign prostate voxels, Kruskal–Wallis tests were used to evaluate, on a per–patient basis, whether \( K^{\text{trans}} \) and \( v_e \) values are significantly different between cancerous and benign prostate voxels. Out of 27 patients with lesions in the TZ, significant differences between PBM estimates of \( K^{\text{trans}} \) and \( v_e \), associated with cancerous and benign TZ were found in 26 and 23 patients, respectively (see Table
8.2). The one patient for which significant differences between benign and cancerous TZ were not found had a small lesion (< 10 mm). Out of 38 patients with PZ lesions, 33 patients manifested significant differences in $K_{\text{trans}}$ associated with cancerous and benign tissue, and 30 manifested significant differences in $v_e$ between cancerous and benign PZ voxels. The five patients that did not exhibit significant differences between cancerous and benign PZ included two with low grade cancer, one with intermediate grade cancer, and two with high grade cancer, all with lesions >10 mm in size. Fewer patients exhibited significant differences between cancerous and benign PZ and TZ tissue when pharmacokinetic parameters were estimated via the AIF–based approach (see Table 8.2). Additionally, $K_{\text{trans}}$ discriminated between cancerous and benign TZ with an average AUC of 0.53 and between cancerous and benign PZ with an average AUC of 0.59, while AUC values associated with $v_e$ were lower (0.50 in the TZ and 0.56 in the PZ). AUC values were similar when pharmacokinetic parameters estimated via the AIF–based method were used to discriminate cancerous from benign voxels (see Table 8.2).

Figure 8.3 displays heatmap representations of $K_{\text{trans}}$ for three prostate lesions.

<table>
<thead>
<tr>
<th>Method</th>
<th>$K_{\text{trans}}$</th>
<th>$v_e$</th>
<th>TZ cancer</th>
<th>PZ</th>
<th>PZ cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIF (6.9 s)</td>
<td>$K_{\text{trans}}$</td>
<td>0.40± 0.50</td>
<td>0.36± 0.53</td>
<td>0.32± 0.47</td>
<td>0.29± 0.41</td>
</tr>
<tr>
<td>AIF (6.9 s)</td>
<td>$v_e$</td>
<td>0.22± 0.22</td>
<td>0.24± 0.20</td>
<td>0.25± 0.22</td>
<td>0.30± 0.22</td>
</tr>
<tr>
<td>AIF (13.8 s)</td>
<td>$K_{\text{trans}}$</td>
<td>0.41± 0.44</td>
<td>0.37± 0.48</td>
<td>0.35± 0.41</td>
<td>0.35± 0.36</td>
</tr>
<tr>
<td>AIF (13.8 s)</td>
<td>$v_e$</td>
<td>0.47± 0.30</td>
<td>0.48± 0.28</td>
<td>0.39± 0.29</td>
<td>0.40± 0.30</td>
</tr>
<tr>
<td>PBM (6.9 s)</td>
<td>$K_{\text{trans}}$</td>
<td>0.22± 0.19</td>
<td>0.24± 0.14</td>
<td>0.87± 0.54</td>
<td>0.81± 0.51</td>
</tr>
<tr>
<td>PBM (6.9 s)</td>
<td>$v_e$</td>
<td>0.13± 0.10</td>
<td>0.13± 0.07</td>
<td>0.35± 0.20</td>
<td>0.35± 0.19</td>
</tr>
<tr>
<td>PBM (13.8 s)</td>
<td>$K_{\text{trans}}$</td>
<td>0.12± 0.12</td>
<td>0.13± 0.07</td>
<td>0.44± 0.35</td>
<td>0.55± 0.28</td>
</tr>
<tr>
<td>PBM (13.8 s)</td>
<td>$v_e$</td>
<td>0.13± 0.10</td>
<td>0.14± 0.07</td>
<td>0.37± 0.20</td>
<td>0.37± 0.19</td>
</tr>
</tbody>
</table>

**Table 8.1:** For forty subjects and two methods of pharmacokinetic parameter estimation, parameters $K_{\text{trans}}$ and $v_e$ are listed as mean ± standard deviation.
Chapter 8. Pharmacokinetic Features for Prostate Cancer Detection on MRI

Table 8.2: For both the AIF–and prostate–based methods, the number of patients with significant differences between normal prostate zones and cancer, evaluated via Kruskal–Wallis significance tests, are listed. Additionally, AUC values for cancer detection on the basis of pharmacokinetic parameters are shown as mean (min–max).

