AN INTEGRATED COMPANION DIAGNOSTICS ASSAY FOR PREDICTING BIOCHEMICAL RECURRENCE FOLLOWING RADICAL PROSTATECTOMY

BY GEORGE C. LEE

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

Graduate Program in Biomedical Engineering

Written under the direction of

Anant Madabhushi

and approved by

_____

_____

_____

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2014

ABSTRACT OF THE DISSERTATION

An Integrated Companion Diagnostics Assay for Predicting Biochemical Recurrence following Radical Prostatectomy

by George C. Lee

Dissertation Director: Anant Madabhushi

The most common treatment of prostate cancer (CaP) is via radical prostatectomy (RP), of which 75,000 are performed in the United States each year. However, within the current paradigm, 15-40% of RP treatments ultimately fail in the form of biochemical recurrence (BCR) within 5 years. Gleason scoring, derived from visual inspection of tissue morphology, has been the gold standard for distinguishing aggressive CaP for over 40 years. Furthermore, the current initiative towards personalized health care has attempted to utilize an integrated predictor via molecular markers such as prostate specific antigen (PSA) to identify men with aggressive localized CaP. However, the non-specificity of these tests has led to an over-treatment of CaP, which is responsible for increased morbidity that is both stressful and costly for the patient.

This dissertation attempts to develop the algorithms that could pave the way for a new class of integrated predictors, which can combine histomorphometric and molecular features into an integrated biomarker and present the information needed for better patient care. Our overall goal was to predict BCR in CaP patients following RP treatment. A host of novel machine learning tools were developed to create integrated diagnostic tests, including dimensionality reduction (Adaptive Dimensionality Reduction with

Semi-Supervision (AdDReSS)) and data integration (Supervised Multi-view Canonical Correlation Analysis (sMVCCA)) methodologies to handle complex, non-linear, high dimensional and heterogeneous biomedical data. Furthermore, the development and discovery of unique discriminatory features for differentiating aggressive CaP were necessary for the understanding of cancer progression and the foundation of an integrated biomarker. Novel histomorphometric features (Co-occurring Gland Tensors (CGTs) and Cell Orientation Entropy (COrE)) were developed to quantify important differentiating image-based characteristics of CaP morphology. These methods were shown to outperform Kattan nomogram and Gleason scoring for predicting BCR following RP. Lastly, fusion of histomorphometry and protein expression into an integrated signature was performed via sMVCCA, and demonstrated improved identification of men with BCR following RP compared to histomorphometric and proteomic signatures alone.

Preface

This dissertation represents the collective published and unpublished works of the author. It is primarily composed from the content of peer-reviewed journal [1–4] and conference [5–9] articles (both published or under review), that were written by the author of this dissertation over the course of his thesis work.

Acknowledgements

This work has been a collaborative effort between Rutgers University and our clinical collaborators from the Hospital at the University of Pennsylvania (HUP): Drs. Michael D. Feldman, Stephen R. Master and Natalie N.C. Shih, SUNY Buffalo's Dr. John E. Tomaszewski and my longtime advisor, Dr. Anant Madabhushi who has since moved onto Case Western Reserve University. Your efforts and guidance have been essential to the success of this work.

Sincere gratitude to each of my committee members for their interest and insightful comments on my dissertation work. They have contributed greatly towards the quality of this work.

Thanks to all the faculty, students, and staff at Rutgers University who have provided me guidance or have otherwise made my life easier during my time at Rutgers, as well as during my time at Case Western Reserve University. Your support has been most appreciated.

Special thanks to all my colleagues from the former Laboratory of Computational Imaging and Bioinformatics (LCIB) in New Jersey, and now at the Computational Center of Imaging and Personalized Diagnostics (CCIPD) in Cleveland, who have been largely responsible for the sanity of the author's work, and the author himself.

Lastly, I must acknowledge my funding resources which has supported me throughout the years. This work was made possible via grants from the Department of Defense W81XWH-12-1-0171 and the National Cancer Institute of the National Institutes of Health under award numbers R01CA136535-01, R01CA140772-01, R43EB015199-01, and R03CA143991-01.

Dedication

To my dearest family and friends who have provided me motivation and support along this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

The future of diagnostics will not reside in the discovery of any single predictive variable, but in the ability to analyze a large host of predictors. Thus, the goal of this work is to develop the framework for integrated diagnostics to guide treatment and improve patient outcomes. Much of the work in this dissertation has been applied towards the prediction of aggressive prostate cancer (CaP), however the proposed data representation and integration methodologies are extensible towards a breadth of applications. Clinicians with access to many different channels of information may see their patients benefit from computational analysis of the multivariate data.

This dissertation is structured around the development of an integrated biomarker which combines information from quantitative histomorphometry (QH) and protein expression for an integrated diagnostics test of CaP recurrence following radical prostatectomy (RP) treatment, as defined by biochemical recurrence (BCR). This dissertation consists of 3 main aims summarized by Figure 1.1.

- In Aim 1, we develop machine learning tools needed to create integrated diagnostic tests.
- In Aim 2, we develop quantitative histomorphometric features to quantify characteristics of prostate tissue for predicting biochemical recurrence.
- In Aim 3, we leverage the tools from the previous aims to create an integrated biomarker for predicting biochemical recurrence in prostate cancer patients following radical prostatectomy.

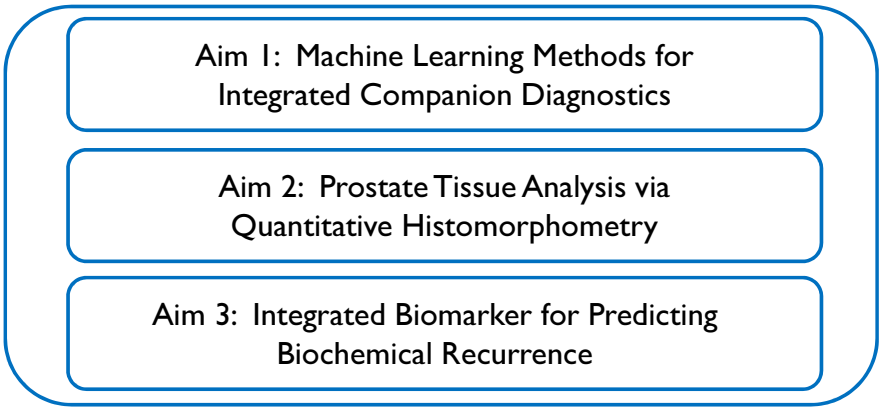


Figure 1.1: Flowchart of the proposed thesis objectives. Aim 1 covers the development of machine learning tools such as data representation and data integration, necessary for creation of an integrated diagnostics test. Aim 2 represents the identification of QH features from prostate whole mount histology. Aim 3 leverages previous methodologies and combines quantitative histomorphometry and protein expression to construct an integrated biomarker for predicting biochemical recurrence in CaP patients following RP.

1.2 Prognosis of men undergoing Radical Prostatectomy treatment for Prostate Cancer

Each year in the United States, nearly 75,000 patients diagnosed with prostate cancer (CaP) undergo radical prostatectomy (RP) treatment [11]. In cases for which there is no prior evidence of spread, treatment of CaP with RP has been largely successful [12]. However, for 15-40% of RP patients, biochemical recurrence (BCR) occurs within 5 years of surgery [13]. BCR is commonly defined as a detectable persistence of prostate specific antigen (PSA) of at least 0.2 ng/ml and is suggestive of recurring aggressive CaP necessitating further treatment [14]. Therefore, it is important to identify aggressive recurrent cancer prior to BCR in order to provide best treatment.

Gleason grading [15,16] is a qualitative system based on the visual analysis of glandular and nuclear morphology to grade CaP aggressiveness. The breakdown in gland shape and organization represents the hallmark of aggressive CaP. The Gleason grading [15,17] system is the most commonly used system in the United States for diagnosis of aggressivity of CaP, quantifying histological patterns of decreasing differentiation, from 1 (most differentiated and benign) and to 5 (least differentiated and malignant).

Gleason score (GS) is currently the gold standard for assigning CaP aggressiveness and is one of the main predictors of BCR. Gleason scoring combines the grade of the most common and second most common pattern, resulting in a Gleason sum ranging from 2 (least aggressive) to 10 (most aggressive).

High Gleason score 8-10 tends to be correlated with more biologically aggressive disease and worse prognosis for long-term, metastasis-free survival [13, 18, 19]. Similarly, GS 8-10 is correlated with BCR [19] and often secondary treatment is provided to accompany RP based on the identification of high GS. Meanwhile, patients with GS 6 typically have a very low incidence of BCR and would not indicate a need for secondary treatment. Unfortunately, outcomes of intermediate GS 7 cancers can vary considerably [20], and statistical tables suggest a 5-year BCR-free survival rate as low as 43% in these men [21]. Furthermore, GS is subject to considerable inter-reviewer variability [22], which can make accurate prognosis of disease more difficult. Therefore, we can determine that prognostic value of GS alone for predicting BCR in RP patients with intermediate-risk GS appears to be limited.

In addition to GS, many postoperative nomograms have been developed to incorporate additional clinical variables such as tumor staging, pre-operative PSA, or positive surgical margins [21, 23–25]. The Kattan nomogram [23] incorporated these parameters to predict 80 month BCR free survival following radical prostatectomy. Han et al. [21] incorporated this information for Johns Hopkins Hospital (JHH) into a series of probability tables, known as the Han Tables, based on Gleason sum, tumor stage, and pre-operative PSA. Subsequently, the Stephenson nomogram [24] added the date of surgery as a prognostic variable. The University of California at San Francisco built their own risk score predictor (CAPRA) [26, 27] for postoperative CaP patients differentiating patients into low, medium, and high risk categories, which also included the percentage of positive biopsy cores into their risk assessment. Hinev et al. [25] performed an independent study advocating the use of the Memorial Sloan Kettering Cancer Center (MS-KCC) nomogram, developed by Kattan and Stephenson, suggesting superior prediction of 5-year BCR compared to JHH predictive tables. The MS-KCC

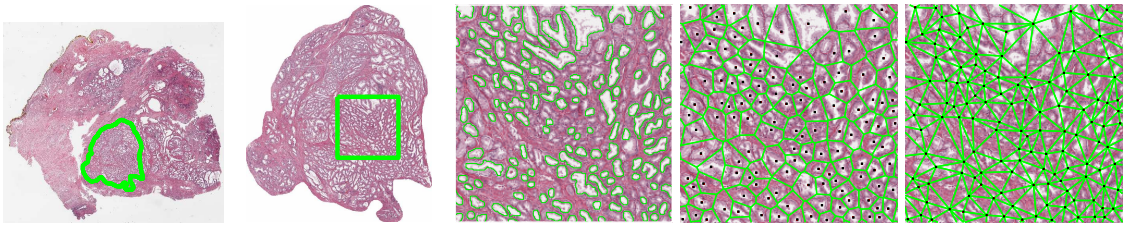
nomogram adds additional variables such as age and time free of cancer. These nomograms represent the state-of-the-art in post-operative CaP prediction of BCR, but still rely heavily on rudimentary clinical variables such as Gleason sum or age.

While current predictive models incorporating prominent clinical markers play an important role in guiding CaP patients towards treatment with better outcomes, previous work has shown that clinical markers yielded just 71% classification accuracy in the prediction of BCR [28, 29]. The technology exists for patients to obtain multi-modal, multi-scale imaging and molecular profiles to characterize prostate cancer. Given availability of more sophisticated diagnostic modalities and the uncertainty involved in predicting BCR using current models, there is an opportunity to develop stronger predictors of prostate cancer outcome.

1.3 Histologic image-based evaluation of prostate cancer

Digital pathology has allowed computer vision and image analysis tools to quantitatively assess tissue morphology and architecture without the aid of a pathologist. Quantitative histomorphometric (QH) features are computationally derived and are therefore not subject to the intra- and inter-observer variability associated with Gleason scoring [22]. In prostate histology (Figure 1.2 (a)), glandular shape and arrangement are key characteristics of the Gleason scoring system [30]. Automated tools for gland segmentation can be used as shown in Figure 1.2 (c) to capture the shape and size of glands, while additional features to quantify various arrangements of glands can be obtained via the construction of Voronoi and Delaunay graphs shown in Figures 1.2 (d) and (e). In contrast to Gleason scoring, QH captures these characteristics in a repeatable, and automated manner.

These features derived from the Gleason grading system have been leveraged to build automated systems for Computer Aided Diagnosis capable of predicting Gleason grade [31]. Previously, commercial QH-based prognostic tests such as Prostate Px+ and Post-Op Px (by Aureon Biosciences Inc.) [28, 32] have indicated the ability to assess survival prognosis from CaP within 5 to 7 days via analysis of digital pathology,



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

Figure 1.2: (a) Prostate whole mount histology with cancerous region (shown in green) annotated by a pathologist. (b) Annotated region of interest with a green bounding box to demonstrate QH features. (c) Gland morphology captured by automated segmentation of the interior lumen boundary. (c) Voronoi diagram and (d) Delaunay triangulation of gland centroids (shown in red) describe architecture.

marking the viability of QH features for predicting BCR.

1.4 Advancement of molecular markers for prostate cancer diagnostics

Many prominent markers for prostate cancer that have been discovered have been genetic-based. PCA3 has gained substantial notoriety as a new oncogene capable of predicting CaP [33]. Another notable genetic biomarker is Annexin A3 (ANXA3), which has shown an inverse relationship to prostate cancer progression [34]. Systems such as ConfirmMDx (by MDx Health^{TM}), are epi-genetics based used for identifying prostate cancers following negative biopsies.

In addition to genetic markers, many researchers have begun to explore mass spectrometry-based proteomic biomarkers [35–38]. Mass spectrometry is an analytical technique that produces spectra of the masses of particles from a sample of interest. The peptide or protein can be subsequently determined based on the spectra, as it provides an isotopic signature of the sample.

The reasons for the interest in mass spectrometry are as follows. Firstly, given the availability of the prostate tissue following prostatectomy, a direct snapshot of the post-transcriptional state offers a stronger inference to the functional state of the cancer [39]. Secondly, routine clinical diagnostics have focused on immunohistochemistry (IHC) of

formalin-fixed paraffin embedded tissue rather than nucleic acid based analysis, allowing for a smoother translation into clinical practice [40]. Therefore, while genetic assays have been successful for producing prognostic tests (ie. Mammaprint [41] and Oncotype DX) [42, 43], mass spectrometry represents a more appropriate modality for both exploratory and prognostic purposes regarding BCR in CaP.

As such, researchers have been looking at proteomic markers for predicting CaP [36, 38, 44]. Ki-67, a cell proliferation protein has been found to be correlated with Gleason scoring [38]. More recently, CD34, a vascular protein, has been found to predictive of BCR in prostate cancer. [45]. Furthermore, PSA and other proteomic markers [46–48] have been shown to be predictive of BCR. It is will be of great importance to continue to explore mass spectrometry methods to mine and identify biomarkers of BCR from thousands of proteins.

1.5 A Call for integrated diagnostics for personalized medicine

The future of personalized medicine will be dependent on integrated diagnostics to leverage the vast amount of medical data available to us and provide better treatments for our patients. Recently, there has been a call to combine multiple prognostic markers to create an integrated biomarker, with potentially greater accuracy in predicting BCR compared to any individual marker [32, 49–52]. Clarke et al. [34] has noted in his investigation of prostate cancer biomarkers that 'no one marker by itself is adequate for detecting all cases'. This suggestion has also been found in other domains such as for the prediction of cardiovascular [53, 54] and pulmonary disease [55]. These studies have all suggested that the integration of biomarkers from a wide-range of modalities can provide better predictors and reduce costs and mortality [56].

However, few attempts have been made in integrated diagnostics to identify BCR in prostate cancers. Post-Op Px has previously been developed by Aureon [57] to estimate BCR following surgery and analyze both the prostate image and its stained immunofluorescence. However, this methodology relies heavily on immunohistochemistry for which antibodies must be developed in order to properly stain the desired protein. Although

integrated diagnostics tests remain in its infancy, given the predictive value of imaging and molecular profiling, the joint predictive value of emerging tests has the potential to provide oncologists with a much more accurate assessment of patient outcomes.

1.6 Summary of the major goals of this thesis

This dissertation consists of a body of work which aims to develop the tools for an integrated diagnostics framework in order to identify BCR in men following RP treatment of CaP. Figure 1.1 illustrates the interplay between aims towards developing an integrated diagnostic test. Investigation of quantitative histomorphometry and protein biomarkers as well as machine learning methods for data representation and integration are all essential components towards creating an integrated diagnostic test for biochemical recurrence in prostate cancer patients following radical prostatectomy. We are able to demonstrate in a preliminary test cohort of radical prostatectomy patients that an integrated biomarker can yield improved classification accuracy compared to the state-of-the-art data integration strategies and compared to QH and protein biomarkers alone for predicting aggressive CaP.

1.7 Organization of this dissertation

The remainder of this dissertation is organized as follows. In Chapter 2, we provide background on the technical motivations and previous work used to support the innovations in this dissertation. As such, in Chapter 3, we discuss a novel dimensionality reduction method, Adaptive Dimensionality Reduction with Semi-Supervision (AdDReSS), for data representation. In Chapter 4, we provide theory and intuition behind supervised Multi-view Canonical Correlation Analysis (sMVCCA) for data integration. The machine learning methods described in the previous chapters will be integrated towards the task of integrated diagnostics for predicting aggressive prostate cancer via quantitative histomorphometry and protein expression. In Chapter 5, we introduce novel developments in quantitative histomorphometry via COrE and CGT algorithms. Background and previous work of QH and its application in prostate cancer prediction

are provided followed by a methodological overview of COrE and CGTs. Evaluation of these methodologies is performed by testing their ability to identify aggressive CaP in patients with radical prostatectomies as compared to state-of-the-art predictors such as Gleason score and Kattan nomograms.

Chapter 6 includes experimental results for using sMVCCA, the developed data integration methodology, to build integrated biomarkers for prostate cancer from QH features and protein expression. The integrated biomarker is subsequently used to construct an integrated diagnostics system for predicting biochemical recurrence in following radical prostatectomy. Lastly, in Chapter 7, we provide concluding remarks and future directions pertaining to this work.

Chapter 2

Technical challenges in developing an integrated diagnostics test

Incorporation of thousands of features to predict CaP inevitably necessitates computational intervention. Development of machine learning tools will be necessary to tackle 2 main issues that will arise from attempts to create an integrated diagnostic test via an integrated biomarker: *Data Representation*, *Data Integration*, and *Classification*.

2.1 Challenges in data representation of high-dimensional, non-linear biomedical data

2.1.1 Combating the 'curse of dimensionality'

Prediction using a large amount of features is computationally expensive and often produces poor predictive accuracy in practice, particularly if many of these features are redundant and uninformative [58, 59]. This has become an increasingly relevant problem given the emergence of gene- and protein-expression profiling for disease prognostication [60–62]. Attempts at analyzing several thousand dimensional gene- and protein- profiles have been primarily motivated by two factors; (a) identification of individual informative genes and proteins responsible for disease characterization [63–66], and (b) to classify patients into different disease cohorts [67–73]. Several researchers involved in the latter area have attempted to use different classification methods to stratify patients based on their gene- and protein-expression profiles into different categories [67–69, 74–86]. While the availability of studies continues to grow, most protein- and gene-expression databases contain no more than a few thousand patient samples. Thus, the task of stratifying these patients based on the gene/protein profile is subject

to the 'curse of dimensionality' problem [58, 87], owing to the relatively small number of patient samples compared to the size of the feature space. In many cases, such as biomedical data, where there exists a relatively sparse database of only a few hundred patients, a low dimensional embedding representation can alleviate the problems of the curse of dimensionality [1, 88]. For example, machine learning classifiers often are unable to generalize a training model that can predict on the high-dimensional data as the large feature space is difficult to generalize for future samples [89, 90]. Additionally, many of the features within the expression profile may be non-informative or redundant, providing little additional class discriminatory information [70, 71] while increasing computing time and classifier complexity.

2.1.2 The role of feature selection for pruning the high-dimensional feature space

Feature selection refers to the identification of the most informative features and have been commonly utilized to precede classification in gene- and protein-expression studies [70, 73, 74]. For the purpose of discriminating object classes, feature selection methods [91–96] have been proposed to limit the overall size of feature set to features which will contribute towards improved classification accuracy. However, since a typical gene or protein microarray records expressions from thousands of genes or proteins, the cost of finding an optimal informative subset from several million possible combinations becomes a near intractable problem. Further, genes or peptides that were pruned during the feature selection process may be significant in stratifying intra-class subtypes. Due to the existence of slightly overlapping features, the task of removing redundant features while retaining informative features via feature selection may lack a complete solution.

2.1.3 The need for dimensionality reduction for a low dimensional transformation

Dimensionality reduction (DR) refers to a class of methods that transforms the high-dimensional data into a reduced subspace to represent data in far fewer dimensions.

These low dimensional embedding representations are useful for presenting high dimensional data. In Principal Component Analysis (PCA), a linear DR method, the reduced dimensional data is arranged along the principal eigenvectors, which represent the direction along which the greatest variability of the data occurs [97]. Note that unlike with feature selection, the samples in the transformed embedding subspace no longer represent specific gene- and protein-expressions from the original high-dimensional space, but rather encapsulate data similarities in low-dimensional space. Even though the objects in the transformed embedding space are divorced from their original biological meaning, the organization and arrangement of the patient samples in low-dimensional embedding space lends itself to data visualization and classification. Thus, if two patient samples from a specific disease cohort are mapped adjacent to each other in an embedding space derived from their respective high-dimensional expression profiles, then it suggests that the two patients have a similar disease condition. By exploiting the entire high-dimensional space, DR methods, unlike feature selection, offer the opportunity to stratify the data into subclasses (e.g. novel cancer subtypes). Furthermore, by using 3 or fewer dimensions, the information can be visually interpreted by plotting the information. Thus, the "curse of dimensionality" problem for classification combined with a visual interpretation of data suggests that it is useful to reduce the number of features M to an $m << M$ feature representation in order to obtain meaningful results.