<table>
<thead>
<tr>
<th></th>
<th>AIF–based</th>
<th></th>
<th>PBM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N patients</td>
<td>TZ</td>
<td>PZ</td>
<td>TZ</td>
<td>PZ</td>
</tr>
<tr>
<td>27</td>
<td>38</td>
<td>27</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>K\text{trans}: p&lt;0.05</td>
<td>26</td>
<td>30</td>
<td>26</td>
<td>33</td>
</tr>
<tr>
<td>K\text{trans}: p&lt;0.01</td>
<td>25</td>
<td>28</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>K\text{trans}: AUC</td>
<td>0.55 (0.18–0.82)</td>
<td>0.59 (0.30–0.94)</td>
<td>0.53 (0.18–0.76)</td>
<td>0.59 (0.34–0.90)</td>
</tr>
<tr>
<td>v\text{e}: p&lt;0.05</td>
<td>21</td>
<td>29</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>v\text{e}: p&lt;0.01</td>
<td>20</td>
<td>26</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>v\text{e}: AUC</td>
<td>0.50 (0.17–0.79)</td>
<td>0.54 (0.28–0.81)</td>
<td>0.50 (0.21–0.78)</td>
<td>0.56 (0.28–0.86)</td>
</tr>
</tbody>
</table>

Table 8.3: Correlation coefficients obtained via Pearson’s correlation between lesion size and 74th and 100th percentiles of K\text{trans} and v\text{e} are shown. Correlations that are statistically significant at a level of \( \alpha = 0.01 \) are shown with an asterisk.

<table>
<thead>
<tr>
<th></th>
<th>AIF–based</th>
<th></th>
<th>PBM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K\text{trans}(100)</td>
<td>v\text{e}(100)</td>
<td>K\text{trans}(75)</td>
<td>v\text{e}(75)</td>
</tr>
<tr>
<td>TZ + PZ</td>
<td>0.52*</td>
<td>0.45*</td>
<td>0.24*</td>
<td>0.20*</td>
</tr>
<tr>
<td>TZ</td>
<td>0.34*</td>
<td>0.23</td>
<td>-0.06</td>
<td>-0.05</td>
</tr>
<tr>
<td>PZ</td>
<td>0.56*</td>
<td>0.48*</td>
<td>0.31*</td>
<td>0.22*</td>
</tr>
</tbody>
</table>

The PBM accounts for differences between enhancement patterns in the TZ and PZ; as a result, K\text{trans} tended to be elevated in the TZ only when cancer was present there. In contrast, the AIF–based approach sometimes yielded ambiguous results in the TZ due to heterogeneous enhancement in the TZ. The 75th and 100th percentiles of pharmacokinetic parameters associated with lesions tended to increase significantly with lesion size (see Table 8.3).

8.3.2 Predicting Prostate Cancer Risk Group

The ability of voxel–wise K\text{trans} and v\text{e} values to discriminate between high, intermediate, and low grade lesions was assessed by ROC analysis (see Figure 8.4). When the AIF–based approach was used to estimate pharmacokinetic parameters,
Figure 8.3: Ground truth extent of prostate cancer is shown for three lesions on the top row. Heatmap representations of $K^{\text{trans}}$ are displayed for the AIF–based approach (middle) and PDM (bottom).
areas under the ROC curves (AUCs) of 0.59 are achieved for distinguishing low grade from high grade lesions in the TZ on the basis of $K_{trans}$ and $v_e$, respectively. When PBM was used, however, AUC values increase to 0.61 and 0.69 for $K_{trans}$ and $v_e$, respectively. When discriminating between low and high grade PZ lesions, AUC decreased from 0.67 to 0.61 for $K_{trans}$ for the PBM but increased from 0.56 to 0.57 for $v_e$ when the PBM is used. In distinguishing intermediate from high grade lesions, $K_{trans}$ estimated using the AIF–based approach led to AUC values of 0.40 and 0.68 in the TZ and PZ, respectively. When the PBM was used to estimate $K_{trans}$, the AUC increased to 0.57 in the TZ but decreased to 0.55 in the PZ (see Figure 8.4(b)). In contrast, $v_e$ provided lower AUC values of 0.35 and 0.52 in the TZ and PZ when the AIF–based approach was used and AUCs of 0.65 and 0.52 in the TZ and PZ when the PBM was used. Thus, whereas the PBM yields higher AUC values for all cancer grading in the TZ, the AIF–based approach yields higher AUC values for cancer grading in the PZ in most cases.