The most popular method for DR for bioinformatics related applications has been PCA [62,98–104]. Originally developed by Hotelling [105], PCA finds orthogonal eigenvectors along which the greatest amount of variability in the data lies. The underlying intuition behind PCA is that the data is linear and that the embedded eigenvectors represent low-dimensional projections of linear relationships between data points in high-dimensional space. Linear Discriminant Analysis (LDA) [97], also known as Fisher Discriminant Analysis, is another linear DR scheme which incorporates data label information to find data projections that separate the data into distinct clusters. Multidimensional Scaling (MDS) [106, 107] reduces data dimensionality by preserving

the least squares Euclidean distance in the low-dimensional space. Classifier performance with linear DR schemes for biomedical data has been a mixed bag. Dawson et al. [100] found that there were biologically significant elements of the gene expression profile that were not seen with linear MDS. Ye et al. [87] found that LDA gave poor results in distinguishing disease classes on a cohort of 9 gene expression studies. Truntzer et al. [101] also found limited use of LDA and PCA for classifying gene- and protein-expression profiles of a diffuse large b-cell lymphoma dataset since the classes appeared to be linearly inseparable. The afore-mentioned results appear to suggest that biomedical data has a nonlinear underlying structure [100, 101] and that DR methods that do not impose linear constraints in computing the data projection might be more appropriate compared to PCA, MDS, and LDA for classification and visualization of data classes in gene- and protein-expression profiles.

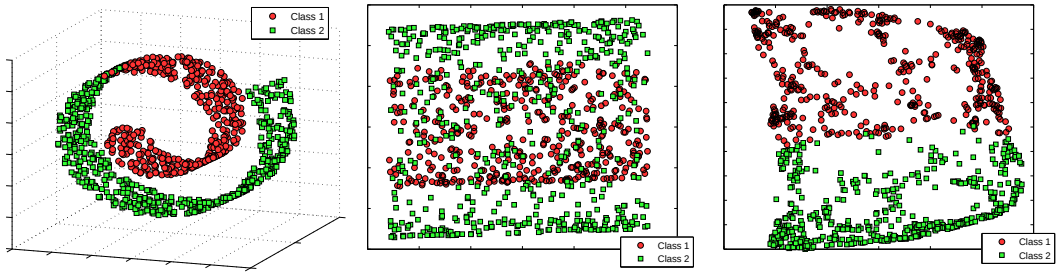
2.1.4 Accounting for non-linearity in biomedical data

Recently, nonlinear DR methods such as Graph Embedding [108], Isometric mapping (Isomap) [10], Locally Linear Embedding (LLE) [109], and Laplacian Eigenmaps (LEM) [110] have been developed to reduce data dimensionality without assuming a Euclidean relationship between data samples in the high-dimensional space. Shi and Malik's Spectral Clustering algorithm utilizes Graph Embedding [108]) to partition the graph into clusters and segment images accordingly. Madabhushi et al. [111] demonstrated the use of graph embedding to detect the presence of new tissue classes on high-dimensional prostate MRI studies. The utility of this scheme has also recently been demonstrated in distinguishing between cancerous and benign magnetic resonance spectra (MRS) in the prostate [112] and in discriminating between different cancer grades on digitized tissue histopathology [31]. Tenenbaum (Isomap) [10] presented the Isomap algorithm for nonlinear DR and described the term 'manifold' for machine learning as a nonlinear surface embedded in high-dimensional space along which dissimilarities between data points are best represented. The Isomap algorithm estimates geodesic distances, defined as the distance between two points along the manifold, and preserves

the nonlinear geodesic distances (as opposed to Euclidean distances used in linear methods) while projecting the data onto a low-dimensional space. Locally linear embedding proposed by Roweis and Saul [109] uses local weights to preserve local geometry in order to find the global nonlinear manifold structure of the data. The geodesic distance between data points is approximated by assuming that the data is locally linear. Recently, Belkin et al. presented the Laplacian Eigenmaps algorithm [110], which like Spectral Clustering, Isomap, and LLE, makes local connections, but uses the Laplacian to simplify determination of the locality preserving weights used to obtain the low-dimensional data embeddings. Graph Emedding, LLE, Isomaps, and LEM, all aim to nonlinearly project the high-dimensional data in such a way that 2 objects x_a and x_b that lie adjacent to each other on the manifold are adjacent to each other in the low-dimensional embedding space, and likewise, 2 objects that are distant from each other on the manifold are far apart in the low-dimensional embedding space.

As previously demonstrated by Tenenbaum [10], Figure 2.1 reveals the limitations of using a linear DR for highly nonlinear data. Figure 2.1 shows the embedding of the swiss roll dataset shown in Figure 2.1(a) obtained by a linear DR method (MDS) in Figure 2.1(b) and a nonlinear DR scheme (LEM) in Figure 2.1(c). MDS, which preserves Euclidean distances, is unable to capture the non-linear manifold structure of the swiss roll, but LEM is capable of learning the shape of the manifold and representing points in the low-dimensional embedding by estimating geodesic distances. Thus, while MDS (Figure 2.1(b)) shows overlap between the two classes that lie along the swiss roll, LEM (Figure 2.1(c)) provides an unraveled swiss roll that separates the data classes in two-dimensional embedding space. Table 2.1 summarizes some popular linear and non-linear DR methods.

While PCA remains the most popular DR method for bioinformatics related applications [98, 100–104], nonlinear DR methods have begun to gain popularity [62, 74, 111, 113]. Liu et al. [74] found high classification accuracy in the use of kernel PCA (nonlinear variant of PCA) for gene expression datasets while Weng [113] recommended



(a) (b) (c)

Figure 2.1: (a) Nonlinear manifold structure of the Swiss Roll dataset [10]. Labels from 2 classes (shown with black circles and red crosses) are provided to show the distribution of data along the manifold. (b) The low-dimensional embedding obtained via linear MDS on the Swiss Roll reveals a high degree of overlap between samples from the two classes due to the use of Euclidean distance as a dissimilarity metric. The embedding obtained via LEM on the other hand, is able to almost perfectly distinguish the two classes by projecting the data in terms of geodesic distance determined along the manifold.

the use of Isomap for medical data analysis. Shi and Chen [62] found that LLE outperformed PCA in classifying 3 gene expression cancer studies. Dawson et al. [100] compared Isomap, PCA, and linear MDS for oligonucleotide datasets, and Nilsson et al. [114] compared Isomap with MDS in terms of their ability to reveal structures in microarray data. In these and other related studies [62, 100, 113, 114], the nonlinear methods were found to outperform linear DR schemes.

2.2 Challenges in data integration of high dimensional, multi-scale medical data

2.2.1 Previous work in data integration

Disease diagnosis, more and more, routinely involves acquiring information from multiple streams and it is now widely appreciated that the quantitative integration of different, independent channels of data into a single classifier could potentially allow for more accurate disease prognosis and response prediction. However, some major problems in data integration have been discovered in reconciling the large diversity in dimensionalities and scales across the different heterogeneous modalities while simultaneously improving separation between the desired object classes.

Type	Citation	Dimensionality Reduction Method
Linear	Hotelling (1933) [105]	Principal Component Analysis (PCA)
	Cox et al. (1988) [107]	Multi-Dimensional Scaling (MDS)
	Fisher et al. (1958) [115]	Linear Discriminant Analysis (LDA)
	Bingham et al (2001) [116]	Random Projections (RP)
Non-linear	Hinton et al. (2002) [117]	Stochastic Neighbor Embedding (SNE)
	van der Maaten et al. (2008) [118]	t-SNE
	Tenenbaum et al. (2000) [10]	Isometric Mapping (ISOMAP)
	Roweis et al. [109] (2000)	Locally Linear Embedding (LLE)
	Brand et al. [119] (2002)	Manifold Charting
	Belkin et al. (2002) [110]	Laplacian Eigenmaps
	Shi et al. (1999) [108]	Graph Embedding
	Kakkonen et al. (1999) [120]	Self-Organizing Maps (SOM)
	Scholkopf et al. (1997) [121]	Kernel PCA
	Zhang et al. (2004) [122]	Local Tangent Space Alignment (LTSA)
	Demartines et al. (1997) [123]	Curvilinear Component Analysis (CCA)

Table 2.1: Summary of linear and non-linear dimensionality reduction methods

As such, methods for combining different modules of information for classification have been suggested [5, 124–126]. In such cases, a meta-space representation is useful to fuse data sources into a combined feature space. Much like dimensionality reduction, integration of independent modalities to a low dimensional integrated data representation offers the advantage of lower computational complexity in the data representation while retaining the discriminatory information from both data sources.

A major barrier towards constructing classifiers from an integration of biomedical data is that the information typically is of high dimensionality [88]. High throughput mining of -omics data via mass spectrometry and DNA microarrays has allowed for the quantification of the expression levels of tens of thousands of molecules [1]. Meanwhile, the increasing use of digital images from radiologic and histologic imaging scans have led to the use of computers to automatically analyze the images [31, 125]. The automated nature of quantitative morphometry allows for calculation of hundreds of statistics to describe the texture, shape, and organization of objects in the image [31]. A combination of data (COD) approach [124], creates a concatenated feature vector comprising all features extracted. However, the combination of high-dimensional features further exacerbates the 'curse of dimensionality' [58] problem of having too many features compared to sample size, which makes building a generalizable classifier difficult. Principal Component Analysis (PCA) [105] has been utilized to reduce the dimensionality of high

dimensional data [98], however, for data fusion, the differences in dimensionalities can bias the attribute vector towards a single modality [88].

Another major limitation in constructing classifiers that can leverage multiple data streams lies in the fact that the data exists at different scales [52]. Thus, the ability to harness useful information from these high dimensional modalities is a non-trival task, necessitating computational machine learning methodologies for heterogenous and orthogonal data integration [52]. Another approach utilizes a combination of interpretations (COI) [124], which collects weak classifier outputs from each modality, to form a joint decision. An example of this is evidence accumulation [127], which combines different types of information via a similarity matrices formed by multiple data clusterings. However, this approach may not be able to leverage the synergy between different modalities [5, 7].

2.2.2 A 'meta-space' representation for integrated diagnostics

A possible solution to overcome these representational differences is to first project the data streams into a meta-space where all data is represented is an equal scale and dimensionality [5]. Generalized Embedding Concatentation (GEC) [5, 7] is a data fusion method which leverages DR methods form a homogenous meta-space. Supervised implementations of embedding fusion include consensus embedding (CE) [128] and boosted embedding concatenation (BEC) [7]. CE aims to combine embeddings via a majority voting scheme, selecting only discriminatory projections. BEC borrows from the Adaboost classifier [129], which evaluates and weights weak embeddings prior to fusion. However, combination of embedding space may still result in potentially noisy or redundant features which may compromise the final fused representation [88].

2.2.3 A multi-kernel learning framework for data integration

Another possible alternative to using embeddings is via the use of high-dimensional kernels [130]. Kernels represent a dot product representation of each modality which can be combined to create a fused representation of heterogenous data [125]. [130] investigated different kernel representations for the purpose of combining data from amino

acid sequences, protein complex, gene expression, and protein interactions. [126] employed a multi-kernel learning (MKL) framework which incorporates labels to model contributions of semantic, boundary support, and contextual sources for the purpose of object localization. [125] proposed semi-supervised multi-kernel (SeSMiK) graph embedding, which combines MKL with a non-linear DR scheme, graph embedding [108], for constructing a fused meta-space representation of multi-protocol MRI data. Semi-supervised graph embedding subsequently incorporates class label information to achieve greater separation of cancer and non-cancerous regions in the low dimensional, fused data representation. However, overfitting can occur with small training samples, leading to inaccurate weighting of the kernels and reduced classifier performance [88, 131].

2.2.4 Canonical correlation analysis for integration of heterogeneous features

To overcome issues with dimensionality, scale, and kernel-based weighting, we explore a family of methods related to Canonical Correlation Analysis [132] which aim to analyze the inter-dependencies between the data sources. CCA [132] is a multivariate statistical method which finds a linear subspace in which correlation between two sets of variables are maximized. In finding a correlated meta-space for data fusion, CCA provides a representation that maximizes the signal, which is likely to be common to data from multiple modalities/views, while minimizing noise which is more likely to be modality specific. Previous work [133] has shown that CCA converges to linear discriminant analysis (LDA) when class labels are transformed into a matrix and provided as one of the two views. However, CCA and LDA can only account for two sets of variables, and thus can only maximize either signal or class separation. Regularized CCA (RCCA) [134, 135], attempts to prevent overfitting in smaller data samples by adding a small positive constant to elements within ill-conditioned matrices. However, regularization is computationally very expensive [88]. Furthermore, like CCA and LDA, RCCA is limited in their ability to consider only two sets of variables at a time. Multi-view (MV) CCA [136, 137] extends the traditional CCA for more than two views by finding a linear subspace which maximizes pairwise correlations between all views.

While these methods have been useful for combining images with text [138], by finding the most correlated features, its utility towards fusing heterogeneous data for the purpose of classification has been limited since there is no guarantee that correlated features are necessarily discriminative. Supervised RCCA (SRCCA) [88] instead utilizes statistical tests to optimize the regularizers in order to achieve greater classification performance compared to RCCA. However, SRCCA is also computationally expensive due to the use of regularization. In Chapter 4, we present a new scheme entitled Supervised Multi-View Canonical Correlation Analysis (sMVCCA) that simultaneously seeks to maximize the signal from any number of heterogeneous data channels and to find a subspace that is also discriminative of the object classes.

2.3 Classification of High Dimensional Biomedical Data

2.3.1 Supervised dimensionality reduction for object class separation of data

While unsupervised methods of dimensionliaty reduction have been utilized for preliminary analysis of data, for classification tasks, it is desirable to incorporate available object class labels to optimize the embedding for class separation, as opposed to basing the affinities solely based off the pre-defined similarity criterion [88,139,140]. Recently, there has been a great deal of interest in semi-supervised dimensionality reduction (SSDR) methods, which utilize labeled instances to improve separation of object classes in the low dimensional embedding [141–144]. This is typically done by extending the pairwise affinity matrix of previous DR methods to incorporate class label information, such that if a pair of objects are of the same class, they are weighted to be more similar and will be mapped to be closer together in the embedding. Similarly, if a pair of objects are of different classes, they are weighted to be less similar and will be mapped further away in the embedding. Sugiyama et al. [141] applied semi-supervised learning (SSL) to Fisher's discriminant analysis in order to find the linear projection that maximized object class separation. Verbeek et al. [145] utilized a method for semi-supervised learning using Gaussian fields with locally linear embedding for object pose recognition.

Yang et al. [142] similarly applied SSL toward manifold learning methods. Zhao [143] presented a semi-supervised method for graph embedding which utilizes weights to simultaneously attract samples of the same class labels and repel samples of different class labels given a neighborhood constraint. Zhang [144] employed a similar approach to SSDR as Zhao, but without utilizing neighborhood constraints.

2.3.2 Constructing better classifiers via active learning

In addition to the high dimensionality and small sample size, another challenge with building predictors for biomedical data is that very often, biomedical datasets are not adequately labeled or annotated [55]. This is due to the significant overhead involved in procuring well-documented biomedical datasets and also due to the fact that invariably an expert is required to perform this task [146]. Hence, if one is attempting to build a predictor to identify disease aggressiveness or predict long term outcome in a patient, one would need a well curated and annotated dataset to provide training labels for the predictor. Active learning can reduce the number of samples needed to train an accurate predictor.

Active learning (AL) is a specific instance of semi-supervised learning, where the learning algorithm may interactively query the desired labels from a user or other source [147]. AL differs from random sampling, which queries training instances randomly from an unlabeled pool [148]. The objective of AL is to find an optimal training set. The benefits of using AL are twofold as 1) classifier accuracy can be improved, and 2) the number of training labels necessary to achieve a classification goal is reduced. For example, several studies [149,150] have found that training with difficult to classify samples can reduce generalization error. This is similar to the idea behind support vector machines (SVMs) [151], where borderline samples can provide a decision boundary which is more generalizable for classification. Chen et al. [55] showed that using a probability based uncertainty sampling approach can reduce the number of annotations for the purpose of clinical text classification. The Query by Committee (QBC) approach [149] uses disagreement across several weak classifiers to identify hard to classify samples. Doyle et al. [146], successfully used QBC [149] to query difficult to classify

regions of mostly cancerous prostate tissue, which resulted in fewer manual pathologist annotations and improved cancer classification. In [152], a geometrically based AL approach utilized SVMs to identify confounding samples, defined as those that lay closest to the decision hyperplane. Liu et al. [148] showed how SVM-based active learning can outperform random sampling for classifying gene expression datasets.

2.4 Novel contributions of this dissertation

We summarize the novel contributions included in this dissertation.

- Data representation via semi-supervised dimensionality reduction is an important problem. We explore the effects and contributions afforded via active learning to improve upon data representation of biomedical data for the purpose of classification and object class discrimination.
- Novel quantitative histomorphometry tools for prostate cancer via co-occurring gland tensors (CGTs) and cell orientation entropy (COrE) are detailed. Both methods have demonstrated better predictive accuracy compared to previously developed QH methods for prostate cancer.
- A data integration framework is proposed for leveraging the information in quantitative histomorphometry and protein expression levels from mass spectrometry. An integrated histologic and proteomic biomarker can be constructed using a new methodology, supervised multi-view canonical correlation analysis (sMVCCA), for the purpose of creating an integrated diagnostics test. Its efficacy in predicting biochemical recurrence is demonstrated in this work.

The development of these tools serve to create the groundwork for future integrated companion diagnostic tests for prostate cancer and other applications and to foster the potential of personalized medicine.

Chapter 3

Adaptive Dimensionality Reduction with Semi-Supervision (AdDReSS) for Classifying Multi-attribute Biomedical Data

3.1 Overview

In this chapter, we present a novel dimensionality reduction (DR) method, AdDReSS (Adaptive Dimensionality Reduction with Semi-Supervision), which aims to seamlessly integrate semi-supervised dimensionality reduction (SSDR) and active learning (AL). This allows AdDReSS to construct low dimensional data representations to improve classification of high dimensional biomedical data while using fewer labels compared to previous SSDR methods. The spirit of AdDReSS is embodied in Figure 3.1. The goal is to separate these two classes in a lower dimensional embedding representation such that each class is in a distinct region of the low dimensional embedding space.

While active learning has been used for providing fewer, optimal instances for training a classifier for biomedical classification, its extension towards learning the best training instances for improving the quality of low dimensional embedding representations has not been heavily investigated [6, 145]. Zhang et al. [153] has suggested that searching in a locally linear or manifold space could provide more representative points for active learning, while Verbeek et a. [145] has suggested an integrated framework using Gaussian fields. A generalizable evaluation and extension of active learning towards semi-supervised dimensionality reduction methods would be important for prediction and representation of biomedical data.

AdDReSS applies the theory behind AL towards the embedding space by identifying difficult to classify samples from the embedding representation. These samples are subsequently used to train the semi-supervised agglomerative graph embedding (SSAGE)

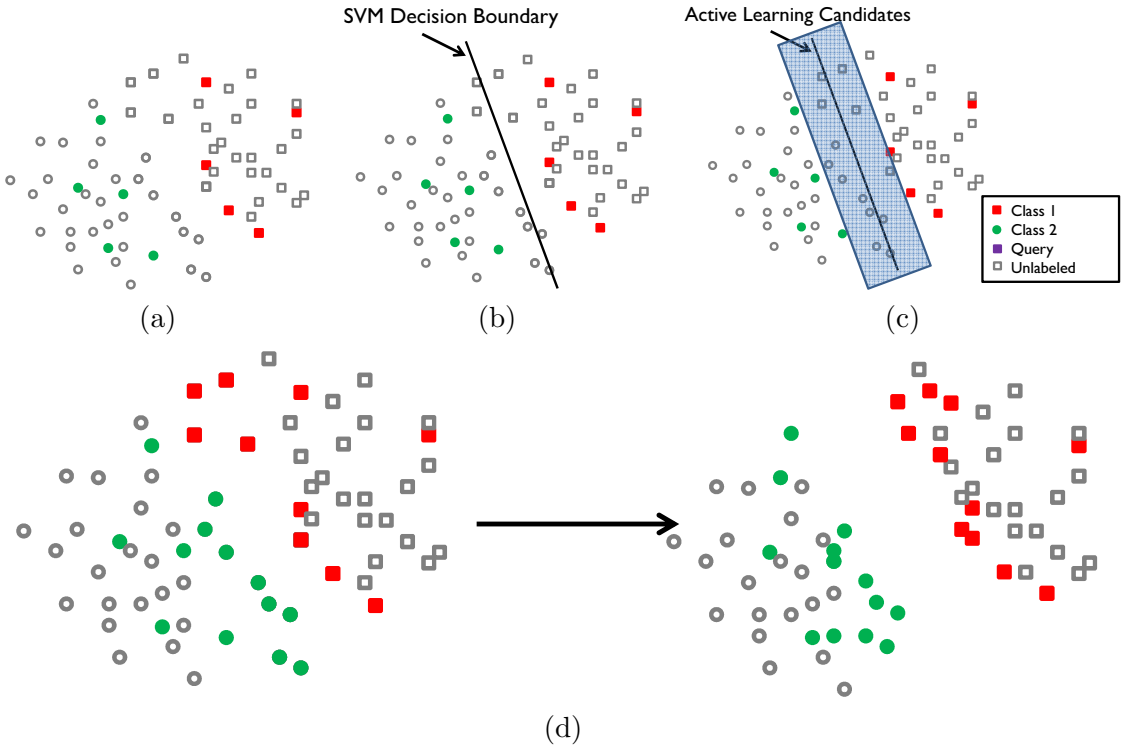


Figure 3.1: An example of how AdDReSS improves embedding by incorporating AL. (a) The original embedding representation given by SSDR. (b) A support vector machine classifier is used as an active learner. (c) samples found to be difficult to classify are selected as candidates for training. (d) SSDR trained on the labels queried by AL provide greater separation of object classes in the embedding.

to produce a more separable representation of the data. This process can be iterated to further refine the embedding representation.

The major contributions and implications of this work are:

1. a novel NLDR method which seamlessly incorporates active learning and semi-supervised learning to guide embedding construction,

2. a demonstration showing the effects of active learning towards improving embeddings generated via SSDR compared to random sampling,

3. a simple framework that could be extensible for other SSDR methods to create more discriminatory low dimensional representations.

We evaluated our methodology on three relevant medical datasets (a Brain MR Imaging dataset [154], a gene expression dataset [76], and a protein expression dataset [78]).

These datasets were chosen to represent varied types of imaging and non-imaging biomedical data - radiologic medical imaging, DNA microarray, and proteomic spectra. We also use two synthetic datasets for showcasing AdDReSS: one imaging (Toy) and one non-imaging (Swiss Roll). Our experimental design was constructed to highlight the differences between embeddings generated via three schemes: (1) AdDReSS, an SSDR method using active learning (2) Semi-Supervised Agglomerative Graph Embedding (SSAGE), an SSDR method which utilizes random sampling and (3) Graph Embedding (GE), an unsupervised NLDR method which does not use any label information.

The rest of this paper is organized as follows. In Section 3.2, we formalize notation and provide an overview of an unsupervised dimensionality reduction method (Graph Embedding), a semi-supervised dimensionality reduction method (Semi-Supervised Agglomerative Graph Embedding), and an active learning strategy (Uncertainty Sampling), thereby providing the theoretical background for AdDReSS. In Section 3.3, we describe our method AdDReSS (Adaptive Dimensionality Reduction with Semi-Supervision) in greater detail. In Section 3.4, we outline the datasets, training parameters, and the evaluation measures used to compare the methodologies described in this work. In Section 3.5, we demonstrate the performance of the comparative methodologies for each of our evaluation measures, followed by concluding remarks in Section 3.6.

3.2 Review of relevant machine learning techniques

3.2.1 Notation

We denote a set $\mathcal{E}$ of samples c_i, $c_j \in \mathcal{E}$, $i, j \in \{1, 2, \ldots, N\}$, where N is the number of samples in set $\mathcal{E}$. Each sample c_i is represented by a $1 \times K$ feature vector $\mathbf{x}_i \in X$. We can formalize a dataset X as a $N \times K$ matrix containing K feature values for each of N samples. The goal of dimensionality reduction is to reduce the $N \times K$ matrix, defined by a $1 \times K$ feature vector $\mathbf{x}_i \in X$, where $k < K$, to a $N \times k$ matrix, where all samples c_i are defined by a $1 \times k$ eigen-feature vector $\mathbf{y}_i \in Y$. Label information may be introduced such that $\ell(c_i)$ denotes the object class label of sample c_i as being a positive class $+1$ or negative class -1. Labels $\ell(c_i) = 0$ denotes that sample c_i is unlabeled.

3.2.2 Graph Embedding

NLDR methods, such as Graph Embedding [108], can be used to reduce samples c_i originally represented as K-dimensional vectors $\mathbf{x}_i \in X$ into k-dimensional vectors $\mathbf{y}_i \in Y$, where $k < K$. To perform this transformation, data X is first represented as an affinity matrix W, which describes the similarity between all pairs of objects $c_i, c_j \in S$ as a graph $G = \{V, E\}$, where V represents all objects c_i and c_j as vertices, and E represents the edges which connect them.

Similarity is computed via the Gaussian diffusion kernel $\gamma = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2}{\sigma}}$, which affects the weighting of the components in W. The kernel allows for a flexible local neighborhood constraint induced based on σ. A small σ narrows the size of the local neighborhood such that fewer points are deemed similar, whereas a large σ increases the size of the local neighborhood such that more points are similar. We set $\sigma = \max_{i,j} ||\mathbf{x}_i - \mathbf{x}_j||_2$.

Alternatively, E, the edges in the graph G, expressed via the affinity matrix, W, can be pruned to further constrain local neighborhoods for NLDR. E can be defined based on a local neighborhood size determined by the number of nearest neighbors κ. For each c_i, if c_j is one of the κ-nearest neighbors of c_i, then we may include c_j in the set $\mathcal{K}_i$ and we can express the edge as $E(c_i, c_j) = 1$. The weight matrix W represents a non-binary extension of the graph G, which takes into account the explicit similarity between objects c_i and c_j in terms of $\mathbf{x}_i$ and $\mathbf{x}_j$ such that

$$W(\mathbf{x}_i, \mathbf{x}_j) \quad = \quad \begin{cases} \gamma, & \text{if } c_j \in \mathcal{K}_i \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

As performed in the normalized cuts algorithm [108], the affinity matrix is normalized such that

$$\tilde{W}(\mathbf{x}_i, \mathbf{x}_j) = \Big(\sum_{ii}^{N} W(\mathbf{x}_{ii}, \mathbf{x}_j) \times \sum_{jj}^{N} W(\mathbf{x}_i, \mathbf{x}_{jj}) \Big)^{-1} W(\mathbf{x}_i, \mathbf{x}_j). \tag{3.2}$$

$\tilde{W}(\mathbf{x}_i, \mathbf{x}_j)$ is used to solve the eigenvalue problem

$$(D - \tilde{W})e = \lambda De, \tag{3.3}$$

where D is a diagonal matrix containing the trace of $\tilde{W}$, and e are the eigenvectors. The embedding Y^{GE} is formed by taking the most dominant eigenvectors e_β, $\beta \in \{1, 2, \ldots, k\}$, corresponding to the k smallest eigenvalues λ_β, where k corresponds to the dimensionality of Y^{GE}.

3.2.3 Semi-Supervised Agglomerative Graph Embedding

Adding semi-supervised learning to DR is performed by modifying the Graph Embedding algorithm to introduce the label information $\ell(c_i)$. A typical strategy for introducing label information into the Graph Embedding framework is to apply an additional set of weighting constraints to describe pairs of c_i and c_i with either the same ($\ell(c_i) = \ell(c_j)$) or different ($\ell(c_i) \neq \ell(c_j)$) labels. We utilize a methodology used by Zhao et al. [143], SSAGE, which includes a multiplier to the Gaussian diffusion kernel $\gamma = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2}{\sigma}}$ such that the affinity matrix is now defined as

$$\hat{W}(\mathbf{x}_i, \mathbf{x}_j) \quad = \quad \begin{cases} \gamma(1+\gamma), & \text{if } \ell(c_i)=\ell(c_j) \text{ and } c_j \in \mathcal{K}_i \\ \gamma(1-\gamma), & \text{if } \ell(c_i)\neq\ell(c_j) \text{ and } c_j \in \mathcal{K}_i \\ \gamma, & \text{if } \ell(c_j)=0 \text{ and } c_j \in \mathcal{K}_i \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

$\hat{W}$ contains the weighted similarities between c_i and c_j based on

1. its position in K-dimensional space via the Gaussian diffusion kernel
2. its proximity to its κ nearest neighbors
3. whether that neighbor is of the same label class or not

$\hat{W}$ is subsequently normalized via Equation 3.2 and the resulting normalized affinity matrix undergoes eigenvalue decomposition as performed in Equation 3.3. As with GE, the embedding Y^{SS} for SSAGE is formed by taking the most dominant eigenvectors e_β, $\beta \in \{1, 2, \ldots, k\}$, corresponding to the k smallest eigenvalues λ_β, where k is the dimensionality of Y^{SS}.

3.2.4 Active Learning by Uncertainty Sampling for Identifying Ambiguous Samples

One can identify samples for active learning by querying difficult to classify samples [55, 146,148,149,155]. While many strategies have been investigated for active learning using different classifiers, ultimately these differences were found not to be heavily correlated with classification performance [146]. For uncertainly sampling, a labeled set S_{tr} is first used to train a classifier. For each sample in the unlabeled set S_{ts}, the classifier predicts the object class label $\ell(c_i)$ with a certain probability that c_i belongs to that particular object class $\ell(c)$ (i.e. $P(\ell(c_i) = 1)$). We can define the most ambiguous samples as those with a probability of $P(\ell(c_i)) = 0.5$. We aim to find samples c_i nearest to $P(\ell(c_i)) = 0.5$ via the objective function

$$\underset{c_i \in S_{ts}}{\operatorname{argmin}} \Big| P(\ell(c_i) = 1) - 0.5 \Big|. \tag{3.5}$$

These samples c_i are assigned to set S_a. Labels $\ell(c_i)$, $c_i \in S_a$ are queried and these ambiguous samples are added to the training set

$$S_{tr} = [S_{tr} \cup S_a]. \tag{3.6}$$

Learning via the updated labels $\ell(c_i)$, $c_i \in S_{tr}$, we endeavor to improve classification performance compared to $S_{tr} \notin S_a$.

3.3 AdDReSS: Adaptive Dimensionality Reduction with Semi-Supervision

The iterative Algorithm *AdDReSS* is presented below. Additionally, we employ the synthetic Swiss Roll example presented in Figure 3.3 to guide the explanation of the AdDReSS algorithm.

Line 0 of the algorithm refers to *Model Initialization*, the construction of the initial embedding Y^{Ad}, and is illustrated in Figure 3.3(c) which shows the application of AdDReSS on the Swiss Roll dataset. The initialized embedding Y^{Ad} is created using data X via GE. In Figure 3.3(d), the revealed labels used for active learning are mapped onto Y^{Ad}.

Algorithm *AdDReSS*

Input: X, $\ell(S_{tr})$

Output: Y^{Ad}

begin

0. Build initial embedding Y^{Ad} using X, $\ell(c) = \{\ \}$
   via Equation 3.3
1. **while** $S_{ts} \neq \{\ \}$
2. Train classifier using Y^{Ad}, $\ell(S_{tr})$
3. Predict $\ell(c_i)$ in S_{ts} using classifier model
   in Step 2
4. Identify ambiguous samples from $c_i \in S_{ts}$
   via Equation 3.5
5. Query labels $\ell(c_i)$, $c_i \in S_a$
6. Update S_{tr} via Equation 3.6
7. Update embedding Y^{Ad} using updated $\ell(S_{tr})$
   via Equation 3.4
8. **end**
9. *return* Y^{Ad}

end

The subsequent illustrations, Figures 3.3(e) and (g), represent successive runs of *Active Learning* and *Model Refinement* via SSDR, respectively, which are contained within the while loop of the algorithm (lines 2-7).

Lines 2-6 of the algorithm represent the *Active Learning* component described earlier in Section 3.2.4, where ambiguous samples are identified based on the results of a trained classifier. Although Doyle et al. [146] have suggested that the particular choice of active learner is not significantly correlated with classifier performance, we have chosen the Support Vector Machine (SVM) classifier to identify the ambiguous samples for the following reasons. Firstly, SVMs have been shown to be highly generalizable to new

unseen testing data, suggesting that the algorithm can consistently identify ambiguous samples [151, 155]. Secondly, SVMs have been heavily investigated and employed for active learning [156, 157]. Finally, SVMs, like GE, operate on a kernel representation of the data, allowing for seamless identification of ambiguous samples derived from the kernel space in construction of the embeddings. A linear kernel was used based on the assumption that the NLDR method GE provides a linearly separable embedding as GE is able to account for non-linear data. We have previously shown the ability of linear kernel SVM to separate biomedical data using low dimensional representations from NLDR methods [1].

Figure 3.3(e) shows a visualization of the ambiguous samples found via SVM classification of Figure 3.3(d). Difficult to classify samples (shown as blue points) are found at the intersection of the two labeled classes (Figure 3.3(f)). New labels are obtained for these samples and added to the training set, completing the active learning phase (lines 2-6).

Line 7 of the algorithm represents the *Model Refinement* component where the updated label set $\ell(S_{tr})$ found via active learning is used to create an improved embedding representation via SSDR (Figure 3.3(g)). This representation demonstrates an improvement upon the previous embedding (Figure 3.3(c)). These steps of identifying samples (Figure 3.3(e)) and generating an optimized representation (Figure 3.3(g)) may be repeated until there are no additional unlabeled samples available for querying or until there is a lack of ambiguous samples to be queried.

3.4 Experimental Design

3.4.1 Embedding Parameters

The goal of these experiments is to understand the performance of AdDReSS with respect to constructing discriminative embeddings. Embeddings Y^{Ad} and Y^{SS} for AdDReSS and SSAGE, respectively, (refer to Sections 3.3 and 3.2.3 for more details) were generated with 20 different randomly selected training sets S_{tr} of training samples.

Measures designed to evaluate each embedding were calculated across multiple iterations of AdDReSS $Y_{l\%}^{Ad}$ corresponding to an embedding for a percentage l of revealed labels $\ell(c_i)$. These trials were repeated across a range of parameters for each dataset $\mathcal{D}_1$-$\mathcal{D}_3$ (as described in Section 3.4). Embeddings Y^{GE} were also generated for unsupervised GE (refer to Section 3.2.2 for more details) for comparison, but since no label information is used, only one embedding is obtained across all label iterations for each parameter set. Optimal κ parameters $\kappa \in \{2, \ldots, n-1\}$ were selected for all experiments.

3.4.2 Training Parameters

Each dataset is divided equally into training and testing pools, $\mathcal{E}_{tr}$ and $\mathcal{E}_{ts}$, respectively, for the purpose of an unbiased evaluation of the resulting Y. Random stratified sampling was performed such that samples for each of $\mathcal{E}_{tr}$ and $\mathcal{E}_{ts}$ are randomly chosen such that the number of positive and negative class labels $\ell(c)$ is the same in both $\mathcal{E}_{tr}$ and $\mathcal{E}_{ts}$. Note that $\mathcal{E}_{tr}$ and $\mathcal{E}_{ts}$ are distinct from the training and testing sets S_{tr} and S_{ts} used for querying samples for active learning. S_{tr} and S_{ts} are solely used for construction of the embedding and make up the entirety of the training pool $\mathcal{E}_{tr}$, described in this section such that $\mathcal{E}_{tr} = [S_{tr} \cup S_{ts}]$. Meanwhile, the labels $\ell(\mathcal{E}_{ts})$ in the testing pool are used only for analysis and are not used for constructing Y.

3.4.3 Evaluation Measures

In this study, two primary methods of quantitative evaluation are used to compare Y: Classifier Accuracy (Acc) and Silhouette Index (SI). These measures and related measures are defined in the following sections.

Evaluation of Classification Accuracy (ϕ^{Acc})

Classifier accuracy (Acc) is calculated to evaluate class separability within the embedding. The labeled set $\mathcal{E}_{tr}$ is used to train a classifier to predict the object class label $\ell(c_i)$ for all $c_i \in \mathcal{E}_{ts}$. If the true label value of $\ell(c_i)$ is of the positive class $\ell(c_i) = 1$ and the classifier predicts correctly, this result is a true positive (TP) classification. If the

classifier predicts this result incorrectly, the result is a true negative (FP). Similarly, if $\ell(c_i) = -1$ and the classifier predicts correctly, the result is a true negative (TN). Otherwise, the result is a false negative (FN). Classification Accuracy (ϕ^{Acc}) measures the ability of a classifier to predict on a new set of testing data provided by the embedding and is calculated as

$$\phi^{Acc} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.7}$$

Specifically, a Random Forest classifier (or bagged decision tree classifiers) [158] has been used due to its robustness and to reduce bias by selecting a different classifier than the one used for query ambiguous samples (in our case, an SVM classifier). The Random Forest classifier is constructed using 50 decision tree classifiers each trained on a random third of the training pool $\mathcal{E}_{tr}$. Classification accuracy ϕ^{Acc} is subsequently calculated based on the consensus of predicted labels $\ell(c_i)$ of the Random Forest classifier on the independent testing pool $c_i \in \mathcal{E}_{ts}$.

Evaluation of Object Class Separation via Silhouette Index (ϕ^{SI})

Silhouette Index (SI) offers an independent measure to quantify the separation of multiple classes in the embedding. SI can detect more subtle changes in the embedding with regards to overall class separation compared to classification accuracy. The Silhouette Index (ϕ^{SI}) [159] is a cluster validity measure which jointly takes into account (1) the compactness of samples belonging to the same object class ($\ell(c_i) = \ell(c_j)$) and (2) the separation of samples belonging to different object classes ($\ell(c_i) \neq \ell(c_j)$). The intra-cluster compactness is measured by $A_i = \sum_{j,\ell(c_j)=\ell(c_i)} \|\mathbf{y}_i - \mathbf{y}_j\|_2$, which represents the average distance of a sample c_i from other samples c_j of the same class in Y. Whereas, inter-cluster separation is measured by $B_i = \sum_{j,\ell(c_j)\neq\ell(c_i)} \|\mathbf{y}_i - \mathbf{y}_j\|_2$, the minimum of the average distances of a sample c_i from other samples in different classes. Thus, the formulation for ϕ^{SI} is as follows,

$$\phi^{SI} = \sum_{i}^{N} \frac{B_i - A_i}{\max[A_i, B_i]}. \tag{3.8}$$

ϕ^{SI} ranges from -1 to 1, where -1 demonstrates the worst, and 1 is the best possible embedding. For each experiment, ϕ^{SI} is calculated using all labels $\ell(c_i)$, $c_i \in \mathcal{E}_{tr}$ in Y.

Evaluation of Embedding Variance via Classification Accuracy (ρ^{Acc})

The rate of learning is affected by the initial training examples S_{tr} provided to the algorithm. It is anticipated that active learning will be able to consistently identify training instances, S_a, which will lead to improved classification, whereas random sampling will show more varied improvement due to the variance in the specific training instances chosen. We test the variance in ϕ^{Acc} of our algorithm (AdDReSS) compared to SSAGE across all runs, each with a unique random initializations S_{tr}. Classification Variance is computed as

$$\rho^{Acc} = \frac{\sum_i^n (\phi_i^{Acc} - \bar{\phi}^{Acc})^2}{n-1}, \tag{3.9}$$

where $n = 20$, representing the number of random initializations, and $\bar{\phi}^{Acc}$ refers to the mean across n values of ϕ_i^{Acc}, $\bar{\phi}^{Acc} = \frac{1}{n}\sum_i^n \phi_i^{Acc}$. A lower ρ^{Acc} suggests greater robustness to initialization via a more consistent ϕ^{Acc}.

Evaluation of Embedding Variance via Silhouette Index (ρ^{SI})

Similar to ρ^{Acc}, we also aim to quantify the variance of the embedding with regards to the Silhouette Index, which reflects the separability of the two object classes in terms of the Euclidean distance between data points in the embedding Y. ρ^{SI} captures the variance in the embedding Y across all runs, each with unique, random initializations, such that Silhouette Variance is computed as

$$\rho^{SI} = \frac{\sum_i^n (\phi_i^{SI} - \bar{\phi}^{SI})^2}{n-1}, \tag{3.10}$$

where $N = 20$, the number of random initializations, and $\bar{\phi}^{SI}$ refers to the mean across n values of ϕ_i^{SI}, $\bar{\phi}^{SI} = \frac{1}{n}\sum_i^n \phi_i^{SI}$. A lower ρ^{SI} suggests greater robustness to initialization in terms of a more consistent ϕ^{SI}.

Evaluation of Overall Embedding Learning Rate via Raghavan Efficiency (ϕ^{Eff})

Raghavan Efficiency [160] describes the rate of learning among active learning algorithms. Figure 3.8 [155] provides a visual interpretation of Raghavan Efficiency, where the region identified by A represents the area between the the Active Learning curve and the maximum achievable performance, and the region defined by B represents the area between the the Active Learning curve and the Random Sampling curve. Raghavan Efficiency is defined by a subtraction of the ratio A/B such that ϕ^{Eff} ranges between 0 and 1 and larger values of ϕ^{Eff} are indicative of a faster learning rate. We use ϕ^{Eff} to compare the overall learning rate between 1) AdDReSS vs GE, 2) SSAGE vs GE and 3) AdDReSS vs SSAGE.

To compare the efficiency of an active learner Y^{Ac} against random sampling Y^{Rd}, ϕ^{Eff} may be expressed as

$$\begin{aligned} \phi^{Eff}(Y^{Ac}|Y^{Rd}) &= 1 - \frac{A}{A+B} \\ &= 1 - \frac{\sum_{t=t_0}^{t_f} \phi^{Acc}(Y^{Rd}_{l=t_f}) - \phi^{Acc}(Y^{Ac}_{l=t})}{\sum_{t=t_0}^{t_f} \phi^{Acc}(Y^{Rd}_{l=t_f}) - \phi^{Acc}(Y^{Rd}_{l=t})}, \end{aligned} \tag{3.11}$$

where t_0 and t_f represent the number of initial training samples used to learn Y, and the final number of training samples used to learn Y, respectively. The empirical maximum accuracy refers to the highest ϕ^{Acc} obtained for any single iteration of Y such that $\phi^{EM} = \max_{i,l}\left[\phi_i^{Acc}(Y_l^{Ac})\right]$, where $i \in \{1, 2, \ldots, n\}$ denotes specific run of Y^{Ac} with a unique initial training set S_{ts}.