The ability of lesion–wide $K_{trans}(75)$, $K_{trans}(100)$, $v_e(75)$ values to discriminate between high, intermediate, and low grade lesions was also assessed by ROC analysis (see Table 8.4). $K_{trans}(75)$, $K_{trans}(100)$, $v_e(75)$ estimated via the AIF–based method yield higher AUC values when discriminating between low and intermediate grade lesions than those associated with the PBM (0.56–0.60 versus 0.47–0.55 in the TZ and 0.56–0.62 versus 0.37–0.50 in the PZ). In contrast, when discriminating between intermediate and high grade lesions, AUC values are high for the PBM (0.59–0.65) and low for the AIF–based approach (0.17–0.42). In discriminating low from intermediate and high grade TZ lesions, AUC values are as high as 0.57 for the PBM; in discriminating low versus intermediate and high grade PZ lesions,
Figure 8.4: ROC curves illustrate the predictive power of $K_{\text{trans}}$ in discriminating (a) low from high grade cancer and (b) intermediate from high grade cancer and $v_e$ in discriminating (c) low from high grade cancer and (d) intermediate from high grade cancer.
Chapter 8. Pharmacokinetic Features for Prostate Cancer Detection on MRI

<table>
<thead>
<tr>
<th>low vs. mid</th>
<th>PZ</th>
<th>TZ</th>
<th>$K_{\text{trans}}(100)$</th>
<th>$v_e(100)$</th>
<th>$K_{\text{trans}}(75)$</th>
<th>$v_e(75)$</th>
<th>PBM</th>
<th>$K_{\text{trans}}(100)$</th>
<th>$v_e(100)$</th>
<th>$K_{\text{trans}}(75)$</th>
<th>$v_e(75)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>low vs.</td>
<td>0.60</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
<td>0.55</td>
<td>0.54</td>
<td>0.47</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mid</td>
<td>0.62</td>
<td>0.61</td>
<td>0.58</td>
<td>0.56</td>
<td>0.50</td>
<td>0.48</td>
<td>0.40</td>
<td>0.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low vs.</td>
<td>0.54</td>
<td>0.51</td>
<td>0.51</td>
<td>0.50</td>
<td>0.57</td>
<td>0.56</td>
<td>0.49</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mid/high</td>
<td>0.58</td>
<td>0.55</td>
<td>0.51</td>
<td>0.52</td>
<td>0.55</td>
<td>0.52</td>
<td>0.46</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mid vs.</td>
<td>0.18</td>
<td>0.17</td>
<td>0.20</td>
<td>0.19</td>
<td>0.60</td>
<td>0.58</td>
<td>0.62</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>0.44</td>
<td>0.42</td>
<td>0.38</td>
<td>0.42</td>
<td>0.59</td>
<td>0.59</td>
<td>0.63</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.4: AUC values for discriminating between lesions of different cancer grades based on the $75^{th}$ and $100^{th}$ percentiles of $K_{\text{trans}}$ and $v_e$ associated with each lesion.

AUC values are as high as 0.58 for the AIF–based method.

8.3.3 Reverse Approximation of AIF Based on Pharmacokinetic Parameters

When pharmacokinetic parameters obtained via the PBM were used as inputs to the Kety–Tofts model to solve for the AIF, the sum–of–squared differences (SSD) between this approximated AIF and the true AIF calculated in the femoral arteries was small (SSD < 0.15). In comparison, when mean pharmacokinetic parameters estimated via the AIF–based approached were used as inputs to the Kety–Tofts model to solve for the AIF, the sum–of–squared differences was higher (SSD < 0.25), although not significantly so.

8.4 Discussion

In general, pharmacokinetic modeling of DCE MRI relies on knowledge of the AIF, which is typically estimated in an artery in the image field–of–view [87]. Nevertheless, in order to accurately estimate the AIF, the DCE MRI must be acquired at a high temporal resolution, which is attained at the expense of lowering the spatial
resolution. Since high spatial resolution is crucial for detecting small tumors, a va-
riety of techniques have been published that circumvent the AIF while estimating
pharmacokinetic parameters. Many of these methods involve using a population–
averaged AIF [98], parameterized AIF [93, 94], or population–averaged phar-
macokinetic parameter values for a reference region [92]. However, AIFs and phar-
macokinetic parameters vary considerably between subjects, and ignoring the in-
herent inter–patient variation leads to errors in estimating patient–specific phar-
macokinetic parameters [90, 91]. In this paper we presented PBM, a novel ap-
proach for estimating pharmacokinetic parameters in the prostate without relying
on knowledge of the AIF or any population–averaged values. Instead, our method
leveraged differences between pharmacokinetic characteristics of the TZ and PZ in
order to estimate $K_{\text{trans}}$ and $v_e$ on a voxel–wise basis throughout the entire prostate.
We evaluated our approach for pharmacokinetic parameter estimation on both a
lesion–wise and a voxel–wise basis, where voxel–wise prostate cancer extent was
known with high accuracy following fusion of MRI with histological sections.