Additionally, to compare AdDReSS and SSAGE against the same baseline comparison, GE, we summarized these results using percentage comparison between for 1) $\phi^{Eff}(Y^{Ad}|Y^{GE})$ and 2) $\phi^{Eff}(Y^{SS}|Y^{GE})$. The percentage change in ϕ^{Eff} for AdDReSS from SSAGE can be expressed as

$$\Delta\phi^{Eff} = \left(1 - \frac{\phi^{Eff}(Y^{Ad}|Y^{GE})}{\phi^{Eff}(Y^{SS}|Y^{GE})}\right) \times 100\%. \tag{3.12}$$

Evaluation of Maximum Query Efficiency (ϕ^{MQE})

While Raghavan Efficiency is useful as an overall measure, there remain important insights that cannot be surmised by the global measure. One example is the cost savings associated with using active learning based dimensionality reduction compared to with traditional SSDR using random sampling. Maximum Query Efficiency is the ratio between the maximum difference in the number of labels necessary to achieve the same classification performance and the number of potential queries such that

$$\phi^{MQE} = \max_{\phi^{Acc}} \Big[\frac{l^{SS} - l^{Ad}}{N} \Big], \tag{3.13}$$

where l^{SS} and l^{Ad} refer to the mean number of labels queried by SSAGE and AdDReSS, respectively, to achieve a classification performance ϕ^{Acc}. N refers to the number of total samples $c_i \in \mathcal{E}$. A larger ϕ^{MQE} is indicative of greater savings in terms of labels queried.

Evaluation of Maximum Information Gain (ϕ^{MIG})

Another useful measure of active learning performance is the maximum information gain from using a particular algorithm of choice. We define maximum information gain as the maximum difference in classification performance ϕ^{Acc} at a given label query amount l, such that

$$\phi^{MIG} = \max_{l} \Big[\phi^{Acc}(\bar{Y}_l^{Ad}) - \phi^{Acc}(\bar{Y}_l^{SS}) \Big]. \tag{3.14}$$

A larger ϕ^{MIG} refers to a larger difference between the classification performance between embeddings constructed by AdDReSS and embeddings generated by SSAGE.

3.4.4 Dataset Description

A total of 5 datasets were used in this study. Two synthetic datasets ($\mathcal{S}_1$ - $\mathcal{S}_2$) were utilized as examples, including 1 imaging and 1 non-imaging. Three additional datasets ($\mathcal{D}_1$ - $\mathcal{D}_3$) were selected for experimentation. These datasets include: $\mathcal{D}_1$: synthetic

Table 3.1: Datasets used for evaluation.

SyntheticDatasets	*Description*	*Features*
$\mathcal{S}_1$: Toy Data 30×50 image	1500 pixels 739 Foreground, 761 Background	RGB intensity (3)
$\mathcal{S}_2$: Swiss Roll	429 Red, 571 Black samples	XYZ coordinates (3)
BiomedicalDatasets	*Description*	*Features*
$\mathcal{D}_1$: BrainWeb 109×131 image	5,975 total Grey Matter and White Matter pixels 2607 Grey Matter, 3368 White Matter	Texture (6)
$\mathcal{D}_2$: Prostate Cancer	52 Tumor, 50 Normal	Gene Expression (12600)
$\mathcal{D}_3$: Ovarian Cancer	162 Tumor, 91 Normal	Protein Expression (15154)

brain image data, $\mathcal{D}_2$: gene-expression of prostate cancer, and $\mathcal{D}_3$: protein expression of ovarian cancer. The datasets are summarized in Table 3.1.

$\mathcal{S}_1$: Toy Data

The synthetic toy dataset ($\mathcal{S}_1$) is a 30×50 RGB color image containing 739 and 761 pixels corresponding to the foreground circle and background respectively as shown in Figure 3.2. The objective is to separate foreground and background pixels on a noisy image. Each pixel is defined by three color intensity channels R,G, and B. The original scene is defined by two types of pixels (foreground and background) which are linearly separable in RGB space. A Gaussian noise function was added to each pixel c_i and to each color channel α such that

$$f_\alpha(c_i) = f_\alpha(c_i) + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{f_\alpha(c_i)}{\sigma})^2}, \tag{3.15}$$

where $\sigma = 150$. The image X is then subsequently normalized such that all color intensity values $f_\alpha(c)$ range between 0 and 255.

Training and testing pools were created based on selecting alternative pixels from a checkerboard pattern. This allowed for a well-represented training and testing pool for evaluation. For initialization of $N = 20$ embeddings Y, we generated multiple embeddings, each with a unique set of 50 pixels $c_i \in S_{tr}$. For AdDReSS, 50 additional pixels $c_i \in S_a$ were added to S_{tr} at each iteration. For SSAGE, 50 additional randomly selected pixels $c_i \in S_{ts}$ were added to S_{tr} at each iteration. In all cases, the $K = 3$ dimensional RGB intensity space is reduced to dimensionality $k = 2$.

$\mathcal{S}_2$: Swiss Roll

The synthetic Swiss Roll dataset [10] is a 1000 sample dataset defined by 3 coordinates commonly used to test NLDR algorithms as shown in Figure 3.3(a). The points have been separated into two classes along the manifold such that a hyperplane in 3D is unable to separate the classes. The goal is to be able to separate these classes within a 2D embedded space. For all results, the $K = 3$ dimensional RGB intensity space is reduced to dimensionality $k = 2$ as shown in Figure 3.3.

$\mathcal{D}_1$: BrainWeb Images

Preprocessing: Synthetic brain images [154] were acquired from the Montreal Neurological Institute[1]. This dataset consists of proton density MRI brain volumes with simulated levels of noise and bias field inhomogeneities. Gaussian noise artifacts were simulated by adding to each pixel in the image, with parameters for Gaussian noise artifacts (NO) ranging between 1% to 9% noise. Inhomogeneity artifacts were simulated by multiplication of each pixel in the image with an intensity non-uniformity field. Intensity non-uniformity (RF) was simulated at from 0, 20 and 40%. Images were acquired at a slice thickness of 1mm. White matter and grey matter regions were labeled for each of the images in the dataset. A single slice is used in this study comprising white and grey matter alone (ignoring other brain tissue classes).

Feature Extraction: 6 texture features [161] were extracted from each image on a per-pixel basis: contrast energy, contrast entropy, intensity variance, correlation, and two features corresponding to information measures. These texture features represent second-order statistics calculated from a gray level intensity co-occurrence matrix constructed from the gray level image intensity values. These features were previously used to discriminate cancerous from non-cancerous prostate regions [111] and different types of brain matter [128, 162] in MRI studies. For all results, the $K = 6$ dimensional RGB intensity space is reduced to dimensionality $k \in \{2, 3\}$.

$\mathcal{D}_2$: Gene Expression of Prostate Cancer

Preprocessing: Gene expression data [76] was acquired from the Biomedical Kent-Ridge Repositories[2], consisting of high quality expression profiles from 52 prostate tumors and 50 non-tumor (normal) prostate samples. The samples are derived from oligonuleotide microarrays containing probes for 12,600 genes.

Feature Extraction: No additional feature extraction was performed and all embeddings were calculated directly from the provided data. For all results, the $K = 12,600$ dimensional dataset was reduced down to dimensionality $k \in \{2, 3\}$.

$\mathcal{D}_3$: Protein Expression of Ovarian Cancer

Preprocessing: The study [78], obtained from the Biomedical Kent-Ridge Repositories[3] uses proteomic spectra extracted from serum to distinguish 91 neoplastic from 162 non-neoplastic disease within the ovary. The proteomic spectra generated by SELDI mass spectroscopy for each sample contains the relative amplitude of 15,154 intensities at each molecular mass / charge (M/Z) identity.

Feature Extraction: No additional feature extraction was performed and all embeddings were calculated directly from the provided data. For all results, the $K = 15,154$ dimensional protein spectra was reduced down to dimensionality $k \in \{2, 3\}$.

3.5 Results and Discussion

3.5.1 Synthetic Example $\mathcal{S}_1$: Toy Data

For $\mathcal{S}_1$ we illustrate the separability of our target classes achievable by AdDReSS and two comparative DR methods, GE, which is unsupervised and SSAGE, which is supervised. In Figure 3.2(a), a simple RGB image consisting of ball and background pixels is shown. Following the addition of Gaussian noise, each pixel in Figure 3.2(a) is plotted in a 3-dimensional RGB space (Figure 3.2(e)). Subsequently, we reduce the 3-dimensional RGB space into a 2-dimensional embedding via GE (Figure 3.2(f)), SSAGE (Figure 3.2(g)), and AdDReSS (Figure 3.2(h)). Figures 3.2(b), 3.2(c), and 3.2(d) represent a pixel-wise binary classification into foreground (ball) and background classes via

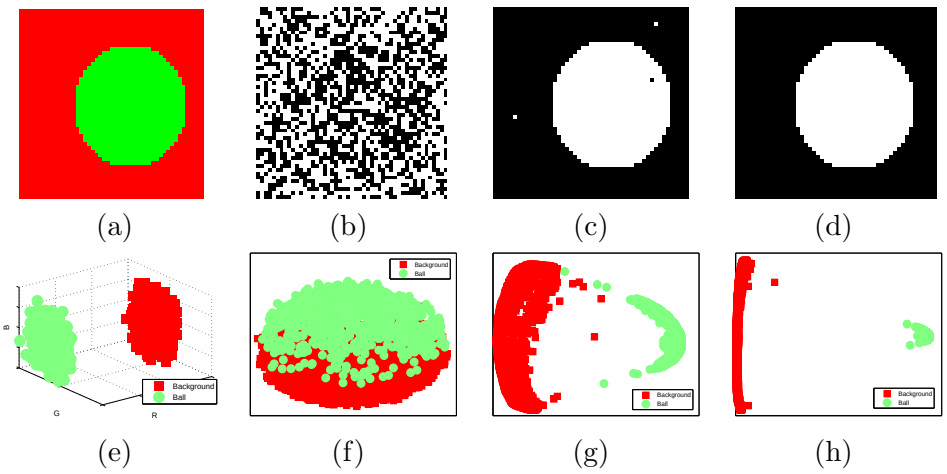


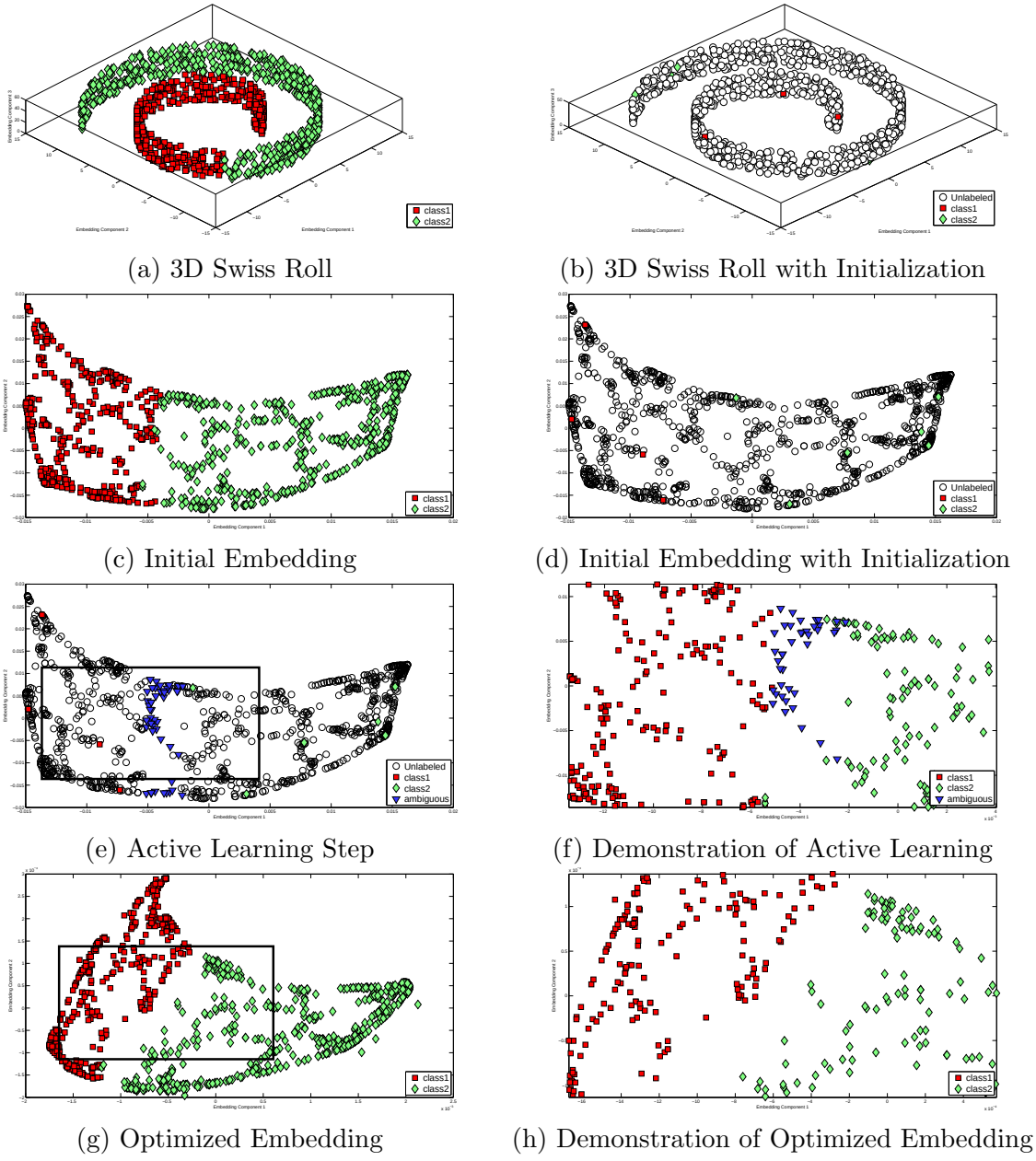
Figure 3.2: (a) RGB image containing ball against colored background pixels. (e) Image pixels plotted in 3D RGB space. Replicated k-means clustering is performed on the reduced embeddings by (f) GE, (g) SSAGE, and (h) AdDReSS, respectively. The resulting binary classifications (b-d) reflect the corresponding quality of embeddings obtained via DR methods (b) GE, (c) SSAGE, and (d) AdDReSS.

GE, SSAGE, and AdDReSS, respectively. These were obtained via replicated k-means clustering on the corresponding DR embeddings, as shown in Figures 3.2(f), 3.2(g), and 3.2(h). We can see that there are differences between the embeddings created via AdDReSS compared to SSAGE and GE, where SSL appears to provide an improvement in separating the foreground and background pixels for SSAGE over unsupervised GE, and the incorporation of active learning appears to provide an embedding with greater separability compared to SSAGE.

3.5.2 Synthetic Example $\mathcal{S}_2$: Swiss Roll

Figure 3.3(a) shows the 3-dimensional representation of the Swiss Roll dataset [10] shown with the two classes. The goal is to separate these two classes in a lower dimensional embedding representation such that each class is in a distinct region of the low dimensional embedding space. Figure 3.3 illustrates how the use of active learning is able to improve upon the separability of the two classes for this dataset.

Difficult to classify examples are identified by the SVM classifier in embedding space and are shown in blue in Figure 3.3(e). The newly identified objects discovered via AL attract towards similarly labeled samples already available to SSAGE and the



(a) 3D Swiss Roll

(b) 3D Swiss Roll with Initialization

(c) Initial Embedding

(d) Initial Embedding with Initialization

(e) Active Learning Step

(f) Demonstration of Active Learning

(g) Optimized Embedding

(h) Demonstration of Optimized Embedding

Figure 3.3: (a) 3D Swiss Roll with all labels revealed. (b) 3D Swiss Roll with initial labels $\ell(S_{tr})$ revealed. (c) Initial 2D embedding with labels. (d) Initial 2D embedding with initial labels $\ell(S_{tr})$. (e) Ambiguous samples (in blue) are determined via active learning. (f) Region of the Swiss Roll at the class boundary (region is shown as a box in (e)). Note the selection of ambiguous samples (in blue) at the boundary between the two classes (in red and green). (g) Subsequent 2D embedding incorporating newly queried labels from the ambiguous samples. (h) Region near the class boundaries (shown as a box from (g)) revealing the increased separation between the two classes (in red and green) following application of the AdDReSS scheme.

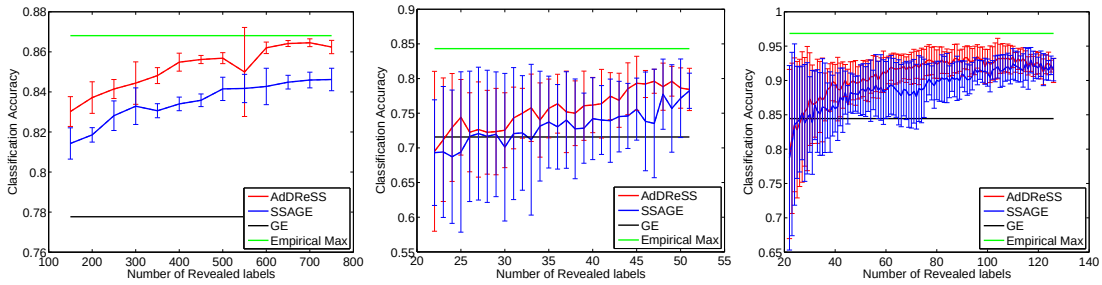
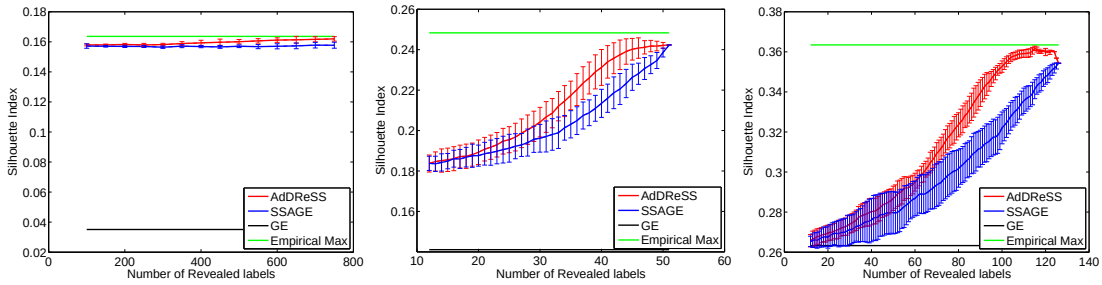


Figure 3.4: Number of instances for which labels were revealed versus mean ϕ^{Acc} for AdDReSS, SSAGE, GE, and the maximum empirically derived ϕ^{Acc} across all runs is shown for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$. Standard deviation of ϕ^{Acc} shown as error bounds at each l.

classifier while repelling from dissimilarly labeled samples, thus creating the separation shown in Figure 3.3(g). Thus, it is clear that the discovery of difficult to classify labels can produce greater separation of the embedding as these samples are leveraged by SSAGE. The use of random sampling would probabilistically provide a uniform sampling of points in the dataset such that SSAGE could not leverage the samples at the classification boundary, resulting in a smaller degree of separation of object classes.

3.5.3 Evaluation via Classifier Accuracy (ϕ^{Acc})

Figure 3.4 shows the classification performance of AdDReSS against SSAGE and GE on three biomedical datasets ($\mathcal{D}_1$ - $\mathcal{D}_3$), where different amounts of labeled data l are revealed to the classifier. We notice greater ϕ^{Acc} for AdDReSS across all amounts of revealed labels l. The accuracy curve corresponding to AdDReSS also approaches the empirical maximum ϕ^{Acc} at a faster rate compared to SSAGE. GE is also shown for each case as a comparison. The use of sufficient labeled instances suggests a clear advantage in employing semi-supervision for DR. Furthermore, the improved performance of AdDReSS over SSAGE across all labeled instances reveals a measurable difference in ϕ^{Acc} at a point between the minimum $l = 10\%$ and the maximum number of revealed labels $l = 50\%$. This is due to the fact that for small training size, $l = 10\%$, there is a significant overlap in S_{tr} for AdDReSS and SSAGE due to the identical initialization S_{tr}. Similarly at $l = 50\%$, training samples are exhausted from the pool $\mathcal{E}_{tr}$, such that



(a) (b) (c)

Figure 3.5: Number of instances for which labels were revealed versus mean ϕ^{SI} for AdDReSS, SSAGE, GE, and the maximum empirically derived ϕ^{SI} across all runs is shown for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$. Standard deviation in ϕ^{SI} shown as error bounds at each l.

$S_{tr} = \mathcal{E}_{tr}$ for both AdDReSS and SSAGE. Therefore, the greatest measurable difference between $\phi^{Acc}(Y_l^{Ad})$ and $\phi^{Acc}(Y_l^{SS})$ can be seen where $10\% < l < 50\%$, reflecting the difference in the active learning and random sampling strategies towards the composition of $c_i \in S_{tr}$, and subsequently, towards the resulting embeddings Y_l^{Ad} and Y_l^{SS}.

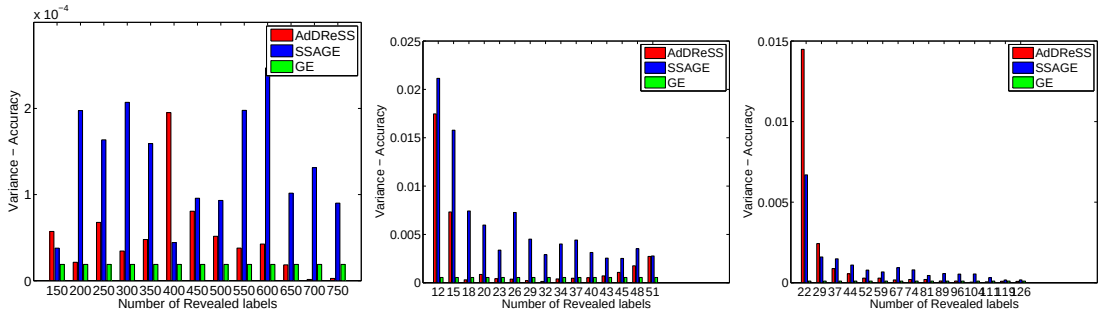
3.5.4 Evaluation via Silhouette Index (ϕ^{SI})

In Figure 3.5, we compared AdDReSS against SSAGE and GE in terms of ϕ^{SI} on datasets ($\mathcal{D}_1$ - $\mathcal{D}_3$) by revealing different amounts of labeled data l. Compared to ϕ^{Acc}, there appears to be greater separation for ϕ^{SI} between the semi-supervised methods compared to GE. This in turn seems to suggest that the separation of the object classes in the embedding space is more pronounced. Furthermore, $\phi^{SI}(Y^{Ad})$ outperforms $\phi^{SI}(Y^{SS})$ across all l. In contrast to ϕ^{Acc}, the improvement in ϕ^{SI} tends to continue with increasing numbers of revealed labeled information l. Only when the revealed labeled information is nearly $l = 50\%$ does ϕ^{SI} approach its empirical maximum ϕ^{SI}.

3.5.5 Evaluation of Variance (ρ^{Acc}, ρ^{SI})

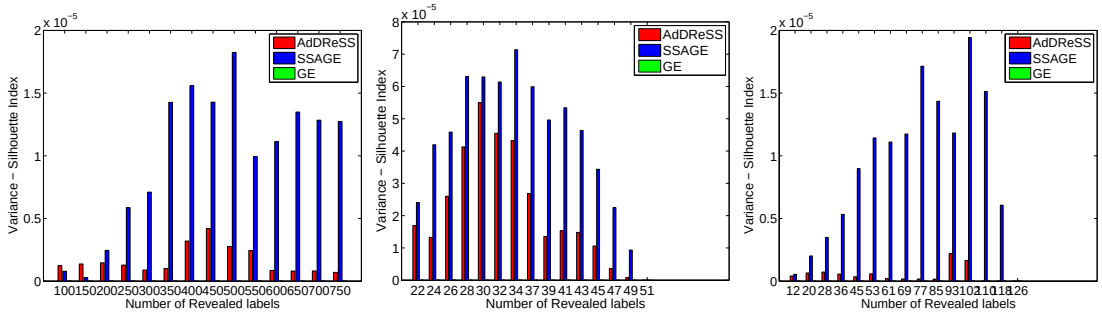
In Figure 3.6, we compare variance ϕ^{Acc} across varied amounts of revealed labels l for Y^{Ad}, Y^{SS} and Y^{GE}. In $\mathcal{D}_1$, we notice very small differences in ϕ^{Acc}, as ρ^{Acc} is found to be on average less than 0.0003 for all values of l. Nevertheless, we can view significant

differences between ρ^{Acc} of AdDReSS and SSAGE, with AdDReSS showing $\rho^{Acc} < 0.0001$ in all but one instance, and most instances of SSAGE showing $\rho^{Acc} > 0.0001$. We notice greater differences in ρ^{Acc} for $\mathcal{D}_2$ and $\mathcal{D}_3$ in Figures 3.6(b) and (c) respectively, as both AdDReSS and SSAGE are more sensitive to the composition of initial training $c_i \in S_{tr}$, reflected in the higher ρ^{Acc} when $l < 10\%$. ρ^{Acc} is subsequently seen to decrease with increasing l as more training samples are queried by the active learner. For all experiments in $\mathcal{D}_2$, AdDReSS shows more consistency in ϕ^{Acc} as demonstrated by lower ρ^{Acc} compared to SSAGE. Furthermore, AdDReSS shows similar ρ^{Acc} values when compared to the unsupervised GE method, which is reflective of the precision of the classifier. The same trends can be seen in $\mathcal{D}_3$ for $l > 28\%$ (Figure 3.6(c)), where over 29 revealed labeled instances were used and AdDReSS shows lower ρ^{Acc} compared to SSAGE.



(a) (b) (c)

Figure 3.6: Variance of ϕ^{Acc} at selected numbers of instances for which labels were revealed for AdDReSS, SSAGE, GE are shown for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$.



(a) (b) (c)

Figure 3.7: Variance of ϕ^{SI} at selected numbers of instances for which labels were revealed for AdDReSS, SSAGE, GE are shown for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$. GE shows zero variance as labeled information does not affect the embedding for GE.

In Figure 3.7, we demonstrate more consistent embeddings Y^{Ad} compared to Y^{SS} as demonstrated by a lower ρ^{SI}. However, unlike with ρ^{Acc}, ρ^{SI} tends to increase with increasing l. In all three datasets $\mathcal{D}_1$-$\mathcal{D}_3$, we notice SSAGE to have greater ρ^{SI} than AdDReSS and up to 3 or 4 times greater for $\mathcal{D}_1$ and $\mathcal{D}_3$. These trends in Figures 3.5 and 3.7 are reflective of the ability of the embedding to converge more quickly with increasing l for AdDReSS compared to SSAGE. The embedding for GE does not change with respect to l, therefore, there is no change in ϕ^{SI}, and $\rho^{SI} = 0$ in any of the cases. These results are suggestive of a embedding representation Y^{Ad} which is more stable than Y^{SS}, and is robust to the specific $c_i \in S_{tr}$ used to initialize AdDReSS.

3.5.6 Evaluation via Raghavan Efficiency (ϕ^{Eff})

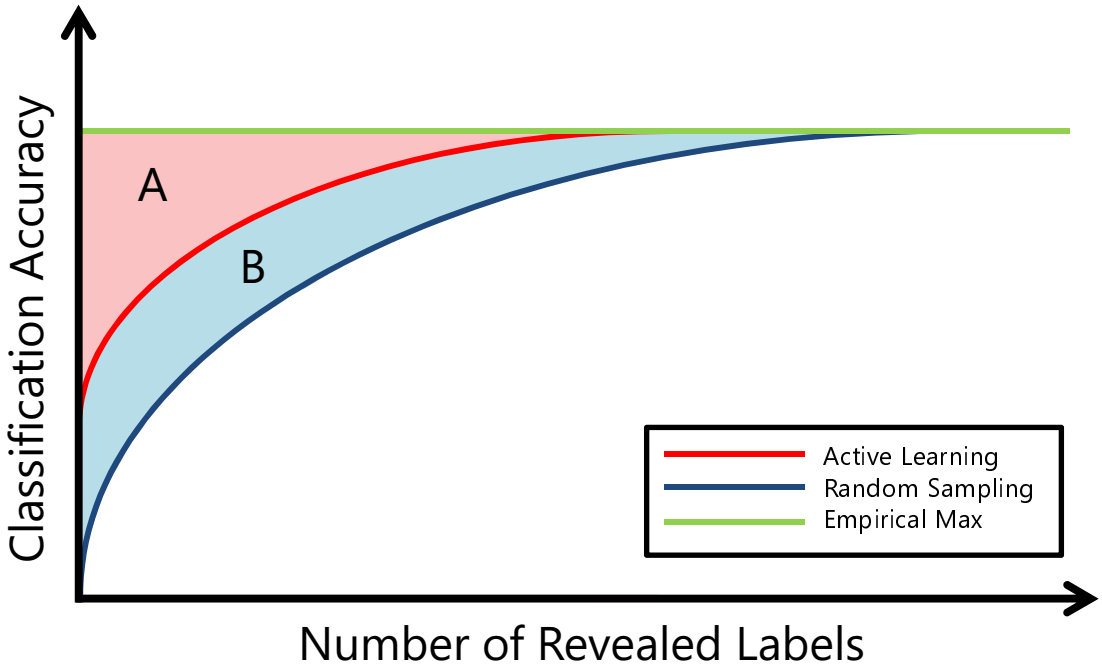
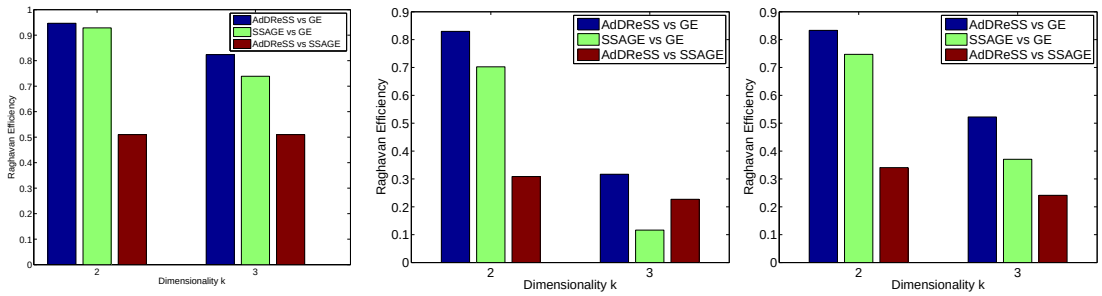


Figure 3.8: Illustration describing Raghavan efficiency. A refers to the area between the Active Learning curve and the empirically-derived maximum accuracy, and B refers to the area between the Random Sampling curve and the Active Learning curve.

In Figure 3.9, we show the overall differences in efficiency between each pair of methods (1) AdDReSS vs SSAGE, 2) AdDReSS vs GE, and 3) SSAGE vs GE) employed for this study via ϕ^{Eff}. In all cases, AdDReSS outperforms SSAGE in terms of ϕ^{Eff}. $\phi^{Eff}(Y^{Ad}|Y^{SS})$ is more pronounced when $k = 2$ for all datasets, suggesting



(a) (b) (c)

Figure 3.9: ϕ^{Eff} for $k \in \{2, 3\}$ shows the comparative efficiency between AdDReSS and GE, SSAGE and GE, and AdDReSS and SSAGE for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$.

Table 3.2: Percent improvement in Raghavan efficiency via AdDReSS over SSAGE

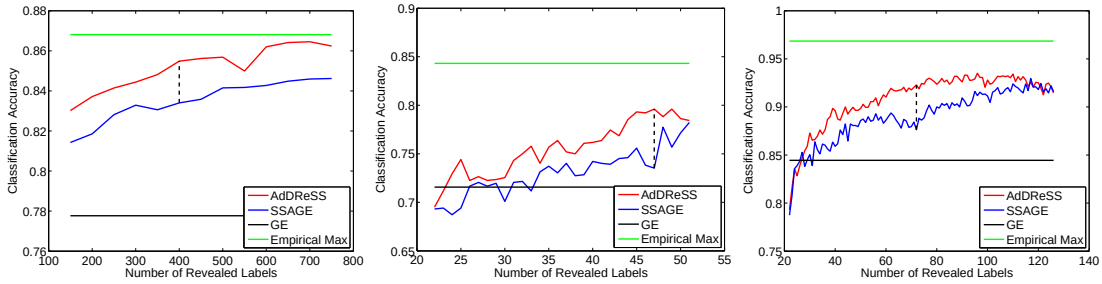
	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	Mean
$k = 2$	+1.94%	+18.09%	+11.53%	+10.52%
$k = 3$	+11.49%	+172.41%	+40.95%	+74.95%
Mean	+6.71%	+95.25%	+26.24%	+42.73%

that AdDReSS is more efficient when fewer dimensions are involved. These trends are consistent to what is seen in Figure 3.4, where AdDReSS shows greater ϕ^{Acc} for varying proportions of revealed labels l.

The improvement in efficiency afforded by AdDReSS compared to SSAGE is summarized in Table 3.2 using GE as the baseline. Table 3.2 shows the percentage increase between $\phi^{Eff}(Y^{Ad}|Y^{GE})$ and $\phi^{Eff}(Y^{SS}|Y^{GE})$ for all datasets $\mathcal{D}_1$-$\mathcal{D}_3$. Overall, the mean percentage improvement in ϕ^{Eff} across all datasets was found to be +10.52% for $k = 2$ and +74.95% for $k = 3$ from using AdDReSS instead of SSAGE, suggesting that AdDReSS appears to outperform SSAGE as the number of dimensions begins to increase.

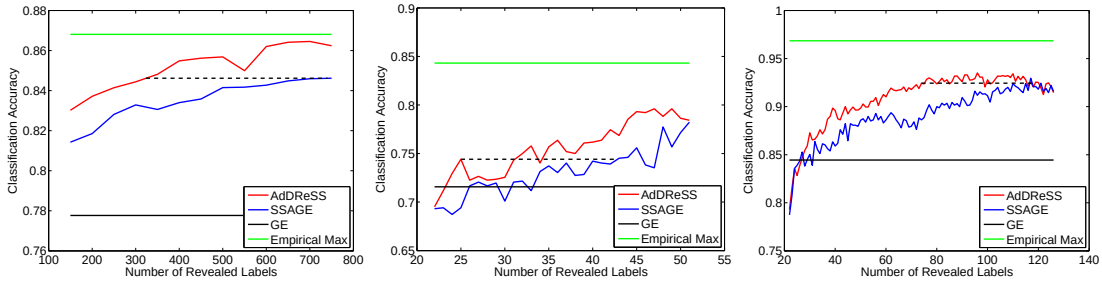
3.5.7 Evaluation via Maximum Information Gain (ϕ^{MIG})

In Figure 3.10, we show the maximum amount of information gain that can be achieved via AdDReSS compared to SSAGE for each dataset. For $\mathcal{D}_1$, $\phi^{MIG} = 0.0208$, which means there is a maximum improvement in ϕ^{Acc} of over 2% (from 0.8340 to 0.8548) due to AdDReSS compared to SSAGE. This improvement in ϕ^{Acc} via Y^{Ad} is equivalent to 60 additional correctly classified samples for $\mathcal{D}_1$ compared to Y^{SS}. In Figure 3.10(b), when



(a) (b) (c)

Figure 3.10: ϕ^{MIG} shows areas of maximum information gain (shown as a dashed black line) in terms of the difference in ϕ^{Acc} between AdDReSS and SSAGE for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$.



(a) (b) (c)

Figure 3.11: ϕ^{MQE} describes the maximum efficiency in terms of queried labels given the same ϕ^{Acc} (shown as a dashed black line) between AdDReSS and SSAGE for (a) $\mathcal{D}_1$, (b) $\mathcal{D}_2$, and (c) $\mathcal{D}_3$.

$l = 46\%$ (47 labels revealed), $\mathcal{D}_2$ shows $\phi^{MIG} = 0.0608$, with over an 8% improvement in ϕ^{Acc} when using AdDReSS compared to SSAGE. For $\mathcal{D}_3$, $\phi^{MIG} = 0.0465$, with an improvement from 0.8764 to 0.9228 in terms of ϕ^{Acc} and the best improvement is found when $l < 30\%$ (less than 72 labels revealed). The results for ϕ^{MIG} suggest a faster rate of learning when using AdDReSS compared to SSAGE.

3.5.8 Evaluation via Maximum Query Efficiency (ϕ^{MQE})

Figure 3.11 illustrates the number of fewer labels required for AdDReSS to achieve the same classification performance ϕ^{Acc} as SSAGE. For $\mathcal{D}_1$, $\phi^{MQE} = 0.0698$, which reflects the fact that AdDReSS requires an average of 417 fewer labels than SSAGE to achieve $\phi^{Acc} = 0.8462$. Stated another way, SSAGE required the use of an additional 6.98% of the labels $l(c_i), c_i \in \mathcal{E}$, to achieve the same performance as AdDReSS. For

$\mathcal{D}_2$, $\phi^{MQE} = 0.1748$. While an average of 25 revealed labeled instances were used to achieve $\phi^{Acc} = 0.74$ for AdDReSS, SSAGE required an average of 43 revealed labeled instances in order to achieve the same ϕ^{Acc}. Similarly, for $\mathcal{D}_3$, $\phi^{MQE} = 0.1730$, such that AdDReSS required, on average, 74 labels to achieve $\phi^{Acc} = 0.9244$ while SSAGE required nearly the entire training pool, $\mathcal{E}_{tr}$, of 126 labels, as shown in Figure 3.11(c). Overall, for $\mathcal{D}_1 - \mathcal{D}_3$, AdDReSS required an average of 45% (and up to 56%) fewer labels to be queried compared to SSAGE to achieve the desired classification accuracy.

3.6 Summary

With the rapid profusion of high dimensional biomedical data, there is a need for 'big data' analytics to efficiently and accurately analyze this data. This analysis very often involves developing classifiers for predicting disease aggressiveness, patient outcome, or appropriate treatment options. Recently, there has been a profusion of companion diagnostic tests in the context of personalized medicine where most of these tests employ a combination of molecular markers. To overcome the curse of dimensionality, there is a need for representing these high dimensional biomedical datasets in reduced dimensional spaces, which are more amenable to classification. In this work, we presented a novel nonlinear dimensionality reduction methodology, Adaptive Dimensionality Reduction with Semi-Supervision (AdDReSS), which attempts to seamlessly integrate active learning into semi-supervised dimensionality reduction (SSDR) to yield low dimensional data representations of high dimensional data. These representations yield greater classification accuracy and class separability while using fewer class labels. AdDReSS attempts to address the problems of classifying 'big data' and the very real problem of often not having class labels or annotations with which to train a classifier. Our scheme employs the use of active learning to query fewer labels which contribute the most towards building low dimensional embeddings with high object class separability and classification performance. We quantified the differences between AdDReSS and SSAGE on problems involving imaging and non-imaging channels from 3 distinct biomedical datasets (MR brain imaging, prostate gene expression, and ovarian proteomic spectra) and 2 synthetic examples (a toy image and swiss roll data). Based on

the results assessed over 8000 experiments, we make the following observations:

- AdDReSS has a greater predictive potential compared to SSAGE and GE based on classification accuracy when different numbers of instances have their labels revealed.
- AdDReSS achieved a higher Silhouette Index compared to SSAGE and GE, suggestive of an embedding with greater separation between the object classes.
- In comparisons of overall efficiency, AdDReSS learns at a faster rate of convergence to the maximum possible accuracy compared to SSAGE and GE, measured by a 42.73% increase in Ragahavan efficiency.
- The potential savings in terms of the number of labels to be queried to achieve the same classification accuracy was shown to be up to 56% for AdDReSS compared to SSAGE across the datasets considered.
- AdDReSS was also found to be more robust to randomized training set initialization, in that it appeared to have a lower variance in terms of classification accuracy and Silhouette Index compared to SSAGE in the datasets considered.

Out findings suggest that active learning has a measurable effect on SSAGE and that AdDReSS could be a powerful data analysis and classification tool for high dimensional biomedical data, especially in scenarios where partial or incomplete annotations and class labels are available.

Chapter 4

Supervised Multi-view Canonical Correlation Analysis for joint correlation and label guided data integration

4.1 Overview

4.2 Multi-modal data integration methods for imaging and non-imaging biomedical data

In this chapter, we introduce methods of data integration and present a methodology, supervised Multi-view Canonical Correlation Analysis (sMVCCA), which aims to integrate infinite views of high dimensional data to provide a more amenable data representation for classification of disease.

The following sections are organized as follows. First, we discuss useful background regarding the implementation of data integration methodologies which may be used or referenced in later parts of the dissertation. These methods include the use of Principal Component Analysis (PCA) as a method for data integration (Section 4.2.1), Generalized Embedding Concatenation (GEC) (Section 4.2.2), as well as methods for Canonical Correlation Analysis (CCA), Regularized CCA (RCCA) and Supervised RCCA (SRCCA) in Sections 4.2.3-4.2.5. Finally, we review a pairwise multi-view canonical correlation analysis (MVCCA) approach which we then extend to derive supervised MVCCA (sMVCCA) in Chapters 4.3 and 4.4, respectively. Table 4.1 provides relevant notation for this chapter.

4.2.1 Principal Component Analysis (PCA) for data integration

PCA [105] provides a low dimensional representation of the data by rotation of the coordinates to a set of orthogonal basis vectors called principal components. The principal

Table 4.1: Summary of Notation for Chapter 4

Symbol	Description
$\mathbf{x}_m$	data matrix of modality $m \in \{1, \ldots, M\}$
f	feature f^j in $\mathbf{x}_m = [f_m^1, f_m^j, \ldots, f_m^{s_m}]$
$\mathbf{X}$	data matrix of all modalities $[\mathbf{x}_1, \mathbf{x}_m, \mathbf{x}_M]$, $\mathbb{R}^{(n_1+n_2+\ldots+n_M)\times s_1+s_m+\ldots+n_M}$
n	number of samples
s_1 and s_2	number of features for modalities 1 and 2, and s is the total number of features across all m
g	number of object classes
$\mathbb{Y}_g$	signifies a particular object class label
$\mathbb{W}$	Object class label
γ_1, γ_2	regularization parameters
$\mathbf{I}_1$ and $\mathbf{I}_2$	Identity matrices of $\mathbb{R}^{s_1\times s_1}$ and $\mathbb{R}^{s_2\times s_2}$ respectively
$\mathbf{C}_{mt}$	Correlation Matrix between modality m and modality t, $m, t \in \{1, \ldots, M\}$
$\bar{\mathbf{C}}$	Summed Correlation Matrix
$\bar{\mathbf{C}}_d$	Diagonal Summed Correlation Matrix
$\hat{\mathbf{C}}$	Summed Correlation Matrix with label encoding
$\hat{\mathbf{C}}_d$	Diagonal Summed Correlation Matrix with label encoding
$\mathbf{Y}$	label matrix

components represent the axes which demonstrate the greatest amount of variance in the data. Given two multi-dimensional modalities of information $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathbf{x}_1 \in \mathbb{R}^{n\times s_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{n\times s_2}$, where n is the number of samples in the dataset, and s_1 and s_2 are the number of features in $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. PCA is performed on the concatenated data matrix, $\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2] \in \mathbb{R}^{n\times(s_1+s_2)}$ [163]. $\bar{\mathbf{X}} \in \mathbb{R}^{n\times(s_1+s_2)}$ is then obtained by subtracting the mean from each feature for a certain sample such that the resultant $\bar{\mathbf{X}}$ contains features of 0 mean. Eigenvalue decomposition of $\bar{\mathbf{X}}$ [105] is shown below as

$$\bar{\mathbf{X}}^T\bar{\mathbf{X}} = V\Sigma V^T \tag{4.1}$$

where $\Sigma \in \mathbb{R}^{(s_1+s_2)\times(s_1+s_2)}$ is diagonal matrix containing the singular values for the eigenvectors in $V \in \mathbb{R}^{(s_1+s_2)\times(s_1+s_2)}$ The singular values in the diagonal of Σ denote the variance of $\bar{\mathbf{X}}$ and determines the principal components based on the corresponding eigenvectors. The embedding components corresponding to the largest largest d singular values are selected for the low dimensional PCA representation of the original data.