Although the PBM used only patient–specific prostate voxels, pharmacokinetic
parameter values estimated by PBM were similar to those estimated via the AIF–
based approach and able to discriminate prostate cancer grades with comparable
performance to the AIF–based approach. In reality, pharmacokinetic parameters
computed via different approaches are not necessarily comparable. Nevertheless,
the pharmacokinetic values associated with normal and cancerous TZ and PZ tis-
sue were similar to $K_{\text{trans}}$ and $v_e$ found in other studies [104, 105, 143]. When the
AIF–based method was used for parameter estimation, both $K_{\text{trans}}$ and $v_e$ were
higher in the TZ than in the PZ; this trend was reversed for parameters estimated
Chapter 8. Pharmacokinetic Features for Prostate Cancer Detection on MRI

via the PBM. We found that for most patients, both $K_{\text{trans}}$ and $v_e$ associated with TZ and PZ cancer were significantly higher than benign TZ and PZ, respectively, for both the PBM and the AIF–based method (see Table 8.2). This finding corroborates numerous studies showing that these pharmacokinetic constants are elevated in prostate cancer [e.g., 104, 105, 143] due to their higher vascularity relative to benign prostate tissue [41–43].

Pharmacokinetic parameters are non–specific for detecting cancer due to vascular leakiness associated with benign prostatic hyperplasia in the TZ and prostatitis in the PZ [144]. Also, post–biopsy hemorrhage in either the TZ or PZ can lead to higher pharmacokinetic parameter values. As a result, it is not surprising that benign regions, as well as cancerous regions, were associated with suspiciously high $K_{\text{trans}}$ and $v_e$ values (see Figure 8.3) and that AUC values for discriminating between cancerous and benign prostate voxels were no higher than 0.55 in the TZ and 0.59 in the PZ. Nevertheless, we found that pharmacokinetic parameters have benefit for discriminating between prostate cancer grades. In general, the PBM yields higher AUC values (up to AUC = 0.65) for discriminating between intermediate and high grade lesions, whereas the AIF–based approach yields higher AUC values (up to AUC = 0.60) for discriminating between low and intermediate grade lesions. $K_{\text{trans}}$ and $v_e$ estimated via PBM each individually distinguished between low and intermediate/high grade prostate cancer with AUC values up to 0.57 in the TZ and 0.55 in the PZ. In comparison, Vos et al. [142] achieved an AUC of 0.72 for discriminating between low and intermediate/high grade cancer when combining $K_{\text{trans}}$ with several other kinetic features derived from DCE MRI. Our finding that the PBM does as well as or better than the AIF–based approach in
most scenarios, while at the same time remaining completely independent of the AIF or any population–averaged measures, suggests that the PBM may be highly desirable.

There are several limitations to our approach. Firstly, there is no reliable gold standard for determining accuracy of pharmacokinetic parameters on DCE MRI. As a result, it is challenging to compare pharmacokinetic constants yielded by disparate approaches and to identify how accurately pharmacokinetic parameter estimates match up with the physiological parameters that they are supposed to measure. Nevertheless, pharmacokinetic parameter values yielded by the PBM are within the range of values obtained by other studies [104, 105, 143]. Furthermore, when these pharmacokinetic parameters are used to estimate the AIF, the AIF is approximated well. Secondly, this preliminary study involved a small cohort of only 40 patients, and further validation of the PBM is necessary on a larger cohort. Thirdly, we found that the effect of decreasing the temporal resolution on estimates of $K_{\text{trans}}$ was different from its effect on estimates of $v_e$. Further investigation into the effect of temporal downsampling is necessary.