4.2.2 Generalized Embedding Concatentation

Generalized Embedding Concatenation (GEC) [5] is an extension of the Generalized Fusion Framework which makes use of multiple embeddings for homogeneous meta-space representation of the features corresponding to each modality. The intermediate 'meta-space' is of equivalent scaling and dimensionality for each modality considered allowing for the representation of 2 or more modalities into a single integrated embedding. The 'meta-space' representation is achieved for each modality m, $\mathbf{x}_m$ via transformation by Principal Component Analysis to eigenvectors $E_m = [e_1, e_2, \ldots, e_d]$. To account for multiple scales E_m is transformed by

$$\hat{E}_m^i = \frac{e_j^i - \min_i[E_j^i]}{\max_i[E_j^i] - \min_i[E_j^i]}, i \in \{1, 2, ..., s_1 + s_2\} \tag{4.2}$$

for each E_j, $j \in \{1, 2, ..., n\}$ to yield normalized embedding vectors. 'Meta-space' fusion is then accomplished via concatenation of $\hat{E}_m$. such that $\mathcal{E} = [\hat{E}_1, \ldots, \hat{E}_m, \ldots, \hat{E}_M]$. The joint space is reduced by PCA on $\mathcal{E}$ to obtain the joint low-dimensional representation $\epsilon_1, \ldots, \epsilon_d$.

4.2.3 Canonical Correlation Analysis

Provided n data samples from two modalities, $m \in \{1, 2\}$, comprising s_1 and s_2 features, respectively: $\{\mathbf{x_1}, \mathbf{x_2}\}_i,\ i \in \{1, 2, ..., n\}, \mathbf{x}_1 = [f_1^1, f_1^j, \ldots, f_1^{s_1}]^T, \mathbf{x}_2 = [f_2^1, f_2^j, \ldots, f_2^{s_1}]^T$. Canonical Correlation Analysis (CCA) seeks a pair of transformations $\mathbf{w}_1$ and $\mathbf{w}_2$ such that correlation of $\mathbf{x}_1$ and $\mathbf{x}_2$ in the transformed space is maximized,

$$\underset{\mathbf{w}_1, \mathbf{w}_2}{\operatorname{argmax}} \frac{\mathbf{w}_1^T \mathbf{C}_{12} \mathbf{w}_2}{\sqrt{\mathbf{w}_1^T \mathbf{C}_{11} \mathbf{w}_1 \mathbf{w}_2^T \mathbf{C}_{22} \mathbf{w}_2}}, \tag{4.3}$$

where $\mathbf{C}_{12} = \sum_i^M \mathbf{x}_1^i \mathbf{x}_2^{iT}$, $\mathbf{C}_{11} = \sum_i^M \mathbf{x}_1^i \mathbf{x}_1^{iT}$, $\mathbf{C}_{22} = \sum_i^M \mathbf{x}_2^i \mathbf{x}_2^{iT}$, and $U = \mathbf{w}^T \mathbf{X}$ is the CCA projection of $\mathbf{X}^T$.

4.2.4 Regularized Canonical Correlation Analysis

RCCA [134, 135] corrects for instability in $\mathbf{x}_1$ and $\mathbf{x}_2$ by using regularization on the covariance matrices $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ to remove singularity in these terms. Correlation matrices $\mathbf{C}_{12}$ and $\mathbf{C}_{21}$ are unaffected, while the matrices $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ become $\mathbf{C}_{11}+\gamma_1\mathbf{I}_1$ and $\mathbf{C}_{22}+\gamma_2\mathbf{I}_2$. RCCA can be solved via the generalized eigenvalue problems [164]

$$\mathbf{C}_{12}(\mathbf{C}_{22}+\gamma_2\mathbf{I}_2)^{-1}\mathbf{C}_{21} = \lambda(\mathbf{C}_{11}+\gamma_1\mathbf{I}_1)\mathbf{w}_1 \tag{4.4}$$

and

$$\mathbf{C}_{21}(\mathbf{C}_{11}+\gamma_1\mathbf{I}_1)^{-1}\mathbf{C}_{12} = \lambda(\mathbf{C}_{22}+\gamma_2\mathbf{I}_2)\mathbf{w}_2. \tag{4.5}$$

The regularization parameters are subsequently selected as described in Golugula et al. [88]. $\mathbf{w}_1^i$ and $\mathbf{w}_2^i$ refer to the weights obtained from RCCA when samples $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ are removed, where $i \in \{1, 2, \ldots, n\}$. γ_1 and γ_2 are varied across $\theta_1 \leq \gamma_1, \gamma_2 \leq \theta_2$ and chosen by grid search [165] optimization of the following cost function [166]:

$$\max_{\gamma_1,\gamma_2} \left| corr\Big([\mathbf{x}_1^i\mathbf{w}_1^i]_{i=1}^n, [\mathbf{x}_2^i\mathbf{w}_2^i]_{i=1}^n\Big)\right| \tag{4.6}$$

where $corr(\cdot,\cdot)$ refers to the Pearson's correlation coefficient [167]. Equation 4.2.4 illustrates a leave one out cross-validation step to find the combination of γ_1, γ_2 and omitted $\mathbf{x}^i$ gives the maximum correlation. γ_1 and γ_2 are chosen using the embedding component with the highest associated eigenvalue λ and then adjusted for the remaining dimensions [166].

4.2.5 Supervised Regularized Canonical Correlation Analysis

Supervised Regularized Canonical Correlation Analysis (SRCCA) [88] is an extension of RCCA, which chooses γ_1 and γ_2 using a supervised feature selection method rather than optimization via correlation. We define $\mathbb{Y}_1$ and $\mathbb{Y}_2$ as the object classes 1 and 2 respectively. Furthermore, we define μ_1, μ_2, σ_{21}, σ_{22}, and n_1, n_2 as the set of means,

variances, and sample sizes for each class $\mathbf{W}_1$ and $\mathbf{W}_2$. The projections $U = \mathbf{x}_1\mathbf{w}_1$ and $U = \mathbf{x}_2\mathbf{w}_2$ can be split across its samples corresponding to their labels $\mathbb{Y}_1$ and $\mathbb{Y}_2$, where $U^{\mathbb{Y}_1}$ and $U^{\mathbb{Y}_2}$, respectively, represent the $\eta_1+\eta_2 = n$ total samples in the dataset. These $U^{\mathbb{Y}_1}$ and $U^{\mathbb{Y}_2}$ can then be used to calculate the discriminatory contribution of each feature given the samples of the two classes. For this work, SRCCA was optimized with the non-parametric Wilcoxon Rank Sum Test (SRCCA-WRST) to choose the regularization parameters, γ_1 and γ_2, as it was found to perform the best in [88]. For each feature in U, the value for each sample is sorted in ascending order and is used to calculate a discriminatory score

$$\max_{\gamma_1,\gamma_2}\left\{\Big(\sum_{i=1}^{\eta_2} b_i - \frac{\eta_2(\eta_2+1)}{2}\Big), \Big(\eta_1\eta_2 - \sum_{i=1}^{\eta_2} b_i - \frac{\eta_2(\eta_2+1)}{2}\Big)\right\}, \tag{4.7}$$

where b_i denotes the rank of sample $i \in \mathbb{Y}$. Similar to RCCA, for SRCCA, γ_1 and γ_2 are selected using the embedding representation which results in the most discriminatory WRST score, and subsequently adjusted for the remaining dimensions.

4.3 Multi-View Canonical Correlation Analysis (MVCCA)

Multi-view CCA (MVCCA) builds upon CCA (refer to Section 4.2.3 for an overview of CCA) by accounting for situations where data samples can be described by greater than two modalities. Since the joint correlation of three or more variables does not formally exist, an alternative solution is to sequentially consider the correlations of each pair of samples [136]. However, this approach is suboptimal and requires an inefficient iterative optimization. An alternative pairwise MVCCA approach is described here which maximizes the sum of the correlations between all pairs of modalities.

Given n data samples, each comprising $s = s_1+s_2+\ldots+s_M$ features: $\{\mathbf{x}_1, ...\mathbf{x}_2, ..., \mathbf{x}_s\}$ from M modalities, $m \in \{1, 2, ..., M\}$, this implementation of pairwise MVCCA seeks a set of linear transformations $\{\mathbf{w}_1 \in \mathbb{R}^{s_1\times 1}, \mathbf{w}_2 \in \mathbb{R}^{s_2\times 1}, ..., \mathbf{w}_M \in \mathbb{R}^{s_M\times 1}\}$ such that the sum of the correlations across all pairs of modalities is maximized,

$$\underset{\{\mathbf{w}_1,...\mathbf{w}_m...,\mathbf{w}_M\}}{\operatorname{argmax}} \sum_{m=1}^{M}\sum_{t=1,t\neq m}^{M} \frac{\mathbf{w}_m^T\mathbf{C}_{mt}\mathbf{w}_t}{\sqrt{\mathbf{w}_m^T\mathbf{C}_{mm}\mathbf{w}_m\mathbf{w}_t^T\mathbf{C}_{tt}\mathbf{w}_t}}, \tag{4.8}$$

where $\mathbf{C}_{mt} = \sum_i^M \mathbf{x}_m^i \mathbf{x}_t^{iT}$, $\mathbf{C}_{mm} = \sum_i^M \mathbf{x}_m^i \mathbf{x}_m^{iT}$, $\mathbf{C}_{tt} = \sum_i^M \mathbf{x}_t^i \mathbf{x}_t^{iT}$. The scaling of $\mathbf{w}$ does not affect the arg max solution, allowing Eq. 4.8 to be written as:

$$\begin{aligned} \underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{argmax}} \quad & \sum_{m=1}^{M} \sum_{t=1,t\neq M}^{M} \mathbf{w}_m^T \mathbf{C}_{mt} \mathbf{w}_t \\ s.t. \quad & \mathbf{w}_1^T \mathbf{C}_{11} \mathbf{w}_1 = 1, \ldots, \mathbf{w}_M^T \mathbf{C}_{MM} \mathbf{w}_M = 1. \end{aligned} \tag{4.9}$$

We define $\mathbf{w} = [\mathbf{w}_1^T \ \mathbf{w}_2^T ... \mathbf{w}_M^T]^T$, $\mathbf{w} \in \mathbb{R}^{(s_1+s_2+\ldots+s_M)\times 1}$, such that Eq. 4.9 can be rewritten in compact matrix form:

$$\begin{aligned} \underset{\mathbf{w}}{\text{argmax}} \quad & \mathbf{w}^T \bar{\mathbf{C}} \mathbf{w} \\ s.t \quad & \mathbf{w}^T \bar{\mathbf{C}}_d \mathbf{w} = 1 \ and \ \mathbf{w}_1^T \mathbf{C}_{11} \mathbf{w}_1 = \ldots = \mathbf{w}_M^T \mathbf{C}_{MM} \mathbf{w}_M, \end{aligned}$$

where

$$\bar{\mathbf{C}} = \begin{bmatrix} \mathbf{0} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1M} \\ \mathbf{C}_{21} & \mathbf{0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{C}_{(M-1)M} \\ \mathbf{C}_{M1} & \cdots & \mathbf{C}_{M(M-1)} & \mathbf{0} \end{bmatrix},$$

$$\bar{\mathbf{C}}_d = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_{MM} \end{bmatrix}. \tag{4.10}$$

For the general situation $\mathbf{W} \in \mathbb{R}^{(s_1+s_2+\ldots+s_M)\times n}, 1 \leq n \leq min(s_1, s_2, \ldots, s_M)$, the corresponding objective function of pairwise MVCCA becomes

$$\begin{aligned} \underset{\mathbf{W}}{\text{argmax}} \quad & trace(\mathbf{W}^T \bar{\mathbf{C}} \mathbf{W}) \\ s.t \quad & \mathbf{W}^T \bar{\mathbf{C}}_d \mathbf{W} = \mathbf{I} \\ & \mathbf{w}_1^T \mathbf{C}_{11} \mathbf{w}_1 = \ldots = \mathbf{w}_M^T \mathbf{C}_{MM} \mathbf{w}_M, \end{aligned}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix and $\mathbf{W} = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ ... \ \mathbf{w}_M^T]^T$. Thus, MVCCA can be formulated into a simple optimization problem for obtaining the weights $\mathbf{w}$, for which we will show a non-iterative solution in the following section.

4.4 Supervised Multi-View Canonical Correlation Analysis (sMVCCA)

4.4.1 Formulation

When used for classification, the MVCCA representation does not guarantee better class separation. We hereby present supervised MVCCA (sMVCCA), which incorporates label information to improve upon classification compared to MVCCA. It has been shown that when correlating data samples with corresponding class labels, we obtain the LDA projection as the solution [168]. sMVCCA leverages this idea with pairwise MVCCA, defining data $\mathbf{X} \in \mathbb{R}^{n \times (s_1 + \ldots + s_m + \ldots + s_M)}$ and class labels encoded as $\mathbf{Y} \in \mathbb{R}^{n \times g}$, where g is the number of object classes. Furthermore, sMVCCA redefines the general weight matrix $\mathbf{W}$ for multi-modality data in MVCCA as $\mathbf{W}_x = [\mathbf{w}_1^T \mathbf{w}_2^T \ldots \mathbf{w}_M^T]$ and treats $\mathbf{W}_y = [\mathbf{w}_{\mathbb{Y}_1}^T \mathbf{w}_{\mathbb{Y}_2}^T \ldots \mathbf{w}_{\mathbb{Y}_g}^T]$ as a special weight matrix for the label information. This yields the following objective function of optimizing the weights $\mathbf{w}$.

$$
\begin{aligned}
\underset{\mathbf{W}_x, \mathbf{W}_y}{\operatorname{argmax}} \quad & trace(\mathbf{W}_x^T \bar{\mathbf{C}} \mathbf{W}_x) + 2trace(\mathbf{W}_x^T \mathbf{X}^T \mathbf{Y} \mathbf{W}_y) \\
= \quad & trace(\begin{bmatrix} \mathbf{W}_x^T & \mathbf{W}_y^T \end{bmatrix} \begin{bmatrix} \bar{\mathbf{C}} & \mathbf{X}^T\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix}) \\
= \quad & trace(\hat{\mathbf{W}}^T \hat{\mathbf{C}} \hat{\mathbf{W}})
\end{aligned}
$$

$s.t.$

$$
\begin{bmatrix} \mathbf{W}_x^T & \mathbf{W}_y^T \end{bmatrix} \begin{bmatrix} \bar{\mathbf{C}}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}^T\mathbf{Y} \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix} = \mathbf{I}
$$

$$
\Leftrightarrow \hat{\mathbf{W}}^T \hat{\mathbf{C}}_d \hat{\mathbf{W}} = \mathbf{I} \tag{4.11}
$$

$$
\mathbf{W}_1^T \bar{\mathbf{C}}_{11} \mathbf{W}_1 = \ldots = \mathbf{W}_M^T \bar{\mathbf{C}}_{MM} \mathbf{W}_M = \mathbf{W}_y^T \mathbf{Y}^T \mathbf{Y} \mathbf{W}_y. \tag{4.12}
$$

4.4.2 Optimization

(A) sMVCCA is optimized based on $\hat{\mathbf{C}}_d$. When $\hat{\mathbf{C}}_d$ is non-singular, we solve sMVCCA in two steps:

1. Ignoring the constraint in Equation (4.12), we have

$$\begin{aligned} \underset{\mathbf{W}_x,\mathbf{W}_y}{\text{argmax}} \quad & trace(\hat{\mathbf{W}}^T\hat{\mathbf{C}}\hat{\mathbf{W}}) \\ s.t. \quad & \hat{\mathbf{W}}^T\hat{\mathbf{C}}_d\hat{\mathbf{W}} = \mathbf{I}. \end{aligned} \tag{4.13}$$

The solution $\mathbf{W}^*$ of this reduced problem consists of the eigenvectors of the d-largest eigenvalues of a generalized eigenvalue system:

$$\hat{\mathbf{C}}\hat{\mathbf{W}} = \hat{\mathbf{C}}_d\hat{\mathbf{W}}\mathbf{\Lambda}. \tag{4.14}$$

where $\mathbf{\Lambda}$ is the diagonal matrix containing eigenvalues. Since $\hat{\mathbf{C}}_d$ is non-singular, we can simply solve

$$\hat{\mathbf{C}}_d^{-1}\hat{\mathbf{C}}\hat{\mathbf{W}} = \hat{\mathbf{W}}\mathbf{\Lambda}. \tag{4.15}$$

Since $\hat{\mathbf{C}}_d^{-1}\hat{\mathbf{C}}$ is non-symmetric, we use SVD decomposition to get the largest eigenvalues and eigenvectors. As compared to direct eigendecomposition, SVD is numerically more stable.

2. We apply the constraint in Equation 4.12 to normalize $\mathbf{W}^*$:

$$\mathbf{W}_m^{**} = \mathbf{W}_m^*(\mathbf{W}_m^{*T}\bar{\mathbf{C}}_{mm}\mathbf{W}_m^*)^{-\frac{1}{2}}, \forall m \in \{1, ..., M\}. \tag{4.16}$$

(B) For dimensionality reduction, it is common that feature dimension is larger than the number of samples. In this case $\hat{\mathbf{C}}_d$ will be singular. Solving sMVCCA with Equation (4.14) will then be numerically unstable. To avoid the problem, we impose a regularization term to make $\hat{\mathbf{C}}_d$ non-singular:

$$\hat{\mathbf{C}}\hat{\mathbf{W}} = \lambda(\hat{\mathbf{C}}_d + \gamma\mathbf{I})\hat{\mathbf{W}}. \tag{4.17}$$

where γ is a small hyperparameter to balance the regularization term, e.g. $\gamma = 0.05$ as in RCCA discussed in Section 4.2.4. Then we can use the solution in **(A)** to solve sMVCCA.

4.4.3 Encoding of Y

The extended form of Eqn 4.15 is

$$\begin{bmatrix} \bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}} & \bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}} \\ \bar{\mathbf{C}}_{yy}^{-1}\bar{\mathbf{C}}_{yx} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix} = \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix} \mathbf{\Lambda},$$

which corresponds to

$$\begin{cases} \bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}}\mathbf{W}_x + \bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}}_{xy}\mathbf{W}_y = \mathbf{W}_x\mathbf{\Lambda}_x \\ \bar{\mathbf{C}}_{yy}^{-1}\bar{\mathbf{C}}_{yx}\mathbf{W}_x = \mathbf{W}_y\mathbf{\Lambda}_y. \end{cases} . \tag{4.18}$$

As compared to the two-view CCA between data $\mathbf{X}$ and labels $\mathbf{Y}$, which is posed as

$$\begin{cases} \bar{\mathbf{C}}_{xx}^{-1}\bar{\mathbf{C}}\mathbf{W}_y = \mathbf{W}_x\mathbf{\Lambda}_x \\ \bar{\mathbf{C}}_{yy}^{-1}\bar{\mathbf{C}}_{yx}\mathbf{W}_x = \mathbf{W}_y\mathbf{\Lambda}_y. \end{cases} , \tag{4.19}$$

Eqn 4.18 considers $\bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}}\mathbf{W}_x$ to couple the correlations within $\mathbf{X}$ with the correlations between $\mathbf{X}$ and $\mathbf{Y}$. When $\mathbf{Y}$ is encoded as

$$\mathbf{Y} = c \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \cdots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \cdots & \mathbf{1}_{n_g} \end{bmatrix}_{n \times g} \tag{4.20}$$

where n is the number of samples, g is the number of classes and c a constant, the projection direction $\mathbf{W}_x$ of two-view CCA is equivalent to performing LDA [133], and the constant c has no influence on $\mathbf{W}_x$. In sMVCCA, from Eqn 4.18 we can easily derive

$$\begin{cases} \mathbf{W}_x^T\bar{\mathbf{C}}\mathbf{W}_x + \mathbf{\Lambda}_y = \mathbf{\Lambda}_x \\ \mathbf{W}_y = \bar{\mathbf{C}}_{yy}^{-1}\bar{\mathbf{C}}_{yx}\mathbf{W}_x\mathbf{\Lambda}_y^{-1} \end{cases} . \tag{4.21}$$

Plugging Eqn. 4.21 into the first equation of Eqn. 4.18 leads to

$$\bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}}\mathbf{W}_x + \bar{\mathbf{C}}_d^{-1}\bar{\mathbf{C}}_{xy}\bar{\mathbf{C}}_{yy}^{-1}\bar{\mathbf{C}}_{yx}\mathbf{W}_x(\mathbf{\Lambda}_x - \mathbf{W}_x^T\bar{\mathbf{C}}\mathbf{W}_x)^{-1} = \mathbf{W}_x\mathbf{\Lambda}_x \tag{4.22}$$

Since $\bar{\mathbf{C}}_{xy}\bar{\mathbf{C}}_{yy}^{-1}\bar{\mathbf{C}}_{yx}$ is independent of the value of c, c has no affect on determining $\mathbf{W}_x$. Therefore, we can use *1-of-Class* (by setting $c = 1$) encoding of $\mathbf{Y}$ for sMVCCA. Alternatively, we can emphasize samples close to the classification boundary by adopting the *Soft-1-of-Class* strategy [168].