In conclusion, we have presented a novel approach for estimating patient–specific pharmacokinetic parameters in the prostate without relying on the AIF or any population–averaged values. We found that $K_{\text{trans}}$ and $v_e$ are beneficial for distinguishing between low, intermediate, and high grade prostate lesions in both the TZ and PZ. These findings suggest that it may be possible to perform pharmacokinetic modeling on prostate DCE MRI without the need to rely on an AIF or population–averaged values for a reference region.
Chapter 9

Concluding Remarks

In this dissertation we have presented novel strategies for building robust and interpretable classifiers when the feature dimensionality is high, a commonality of many medical imaging problems. Specific applications were shown in the context of (a) prostate cancer diagnosis on multi-parametric MRI, (b) predicting risk of post-radiotherapy prostate cancer biochemical recurrence T2w MRI, and (c) predicting prognosis of patients who underwent therapy for breast and prostate cancers based on digital pathology images. Additionally, we also presented novel strategies for prostate zone-specific prostate cancer diagnosis on multi-parametric MRI in a multi-institutional study and for assessing prostate cancer aggressiveness based on DCE MRI via pharmacokinetic modeling. The major goals accomplished in this work include:

- Development of two novel strategies—Variable Importance in Projections (VIP) and Feature Importance in Nonlinear Embeddings (FINE)—that overcome the curse of dimensionality without compromising on the interpretability of a classifier constructed using high-dimensional features. VIP ranks features based on their importance for classification in a linear projection of the data,
while FINE ranks features based on their importance for classification in a nonlinear data projection.

- Application of VIP for (a) identifying multi-parametric MRI features useful for detecting transition zone (TZ) and peripheral zone (PZ) prostate cancers on multi-parametric MRI and (b) identifying T2w MRI texture features that are useful for predicting biochemical recurrence risk following radiotherapy for prostate cancer. To the best of our knowledge, our work is the first to evaluate differences in characteristics of TZ and PZ tumors on multi-parametric MRI and to evaluate risk of biochemical recurrence based on T2w MRI alone.

- Application of FINE for identifying quantitative histomorphometric features useful for predicting (a) treatment outcome and risk assessment of estrogen receptor positive (ER+) breast cancers and (b) risk of biochemical recurrence of prostate cancers following radical prostatectomy.

- Development and multi-institutional evaluation of radiomic feature sets for cancer detection in the TZ and PZ, respectively. To the best of our knowledge, this was the first study to evaluate the performance of radiomic features for prostate cancer detection on MRI in a multi-institutional setting.

- Development and preliminary evaluation of a novel approach for estimating pharmacokinetic parameters in the prostate by leveraging inherent differences between pharmacokinetic characteristics of the TZ and PZ without relying on knowledge of the AIF or any population-averaged values.

In Chapter 3 VIP and FINE were introduced as methods to overcome the limitations of both traditional feature selection schemes and dimensionality reduction.
schemes to identify useful features for classifier construction in high–dimensional spaces. By ranking features based on the strengths of their contributions to classification in a reduced feature subspace obtained via nonlinear dimensionality reduction, FINE can be used for feature selection, much as any other filter method is used for feature selection. Nevertheless, it is important to note that FINE is very different from other filters. Firstly, FINE is the only filter that ranks features based on their roles in (a) defining the geometry of a non–linearly derived embedding and (b) driving accurate classification. Secondly, FINE can be used for feature discovery post facto. That is, once DR has already been done and/or a classifier has already been constructed in a low dimensional embedding space, FINE provides insight into which of the original, high dimensional features actually contributed most prominently to classification in the embedding space.

In Chapters 4-6 we demonstrated the usefulness of VIP and FINE to overcome the curse of dimensionality when building classifiers with high–dimensional features. In Chapter 4, VIP was leveraged to identify stable sets of radiomic features extracted from MRI that are useful for diagnosing prostate cancer in both the TZ and PZ. VIP was implemented in Chapter 5 to identify radiomic features that are useful for predicting the risk of biochemical recurrence of prostate cancer following radiation therapy. In comparison to both the Kattan nomogram and semantic MRI features, the three radiomic features identified by the VIP scheme, in conjunction with a logistic regression model predicting biochemical recurrence risk, led to more accurate differentiation between patients who later developed biochemical recurrence and those who remained recurrence–free. The fact that our methods outperform the Kattan nomogram in spite of the poor quality of the MRI data used
in this study suggests that multi-parametric MRI with improved spatial and temporal resolution will be even more useful for predicting biochemical recurrence following treatment. Chapter 6 described the ability of FINE to identify quantitative histomorphometric features that are useful for predicting the risk of breast and prostate cancers. Measures of nuclear and glandular architecture and clusteredness were found to play an important role in predicting the likelihood of recurrence of both breast and prostate cancers.

Chapter 7 described a multi-institutional study that identified radiomic features extracted from multi-parametric MRI for diagnosing TZ and PZ tumors. The radiomic features identified as useful for cancer detection in the PZ were different from those that were useful for TZ cancer detection. A zone-aware classifier was found to significantly improve the accuracy of cancer detection in the PZ, compared to a classifier that ignorant of differences between prostate zones. This finding suggests that decision support tools for evaluating prostate MRI exams should take into account image and radiomic feature differences between TZ and PZ tumors.

In Chapter 8 we presented a novel method for estimating pharmacokinetic parameters in the prostate without relying on knowledge of the AIF or any population-averaged values. We found that $K_{trans}$ and $v_e$ are elevated in prostate cancer regardless of prostate zone and are beneficial for distinguishing between low, intermediate, and high grade prostate lesions in both the TZ and PZ. These findings suggest that it may be possible to perform pharmacokinetic modeling on prostate DCE MRI without the need to rely on an AIF or population-averaged values for a
reference region. Nevertheless, this study was only preliminary and more extensive validation is necessary.

In conclusion, medical images, including multi-parametric MRI and digital pathology images, have considerable potential for diagnosing prostate cancer, assessing cancer grade, and predicting risk of recurrence following treatment. However, a prerequisite for clinical adoption of computerized decision support tools is feature transparency and classifier interpretability since physicians are typically resistant to opaque "black box" prediction models. The methods developed and evaluated in this work enable computerized decision support tools to be more robust and consistent across institutions— and also interpretable. Thus, these methods have significant translational and clinical benefits towards improving cancer detection and prognosis both prior to and following treatment.
Chapter 10

Future Work

In future work we would extend the feature selection strategies in this work to make them applicable to a wider variety of dimensionality reduction techniques. VIP and FINE, respectively, enable performing linear and nonlinear dimensionality reduction without compromising on the interpretability of the classifier in the reduced feature space. FINE can be applied when dimensionality reduction is performed by solving an eigenvalue problem. In future work we would extend FINE to embeddings obtained by solving a generalized eigenvalue problem (e.g., canonical correlation analysis, weighted extensions of principal component analysis). Additionally, since FINE ranks features based on their contributions to classification on an embedding, features selected based on their FINE scores may be correlated and redundant. In order to make FINE a desirable feature selection method, we would overcome this problem by introducing a redundancy term that penalizes the selection of redundant features.

In this work we demonstrated the benefit of using the VIP and FINE framework to identify radiomic features extracted from multi-parametric MRI and quantitative histomorphometric features extracted from digital pathology features that are
useful for diagnosing and predicting prognosis of prostate and breast cancers. In future work we intend to apply this methodology to identify radiomic features extracted from multi-parametric MRI that are useful not only for diagnosing prostate cancer, but also for distinguishing between prostate cancer grades. Also, future studies will assess the benefit of placing maps of the computer-extracted features identified via FINE alongside multi-parametric MR images in decision support systems to aid radiologists in prostate cancer detection.

Additionally, in this work we performed a preliminary study to identify radiomic features extracted from T2w MRI that predict biochemical recurrence risk following radiation therapy with higher accuracy than the current standard. However, this study involved a cohort of only sixteen subjects, and more extensive evaluation will need to be performed in future work. We would also investigate radiomic features extracted from DCE MRI and DWI, as well as from T2w MRI, for predicting biochemical recurrence following radiotherapy. Additionally, we developed a novel approach for estimating pharmacokinetic parameters associated with prostate cancer. This preliminary study included only 40 patients, most of whom had intermediate grade disease; future work would evaluate this approach on a larger patient cohort.
Chapter 11

Appendix A: Classification Strategies

11.1  Notation

We define a voxel in an image as $z \in Z$, where $Z$ is a 3–dimensional grid of voxels. Each $z \in Z$ is associated with a label $y(z) \in \{0, 1\}$ and a feature vector $F(z) \in \mathbb{R}^m$, which is comprised of $m$ quantitative features extracted from that image voxel. Prior to classifier training, these feature vectors are concatenated and centered to create a data matrix $X \in \mathbb{R}^{n \times m}$, whose rows represent $n$ samples (usually voxels) and whose columns represent $m$ features. A vector $Y$ of outcome labels is comprised of the classification labels associated with each of the $n$ samples. We define a classifier that is trained to predict $Y$ given $X$ as $C \in \{LR, NB, SVM, RF\}$, where LR is a logistic regression classifier, NB is a naive Bayes classifier, SVM is a support vector machine, and RF is a random forest classifier.
11.2 Classifier strategies

11.2.1 Logistic regression (LR) classifier

The LR classifier [15] is a probabilistic classifier that is commonly used to model binary responses in biomedical applications. LR is based on the desire to model the posterior probabilities of a sample belonging to a particular class using linear functions of $x$. When only two classes are considered, the LR model has the form

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \sum_{j=0}^{m} \beta_j x,$$  \hspace{1cm} (11.1)

where the coefficients $\beta$ are estimated using maximum likelihood. The output of the LR classifier is the posterior probability that each sample belongs to class $y = 1$. When this probability is greater than 0.5, a sample is assigned to class $y = 1$; when this probability is less than 0.5 the sample is assigned to class $y = 0$.