4.5 Extension of sMVCCA via Spearman Rank

Previous work on CCA has focused highly on correlation to provide an optimal embedding, as opposed to classification performance, and thus Pearson correlation has been

used as the standard for optimization. However, to the best of the author's knowledge, Spearman correlation [169] has not been investigated for CCA. For Spearman calculation, we provide a transformation of each feature $f_j \in \mathbf{X}$ to a ranked feature $\hat{f}_j$, such that all values $\hat{f}_j \in \hat{\mathbf{X}}$ is an integer in the set $\{1, \ldots, n\}$. This input is used for Equation 4.9 such that $\mathbf{C}_{mt} = \sum_i^M \hat{\mathbf{x}}_m^j \hat{\mathbf{x}}_t^{iT}$, $\mathbf{C}_{mm} = \sum_i^M \hat{\mathbf{x}}_m^j \hat{\mathbf{x}}_m^{iT}$, $\mathbf{C}_{tt} = \sum_i^M \hat{\mathbf{x}}_t^j \hat{\mathbf{x}}_t^{iT}$.

Using ranked features may provide better optimization compared to the original features as Pearson correlation may reject features which otherwise have high predictive value but are simply scaled non-linearly as illustrated in Figure 4.1, where unranked X and Y represent features f_X^j and f_Y^j, respectively while ranked X and Y represent features $\hat{f}_X^j$ and $\hat{f}_Y^j$ respectively. Thus, the use of Spearman rank correlation allows for a less stringent definition of correlation which is more robust to non-linearity and outliers, while retaining the intrinsic discriminatory properties of the original features as Pearson correlation.

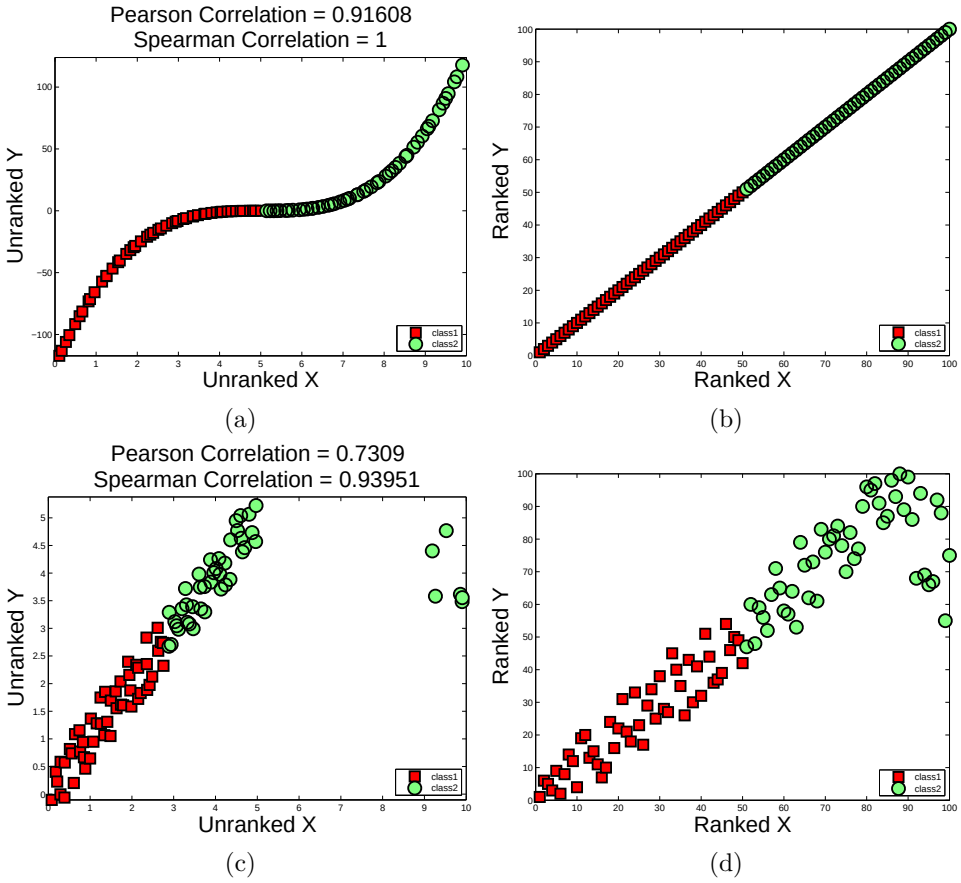


Figure 4.1: Comparison of Pearson correlation versus Spearman rank correlation on (a,b) non-linear data and (c,d) data with outliers. For these datasets, Spearman correlation gives a higher optimization showing $\rho = 1$ and $\rho = 0.939$ for Spearman compared to $r = 0.916$ and $r = 0.731$ for Pearson. Note that the discriminatory properties are retained in both unranked and ranked representations of features X and Y. Thus, the use of Spearman rank can provide better optimization while retaining the ability to discriminate the object classes.

Chapter 5

Novel Strategies in Quantitative Histomorphometry

5.1 Overview

In this chapter, we showcase our work in quantitative histomorphometry for automated analysis of prostate cancer tissue. In Section 5.2, we provide background on QH as it pertains to the diagnosis and prediction of aggressive prostate cancer. We subsequently introduce our methodologies, Cell Orientation Entropy (COrE) and Co-occurring Gland Tensors (CGTs) and demonstrate their efficacy as compared to previous work in two separate validation cohorts of post-operative prostate cancer patients in Chapters 5.4 and 5.9, respectively.

5.2 Role of Quantitative Histomorphometry in Prostate Cancer

The recent advent of digital whole slide scanners has allowed for the development of quantitative histomorphometry (QH), sophisticated computerized image analysis tools for automated scoring of digitized histology images. Gleason scoring (GS), while predictive, is subject to considerable inter-reviewer variability [22] due to the qualitativeness of its evaluation, a problem that is alleviated via QH. Additionally, the automated nature of QH is valuable to pathologists, as the grading of a single patient biopsy typically involves the inspection of at least 20 slides. The inspection of many high resolution whole slides for aggressive prostate cancer can become a laborious task.

Quantitative histomorphometry (QH) has been utilized successfully in a wide range of applications from cancer detection to prognosis [30, 170, 171]. Monaco et al. has utilized markov random fields for detection of prostate cancer on whole mount histology [170]. To differentiate Gleason pattern 3 from Gleason pattern 4 on prostate needle

biopsies, Doyle et al. has been extracting architectural features of nuclei [30]. Furthermore, Veltri et al. [171] has calculated nuclear roundness variance for the prediction of progression following RP.

Prevoiusly, analysis of image pixel intensity has been explored for quantifying CaP [146, 172]. For the purpose of automated CaP grading, Jafari-Khouzani et al. [172] examined the role of image texture features based on co-occurrence matrices. Other texture features such as mean pixel intensity and Gabor filters [146] have also been used to predict CaP. However, these features are based on pixel intensity and lack direct biological significance.

Attempts to segment specific structures relevant for CaP have been widespread. In addition to color and texture, Tabesh et al. [173] also investigated structural morphology to evaluate prostate histopathology in terms of GS. In [174], morphological descriptors such as gland size and perimeter ratio were shown to distinguish benign and malignant histological regions. In [171], Veltri et al. investigated nuclear morphology using a descriptor called nuclear roundness variance. Cell morphology was found to exceed GS for predicting CaP aggressiveness. However, complex spatial relationships between structures are not investigated.

To model tissue architecture, much work has been done in constructing graphs networks around specific structures in the tissue [174–177]. Voronoi- and Delaunay-based graph tessellations have been suggested to describe the architecture of various structures in CaP histology [175]. Previously, Doyle et al. [174] had shown Minimum Spanning Trees, in addition to Voronoi, Delaunay features to be a strong predictors of Gleason grade. However, these features are derived from fully connected graphs, whose edges traverse across epithelial and stromal regions. By connecting globally, fully connected graphs tend to dilute the contribution of the tumor morphologic features specific to the cancer epithelium.

Given that global graphs are not sensitive to local organization, which may be critical for characterizing tumor aggressiveness, the analysis of smaller, local subgraphs may provide a useful alternative. Unlike global graphs (e.g. Voronoi and Delaunay) that aim to capture a global architectural signature for the tumor, subgraph construction

can allow for quantification of local interactions within flexible localized neighborhoods. Bilgen et al. [176] extracted features from different types of local cell graphs for classification breast tissue. Furthermore, Ali et al. [177] examined cell cluster graphs for the prediction of biochemical recurrence in prostate tissue microarrays.

5.3 A need for novel quantitative histomorphometry for predicting aggressive prostate cancer

There remains evidence that we have not yet reached the potential for predicting aggressive prostate cancer from quantitative histomorphometry. There remain updates to the classification of Gleason patterns within the International Society of Urological Pathology (ISUP) for the purpose of improving differentiation of aggressive CaP [16]. Furthermore, novel methodologies in segmentation [178] and feature extraction [177, 179] from histological images continue to be made which improve upon the state-of-the-art.

The following chapters represent two methods of quantitative histomorphometry developed for the purpose of evaluating prostate cancer histology. In Chapter 5.4, we describe the method of Cell Orientation Entropy and evaluate its predictive value on prostate tissue microarrays. In Chapter 5.9, we describe the method of Co-occurring Gland Tensors in Localized Subgraphs and evaluate its predictive value on prostate whole slides.

5.4 Cell Orientation Entropy (COrE) for Predicting Biochemical Recurrence in Prostate Tissue Microarrays

In the following sections, we present a new set of quantitative histomorphometric (QH) features called cell orientation entropy (COrE), which aim to capture the local directional information of epithelial cancer cells. CaP is fundamentally a disease of glandular disorganization and the resulting breakdown in nuclei orientation is related to its grade [16]. Epithelial cells align themselves with respect to the glands, and thus display a coherent directionality. However, cancerous prostate glands are less well formed, resulting in a more chaotic organization and orientation of the surrounding nuclei.

COrE attempts to model this difference between cancerous and benign regions via a novel scheme, unique to digital pathology image analysis. We believe this work to be the first rigorous attempt to quantitatively model cell orientation and explore the linkage between cell orientation and CaP aggressiveness. Firstly, while previous work has focused on global graph networks for characterizing tumor architecture, COrE employs subgraphs to construct local cell networks and thereby quantify second order statistics based on co-occurrence matrices of cell orientations. Secondly, while co-occurrence matrices are commonly used to describe image textures [161], by quantifying second order statistics of image intensities, this is the first instance of the use of the co-occurrence matrix to evaluate local, higher order interactions of nuclear orientations. These second order local statistical features of nuclear orientation yield a rich set of descriptors for distinguishing the different CaP tumor classes.

5.5 Cell Orientation Entropy (COrE)

5.5.1 Automated Cell Segmentation

We employed an energy based segmentation scheme presented in [178] to detect and segment a set of cell/nuclei $\gamma_i, p \in \{1, 2, \ldots, n\}$, where n is the total number of nuclei found. This segmentation scheme is a synergy of boundary and region-based active contour models that incorporates shape priors in a level set formulation with automated initialization based on watershed. The energy functional of the active contour is comprised of three terms. The combined shape, boundary and region-based functional formulation [178] is given below:

$$F = \underbrace{\beta_s \int_{\Omega} (\phi(\mathbf{x}) - \psi(\mathbf{x}))^2 |\nabla\phi| \delta(\phi) d\mathbf{x} +}_{Shape+boundary force} \quad \underbrace{\beta_r \int_{\Omega} \Theta_{in} H_{\psi} d\mathbf{x} + \int_{\Omega} \Theta_{out} H_{-\psi} d\mathbf{x}}_{Region force} \tag{5.1}$$

where $\beta_s, \beta_r > 0$ are constants that balance contributions of the boundary based shape prior and the region term. $\{\phi\}$ is a level set function, ψ is the shape prior, $\delta(\phi)$ is the contour measure on $\{\phi = 0\}$, $H(.)$ is the Heaviside function, $\Theta_r = |I - u_r|^2 + \mu|\nabla u_r|^2$ and $r \in \{in, out\}$.

The first term is the prior shape term modeled on the prostate nuclei, thereby constraining the deformation achievable by the active contour. The second term, a boundary-based term detects the nuclear boundaries from image gradients. The third term drives the shape prior and the contour towards the nuclear boundary based on region statistics.

5.5.2 Calculating Cell Orientation

To determine the directionality for each cell γ_i, we perform principal component analysis on a set of boundary points $[x_i, y_i]$ to obtain the principal components $Z = [z_1, z_2]$. The first principal component z_1 describes the directionality of the cell in the form of the major axis $z_1 =< z_1^x, z_1^y >$, along which the greatest variance occurs in the nuclear boundary. The principal axis z_1 is converted to an angle $\bar{\theta}(\gamma_i) \in [0°180°]$ counterclockwise from the vector $< 1, 0 >$ by $\bar{\theta}(\gamma_i) = \frac{180°}{\pi} \arctan(\frac{z_1^y}{z_1^x})$.

5.5.3 Local Cell Subgraphs

Pairwise spatial relationships between cells are defined via sparsified graphs. A graph $G = \{V, E\}$, where V represents the set of n nuclear centroids $\gamma_i, \gamma_j \in V$, $i, j \in \{1, 2, \ldots, n\}$ as nodes, and E represents the set of edges which connect them. The edges between all pairs of nodes γ_i, γ_j are determined via the probabilistic decaying function

$$E = \{(i, j) : r < d(i, j)^{-\alpha}, \forall \gamma_i, \gamma_j \in V\}, \tag{5.2}$$

where $d(i, j)$ represents the Euclidean distance between γ_i and γ_j. $\alpha \geq 0$ controls the density of the graph, where α approaching 0 represents a high probability of connecting nodes while α approaching ∞ represents a low probability. $r \in [0, 1]$ is an empirically determined edge threshold.

Table 5.1: Representative COrE features

COrE Feature (Θ)	**Description**
Entropy	$\sum_{a,b} -\mathcal{C}(a,b)\ \log(\mathcal{C}(a,b)))$
Energy	$\sum_{a,b} \mathcal{C}(a,b)^2$
Correlation	$\sum_{a,b} \frac{(a-\mu a)(b-\mu b)\mathcal{C}(a,b)}{\sigma_a \sigma_b}$
Contrast (variance)	$\sum_{a,b} \lvert a-b\rvert^2 \mathcal{C}(a,b)$

5.5.4 Calculating Second Order Statistics for Cell Orientation

The objects of interest for calculating COrE features are the cell directions given by a discretization of the angles $\bar{\theta}(\gamma_i)$, such that $\theta(\gamma_i) = \omega \times \text{ceil}(\frac{\bar{\theta}}{\omega})$, where ω is a discretization factor. Neighbors defined by the local cell subgraphs G, allow us to define neighborhoods for each cell. For each $\gamma_i \in V$, we define a neighborhood $\mathcal{N}_i$, to include all $\gamma_j \in V$ where a path between γ_i and γ_j exists in graph G.

An $N \times N$ co-occurrence matrix $\mathcal{C}$ subsequently captures angle pairs which co-occur in each neighborhood $\mathcal{N}_i$, such that for each $\mathcal{N}_i$,

$$\mathcal{C}_{\mathcal{N}_i}(a,b) = \sum_{\gamma_i,\gamma_j}^{\mathcal{N}_i} \sum_{a,b=1}^{N} \begin{cases} 1, & \text{if } \theta(\gamma_i)=a \text{ and } \theta(\gamma_j)=b \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

where $N = \frac{180}{\omega}$, the number of discrete angular bins. We then extract second order statistical features (Contrast energy, Contrast inverse moment, Contrast average, Contrast variance, Contrast entropy, Intensity average, Intensity variance, Intensity entropy, Entropy, Energy, Correlation, Information measure 1, Information measure 2) from each co-occurrence matrix $\mathcal{C}_{\mathcal{N}_i}(a,b)$. Selected formulations are described in Table 5.1. Mean, standard deviation, and range of Θ across all $\mathcal{N}_i$ constitute the set of 39 COrE features.

5.6 Experimental Design

5.6.1 Prostate Cancer Tissue Microarray Data

While COrE is extensible towards the histological analysis of other pathological diseases, we have chosen prostate cancer (CaP) as a test case for this initial work. Our dataset comprised of histologic image samples in the form of tissue microarray (TMA) cores from 19 CaP patients who experienced BCR within 10 years of RP, and from 20 patients

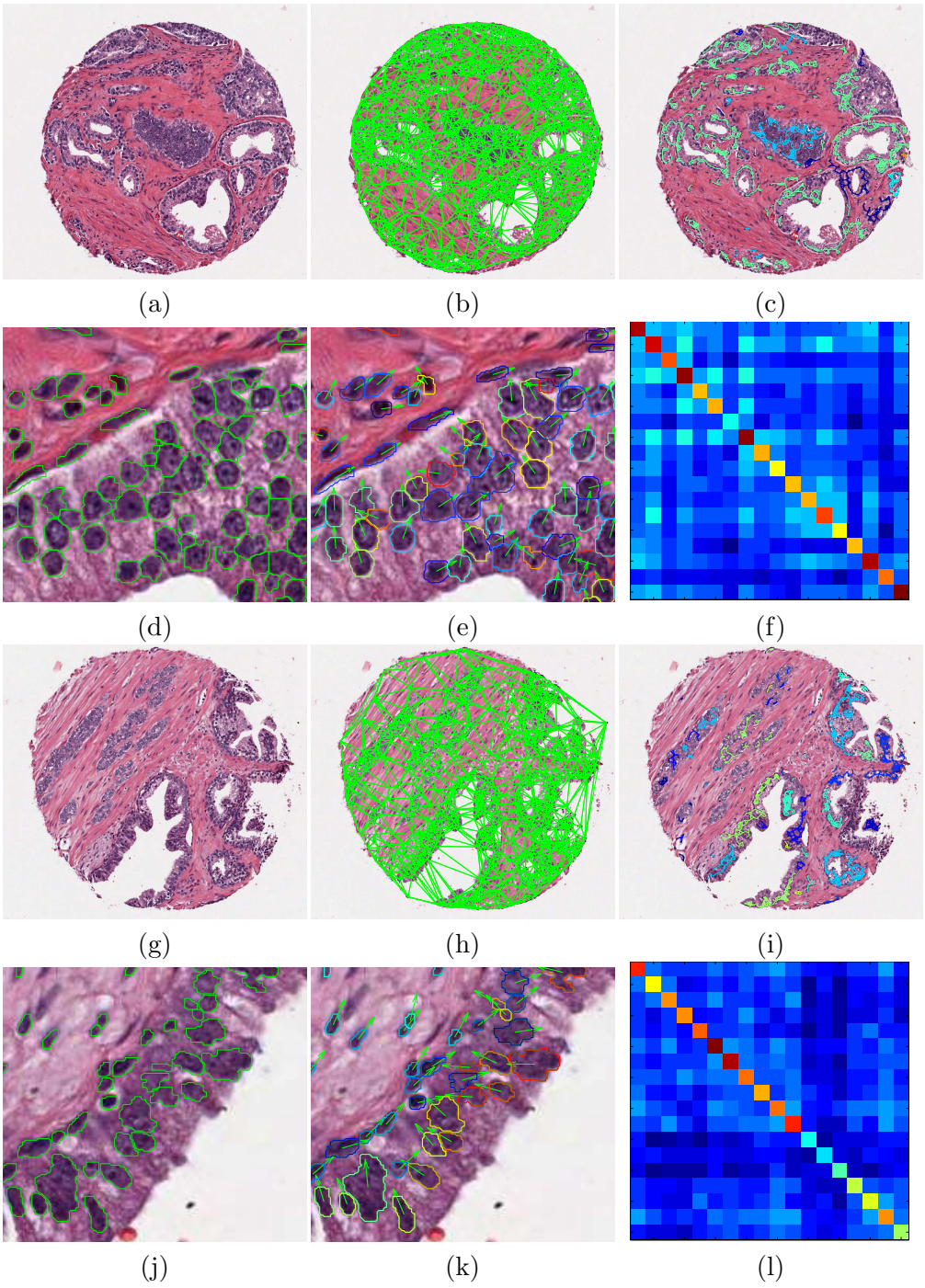


Figure 5.1: Prostate TMAs pertaining to (a)-(f) BCR and (g)-(l) NR case studies. Nuclei are used as nodes for calculation of (b),(h) Delaunay graphs. Automated segmentation (d),(j) defines the nuclear boundaries and locations from the TMA image. (e),(k) Cell orientation vectors are calculated from the segmentated boundaries (illustrated via different boundary colors). (c),(i) Subgraphs are formed by connecting neighboring cells. COrE features calculate contrast in the cell orientation (with dark regions showing more angular coherence and bright regions showing more disorder). Summation of the co-occurrence matrices provide a visual interpretation of disorder, where (f) shows brighter co-occurrence values in the off-diagonal cells, suggesting higher co-occurrence of nuclei of differing orientations compared to (l).