11.2.2 Naive Bayes (NB) classifier

Similar to the LR classifier, the NB classifier is a Bayesian classifier that is commonly used when the feature dimensionality is high [15]. The NB classifier assumes independence between features $x_j$, and although this assumption is often unmet NB classifiers often outperform more complex classification models [15]. When only two classes are considered, the NB model has the form

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \log \frac{\gamma_1 D_1(x)}{\gamma_0 D_0(x)},$$  \hspace{1cm} (11.2)
Chapter 11. Appendix A: Classification Strategies

a direct result of Bayes’ theorem. Here $\gamma_0$ and $\gamma_1$ are the prior probabilities of a sample belonging to class 0 and 1, respectively. $D_1(X) = \prod_{j=1}^{m} D_{j1}$ and $D_0(X) = \prod_{j=1}^{m} D_{j0}$, where $D_{j0}$ and $D_{j1}$ are the individual class-conditional marginal probability density functions associated with classes 0 and 1, respectively. In our implementation, these class-conditional probability density functions were modeled as Gaussian distributions. The output of the NB classifier is the posterior probability that each sample belongs to class $y = 1$. When this probability is greater than 0.5, a sample is assigned to class $y = 1$; when this probability is less than 0.5 the sample is assigned to class $y = 0$.

11.2.3 Support vector machines (SVM)

The SVM [15] tries to find the optimal hyperplane that best separates classes 0 and 1 in the feature space. The SVM first projects the data into a larger space of dimension $\mu$, $\mu > m$ by using a kernel function that computes inner products in the transformed feature space. We use a radial basis kernel to project the data into a higher dimensional space. Then the SVM finds the hyperplane that best separates the two classes in this transformed space. Thus, unlike the LR and NB classifiers, which provide posterior probabilities of a sample belonging to each class, the SVM yields only the predicted class $\hat{y} \in \{0, 1\}$ associated with each sample.

11.2.4 Random forest (RF) classifier

The RF classifier [15] combines classification decisions from multiple decision trees that are bootstrap aggregated (bagged). Each individual decision tree classifier breaks up the feature space in multiple partitions in a multi-stage process. The
decision tree classifier begins by finding a single split point to partition the feature space into two classes; this split point is usually chosen to minimize the sum–of–squares classification error. At each stage, each partition is further split into two partitions. Thus, the decision tree creates a tree–like structure for determining the class associated with each sample. When the RF bags the decisions of $B$ decision trees, the probability that a sample belongs to class 1 is modeled by the RF as

$$P(y = 1) = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b(x),$$

(11.3)

where $\hat{y}_b(x) \in \{0, 1\}$ is the predicted class yielded by the $b$th decision tree classifier. Thus, the more decision tree classifiers that predict a sample belongs to class 1, the higher the RF output. When this probability is greater than 0.5, a sample is assigned to class $y = 1$; when this probability is less than 0.5 the sample is assigned to class $y = 0$. 


Chapter 12

Appendix B: Bias Field Correction

MR image formation can be modeled as

\[ v(x) = u(x)f(x), \quad (12.1) \]

where \( v \) is the measured signal at location \( x \), \( u \) is the true signal, \( f \) is an unknown bias field that varies smoothly, and we assume that the image is noise–free \[113\]. Bias field correction involves estimating \( f \) and then removing its effect.

In order to make the model (12.1) additive, the intensities can be log–transformed so that

\[ \hat{v}(x) = \hat{u}(x) + \hat{f}(x), \quad (12.2) \]

where the notation \( \hat{\cdot} = \log(\cdot) \). Assuming that \( \hat{u} \) and \( \hat{f} \) are approximately independent random variables, the distribution of their sum can be found by convolution, as follows:

\[ V(\hat{v}) = F(\hat{v}) * U(\hat{v}) = \int F(\hat{v} - \hat{f})U(\hat{f})d\hat{f}. \quad (12.3) \]
Given a distribution $U$, the bias field $F$ can be estimated as follows. The expected value of $\hat{u}$ given a measurement $\hat{v}$ at some arbitrary location $x$ is:

$$E(\hat{u}|\hat{v}) = \int_{-\infty}^{\infty} \hat{u} \frac{p(\hat{u}, \hat{v})}{p(\hat{v})} = \frac{\int_{-\infty}^{\infty} \hat{u} F(\hat{v} - \hat{u}) U(\hat{u}) d\hat{u}}{\int_{-\infty}^{\infty} F(\hat{v} - \hat{u}) U(\hat{u}) d\hat{u}}.$$  \hfill (12.4)

Thus $\hat{f}$ can be estimated from (12.4) as follows:

$$\hat{f}_x(\hat{v}) = E(\hat{f}|\hat{v}) = \hat{v} - E(\hat{u}|\hat{v}),$$  \hfill (12.5)

where $\hat{f}_x$ is an estimate of $\hat{f}$ at location $x$.

The N3 bias field correction algorithm involves an iterative approach. First a distribution for $U$ is proposed by sharpening the distribution $V$. Then, a corresponding bias field is estimated via equation (12.5) to yield a distribution that is similar to the proposed distribution $U$. Now that an estimate of the bias field $F$ has been obtained, the distribution $U$ can be estimated by using a de-convolution filter. These steps of estimating $F$ and $U$ are subsequently repeated until some convergence criterion is met. Finally, the estimated bias field is smoothed before its effect is removed to yield a bias field-corrected MR image.
Chapter 13

Appendix C: MR Intensity Standardization

Intensity standardization [62] involves identifying a set of landmarks on image grayscale intensity histograms so that each landmark shares the same tissue-specific meaning across multiple patients and regardless of scanner. First a standardization scheme is learned on a set of template images, and then this standardization scheme is applied to all images in need of intensity standardization. The standardization scheme is learned via the following multi-step approach:

1. For each 3D template prostate MR image, the intensities of interest in the prostate and surrounding regions are selected by dilating a mask of the prostate so that the mask includes the prostate, as well as 30 voxels surrounding the prostate in all directions. Image intensities associated with all un-masked voxels comprise the intensities of interest.

2. For each template image, three landmarks are identified: the minimum intensity, denoted as $p_1$, the maximum intensity, denoted as $p_2$, and the median intensity, denoted as $\mu$. If there are multiple template images, these landmarks
are chosen by averaging the minimum, median, and maximum intensities over all template images.

3. All intensities from the range \([p_1, p_2]\) are linearly mapped to a new “standard” scale with range \([s_1, s_2]\). Each \(x \in [p_1, p_2]\) is mapped to \(x' \in [s_1, s_2]\) via the following equation:

\[
x' = s_1 + \frac{x - p_1}{p_2 - p_1}(s_2 - s_1).
\] (13.1)

Once the standardization scheme has been learned on the template image(s), it can be applied to standardize another MR image by performing a piecewise linear mapping of the image intensities to the standard scale. Each \(x \in [p_1, p_2]\) is mapped to \(x' \in [s_1, s_2]\) via the following equations:

\[
x' = \mu + (x - x_{50})\frac{s_1 - \mu}{p_1 - \mu} \text{ if } x \leq \mu \tag{13.2}
\]

\[
x' = \mu + (x - x_{50})\frac{s_2 - \mu}{p_2 - \mu} \text{ if } x \geq \mu, \tag{13.3}
\]

where \(x_{50}\) is the median intensity within the intensity range of interest for the image that is being standardized.
Bibliography


62. Madabhushi A and Udupa J. New methods of MR image intensity standard-

63. Gurcan M, Bucheron A, Can A, Madabhushi A, Jajpoot N, and Yener B.
Histopathological image analysis: a review. IEEE Reviews in Biomedical En-

64. Wang H, Cruz-Roa A, Basavanhally A, et al. Mitosis detection in breast can-
cer pathology images by combining handcrafted and convolution neural net-

graphs: predicting biochemical recurrence in intermediate–risk prostate can-

66. Huang P and Lai Y. Effective segmentation and classification for HCC biopsy

67. Petushi S, Garcia F, Haber M, Katsinis C, and Tozeren A. Large–scale com-
putations on histology images reveal grade–differentiating parameters for

the prediction of prostate cancer progression after radical prostatectomy. Jour-

arrhythmia using dimensionality reduction methods on transform domain