Table 5.2: Summary of 151 nuclear morphologic features

Cell Morphology	#	**Description**
	100	Area Ratio, Distance Ratio, Standard Deviation of Distance, Variance of Distance, Distance Ratio, Perimeter Ratio, Smoothness, Invariant Moment 1-7, Fractal Dimension, Fourier Descriptor 1-10 (Mean, Std. Dev, Median, Min / Max of each)
Cell Architecture		**Description**
Voronoi Diagram	12	Polygon area, perimeter, chord length: mean, std. dev., min/max ratio, disorder
Delaunay Triangulation	8	Triangle side length, area: mean, std. dev., min/max ratio, disorder
Minimum Spanning Tree	4	Edge length: mean, std. dev., min/max ratio, disorder
Nearest Neighbors	27	Density of nuclei, distance to nearest nuclei

who did not (NR). Patients were matched for GS 7 and tumor stage 3A. CaP tissue included in the TMAs were selected and reviewed by an expert pathologist. For this study, each of 39 patients was represented by a single randomly selected 0.6mm TMA core image, chosen from a set of 4 TMA cores taken for that patient.

5.6.2 Comparative Methods for Evaluating COrE

We compared the efficacy of COrE features with previously studied nuclear features. The shape of individual nuclei has previously been shown to be prognostic of GS [31, 171]. The set of 100 cell morphology features representing mean, standard deviation of nuclear size and shape are summarized in Table 5.2.

Nuclear/cell architecture refers to the spatial arrangement of cells in cancerous and benign tissue. 51 architectural image features describing the nuclear arrangement were extracted as described in [31]. Voronoi diagrams, Delaunay Triangulation and Minimum Spanning Trees were constructed on the digital histologic image using the nuclear centroids as vertices (See Table 5.2).

For all feature sets, the nuclear segmentations from Section 2.1 were used to calculate the cell boundaries and centroids. In total, we investigated the performance of 4 feature cohorts: (1) 100 features describing cell morphology, (2) 51 features describing cell architectures, (3) 39 features describing cell orientation entropy (COrE), and (4) the combined feature set spanning cohorts (1-3).

5.6.3 Random Forest Classifier

In this study, we demonstrate the efficacy of including COrE features for improving classification accuracy and area under the receiver operating characteristic curve (AUC) in predicting BCR in CaP patients from prostate TMAs. Randomized 3-fold cross validation was performed on the top 10 most informative features selected via Student t-test for each of 4 feature cohorts defined in Section 3.2. Classification was performed using a random forest classifier.

5.7 Results and Discussion

5.7.1 Comparison with Nuclear Morphology and Architecture

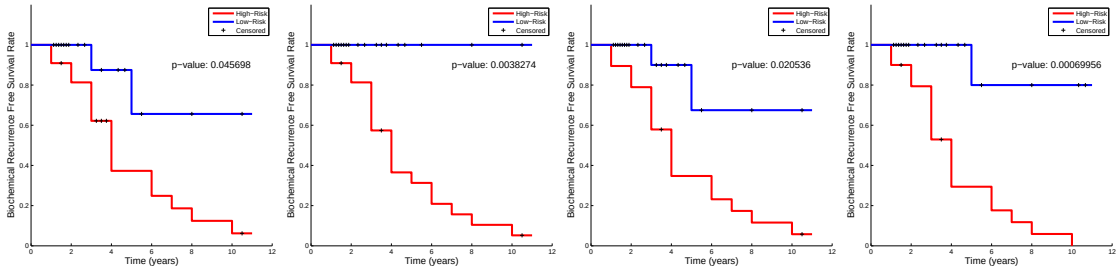
Figure 5.1 reveals the ability of the COrE features to capture the differences in angular disorder across localized cell networks and illustrates the differences between the BCR and NR cases in terms of the COrE features.

In Table 5.3, we can summarize the performance of feature descriptors describing cell architecture and cell morphology which appear to have a maximum BCR prediction accuracy of 79.9%. However, by inclusion of novel cell orientation entropy (COrE) features, the overall classifier accuracy improves to 82.7%. Similar improvements are also observed in terms classification AUC. This reflects the utility of COrE features as a valuable prognostic measurement for predicting BCR in conjunction with previously described nuclear morphologic features.

Classifier improvement following inclusion of COrE features suggests that many of the new COrE features are non-correlated with previously defined cell architectural and morphological feature sets. This distinction is illustrated in Figure 5.1, where we observe the differences between COrE features compared with those obtained from Voronoi and Delaunay graphs. These graphs span across stromal and epithelial regions, while COrE features are limited to subgraphs in localized regions. It is also important to note that the combination of COrE and nuclear morphologic features clearly and significantly outperform the clinical standard of pathologist grade, which classified all cases as GS 7.

Table 5.3: 100 runs of 3-fold Random Forest Classification

	Architecture	Morphology	COrE	Arch + Morph + COrE
Accuracy	71.2 ± 4.2%	79.9 ± 3.7%	74.6 ± 4.1%	82.7 ± 3.1%
AUC	0.641 ± 0.054	0.773 ± 0.042	0.688 ± 0.063	0.809 ± 0.037



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

Figure 5.2: Kaplan-Meier curves, (a) COrE + Morph + Arch (All QH), (b) Gleason Sum + QH, (c) Tumor Stage + QH, and (d) All QH + Gleason Sum + Tumor Stage, illustrate the outcome (biochemical recurrence) of patients stratified into high-risk and low-risk groups via a Random Forest Classifier using both QH and clinical features. These results suggest the use of independent and synergistic QH features which can be used in conjunction with clinical features for improved cancer prediction.

5.7.2 Comparison with Gleason Scoring and Tumor Stage

We compared the classification accuracy of COrE and previously investigated QH methods (Morphological and Architectural features) against Gleason Sum and Tumor Stage. Kaplan-Meier survival curves (Figure 5.2) illustrate the difference in the BCR-free survival outcomes in the predicted high-risk and low risk BCR groups determined via 100 runs of 3-fold Random Forest classifier. Patients predicted to have BCR were placed in the high-risk group, while patients predicted as NR were placed in the low-risk group. Log-rank test determines the significance of the difference between high-risk and low-risk curves (lower p-values represent greater differences in BCR-free survival).

Figure 5.3 demonstrates the successive contribution of these non-overlapping features to offer better prediction of BCR when combined together. Gleason sum and tumor stage provide lower AUC compared to QH features. However, by integrating Gleason sum and tumor stage to the QH features, better classification AUC can be achieved, suggesting that these features capture information that is not present in

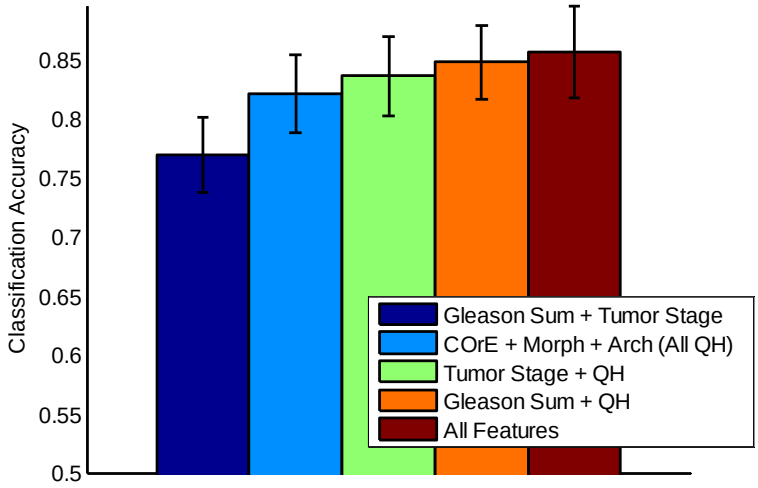


Figure 5.3: Our results suggest improvement in classification performance via QH measures over traditional clinical features (Gleason Sum and Tumor Stage). Furthermore, combining QH and clinical features is shown to yield better classification than each feature type individually.

Gleason sum and tumor staging. This is supported via the Kaplan-Meier curve corresponding to Figure 5.2, showed the largest separation of BCR and NR shown via a p-value of 0.0006.

5.8 Summary

In this work, we presented a new feature descriptor, cell orientation entropy (COrE), for quantitative measurement of local disorder in nuclear orientations in digital pathology images. We demonstrated high accuracy and improvement in predicting BCR in 39 CaP TMAs via the use of COrE features. While COrE features did not outperform other quantitative histomorphometric measurements such as nuclear shape and architecture significantly, the combination of nuclear shape, architectural and COrE features boosted classifier accuracy in identifying patients at risk for BCR following radical prostatectomy. More significantly, the combination of COrE and other image based features significantly outperformed pathologist derived GS, which is 50% for GS 7, and is further known to have at best moderate inter-observer agreement (κ = 0.47-0.7) [22].

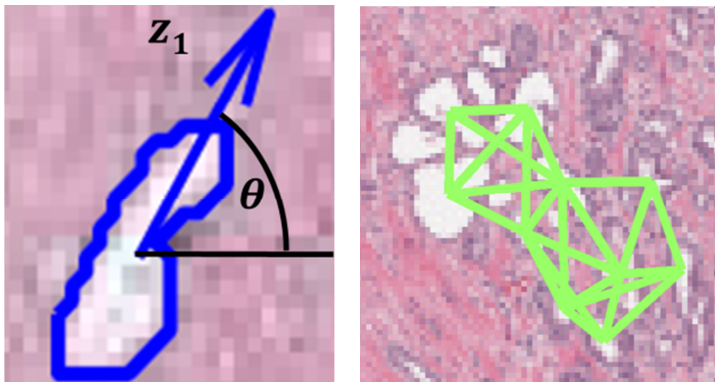
5.9 Co-occurring Gland Tensors in Localized Subgraphs: Quantitative Histomorphometry for Postoperative Prediction of Biochemical Recurrence in Prostate Cancer Patients with Intermediate-Risk Gleason Scores

In the following sections, we present a new quantitative histomorphometric attribute, co-occurring gland tensors (CGT), that aims to capture the directional information in localized gland networks to characterize differences in gland orientation between (a) malignant and benign regions and (b) CaP patients who do and do not experience biochemical recurrence following RP, from excised histopathology sections. We briefly summarize our methodology as follows.

For CGTs, a segmentation algorithm is first employed to individually segment gland boundaries from digitized pathology sections. To each gland, we ascribe a tensor that reflects the dominant orientation of the gland based off the major axis as shown in Figure 5.4(a). A subgraph is then constructed to link together glands proximal to each other into a gland network as illustrated in Figure 5.4(b). The subgraphs, unlike the graphs for Voronoi, Delaunay and minimum spanning trees that have been previously used to characterize global glandular architecture [8], allows for characterization of local gland arrangement. Use of local subgraphs prevent graph edges from traversing heterogeneous tissue regions such as stroma and epithelium.

The co-occurrence matrix, previously used to characterize image intensity textures, is used to capture second-order statistics of gland orientations within each gland network in the image. Hence each co-occurrence matrix captures the frequency with which orientations of two glands proximal to each other co-occur. Co-occurrence features such as entropy are extracted from the co-occurrence matrix associated with each gland network and captures the degree to which orientations are similar or divergent to each other. Hence a neighborhood with a high entropy value would reflect a high degree of disorder among gland orientations while a low entropy value reflects that the gland tensors appear to be aligned roughly in the same direction.

Given that we expect to see glandular tensor disorder in (a) malignant versus benign



(a) (b)

Figure 5.4: (a) Angular calculation of the gland tensor converts z_1 to an angle between 0° to 180°. (b) Subgraphs connect the centroids of neighboring glands into gland networks.

regions and (b) biochemical recurrence cases versus non-recurrence cases, second-order statistical tensor features like entropy represent a novel, reproducible, and interpretable way to characterize disease appearance on histopathology. Unlike first order statistics of tensors, the co-occurring gland tensor features are able to implicitly capture the cyclical properties of gland orientation. The use of local subgraphs generated by a probabilistic decaying function help define local gland networks within which the CGT features can be extracted and analyzed.

In this work, we demonstrate the utility of CGT features on the following classification tasks:

1. Differentiating cancerous and non-cancerous prostate tissue regions, and
2. Distinguishing CaP patients with and without biochemical recurrence following radical prostatectomy.

The remainder of this chapter is structured as follows. We introduce the methodology for Co-occurring Gland Tensors in Localized Subgraphs (CGTs) in Section 5.10. In Section 5.11, we explain the Experimental Design, outlining datasets, workflow, and comparative methodologies used in the study. Experiments and Results are explained in Section 5.12 followed by Concluding Remarks in Section 5.13.

5.10 Quantitative Histomorphometry via the method of Co-occurring Gland Tensors (CGTs)

5.10.1 Notation

For this work, we are interested in predicting biochemical recurrence in prostate cancer patients via the analysis of histological slices of the excised prostate following radical prostatectomy. We define a patient $\mathcal{P}_i$, $i \in \{1, 2, \ldots, n_p\}$, where n_p is the number of patients in the study cohort. To predict BCR on a CaP patient $\mathcal{P}_i$, we analyze a digitized histology image $\mathcal{R}_i$, $i \in \{1, 2, \ldots, n_r\}$. $\mathcal{R}_i$ is composed of pixels $\dot{c}$ from which features can be extracted to represent a patient $\mathcal{P}_i$ or region $\mathcal{R}_i$. We define features as $f = [f_1, f_2, \ldots, f_{n_f}]$, where n_f is the number of features extracted from each $\mathcal{R}_i$ and n_r is the number of images investigated. f extracted from $\mathcal{R}_i$ is used to train a classifier $\mathcal{C}^{CGT}$, which is used to predict the outcome $\ell(\mathcal{P}_i) \in \{-1, +1\}$ of patient $\mathcal{P}_i$, where -1 denotes a patient $\mathcal{P}_i$ whom did not experience BCR and +1 denotes that $\mathcal{P}_i$ did experience BCR. Alternatively, $\mathcal{C}^{CGT}$ can be trained for other prediction tasks distinguishing between cancerous regions from normal regions $\mathcal{R}_i$.

5.10.2 Calculating Gland Tensors

To determine the directionality for each gland γ_i, we perform principal component analysis [105] on a set of boundary points γ_i^b to obtain the principal components $Z = [z_1, z_2]$. The first principal component z_1 describes the directionality of the gland in the form of the major axis $z_1 =< z_1^x, z_1^y >$, along which the greatest variance occurs in the glandular boundary. The principal axis z_1 represents a 1st order tensor, which is converted to an angle $\bar{\theta}(\gamma_i) \in [0°180°]$ calculated counterclockwise from the vector $< 1, 0 >$ by $\bar{\theta}(\gamma_i) = \frac{180°}{\pi} \arctan(\frac{z_1^y}{z_1^x})$. A depiction of the calculated angles is shown in Figure 5.4(a).

5.10.3 Defining Local Gland Subgraphs

Pairwise spatial relationships between glands are defined via sparsified graphs. A graph $G = \{V, E\}$, where V represents the set of n_g gland centroids $\gamma_i^c, \gamma_j^c \in V$,

$i, j \in \{1, 2, \ldots, n_g\}$ as nodes, and E represents the set of edges which connect them. The edges between all pairs of $\{\gamma_i^c, \gamma_j^c\}$ are determined via a probabilistic decaying function

$$E = \{(i,j) : r < d(i,j)^{-\alpha}, \forall \gamma_i^c, \gamma_j^c \in V\}, \tag{5.4}$$

where $d(i,j)$ represents the Euclidean distance between γ_i^c and γ_j^c. $\alpha \geq 0$ controls the density of the graph, where α approaching 0 represents a high probability of connecting nodes while α approaching ∞ represents a low probability. $r \in [0,1]$ is an empirically determined edge threshold. An example of a resulting glandular subgraph network is shown in Figure 5.4(b).

5.10.4 Constructing Tensor Co-occurrence Matrices

The objects of interest for calculating CGT features are the gland tensors given by a discretization of the angles $\bar{\theta}(\gamma_i)$, such that $\theta(\gamma_i) = \omega \times \text{ceil}(\frac{\bar{\theta}}{\omega})$, where ω is a discretization factor. Neighbors defined by the local gland subgraphs G, allow us to define neighborhoods for each gland. For each $\gamma_i^c \in V$, we define a neighborhood $\mathcal{N}_i$, to include all $\gamma_j^c \in V$ where a path between γ_i^c and γ_j^c exists via E in the graph G.

An $N \times N$ tensor co-occurrence matrix $\mathcal{M}$ subsequently captures gland tensor pairs which co-occur within each neighborhood $\mathcal{N}_i$, such that for each $\mathcal{N}_i$,

$$\mathcal{M}_{\mathcal{N}_i}(a,b) = \sum_{\gamma_i, \gamma_j}^{\mathcal{N}_i} \sum_{a,b=1}^{N} \begin{cases} 1, & \text{if } \theta(\gamma_i)=a \text{ and } \theta(\gamma_j)=b \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

where $N = \frac{180}{\omega}$, the number of discrete angular bins. An example of a tensor co-occurrence matrix is shown in Figures 5.5(d) and (l)

5.10.5 Calculating Second Order Statistics

We subsequently extract second order statistical features Θ (Contrast energy, Contrast inverse moment, Contrast average, Contrast variance, Contrast entropy, Intensity average, Intensity variance, Intensity entropy, Entropy, Energy, Correlation, and two measures of information) from each tensor co-occurrence matrix $\mathcal{M}_{\mathcal{N}_i}(a,b)$. Selected

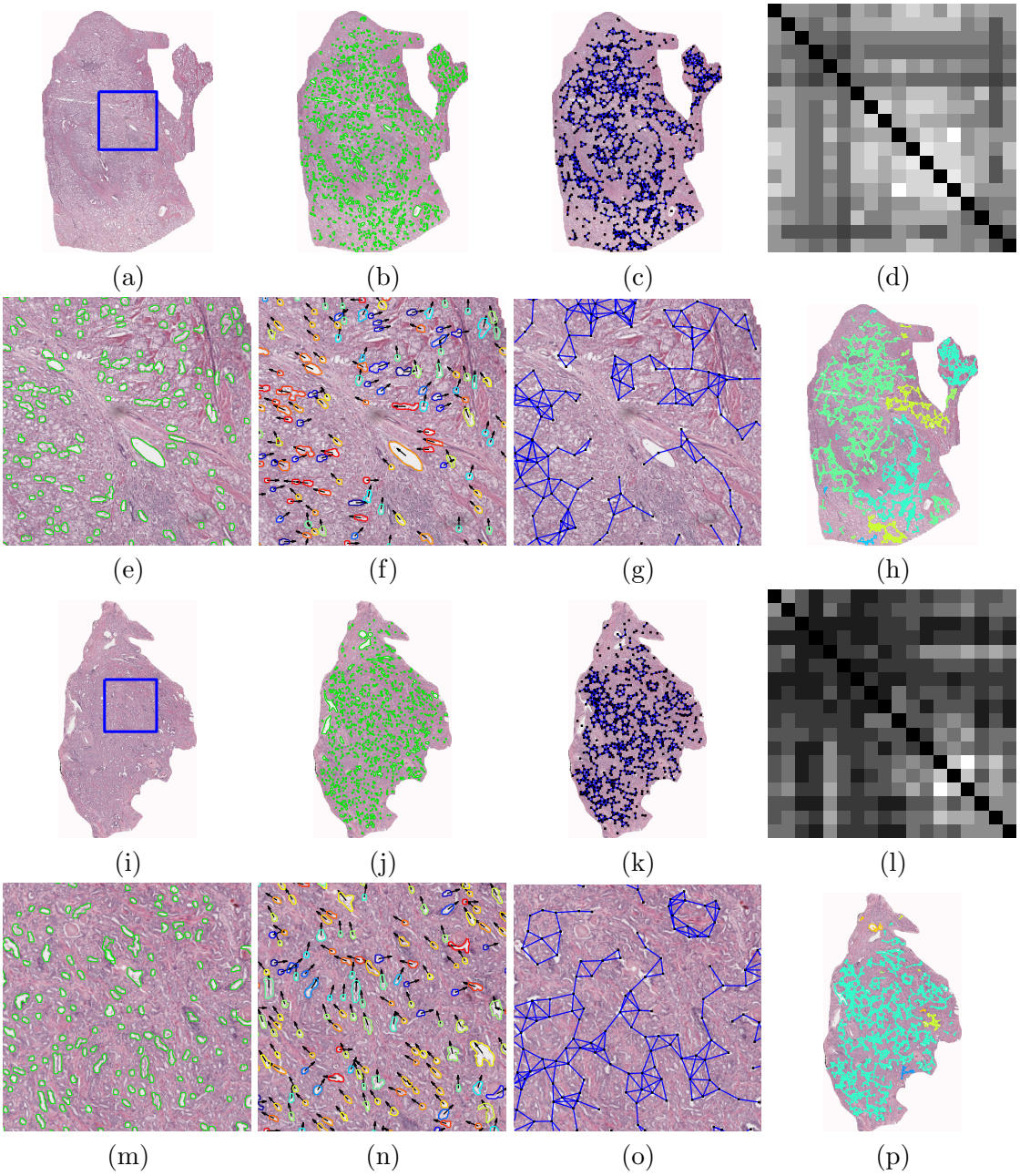


Figure 5.5: (a) and (i) show annotated histological CaP regions pertaining to a BCR (a)-(h) and a NR (i)-(p) case study, respectively. (b,j) Automated gland segmentation of gland boundaries. (c,k) Subgraphs connect neighboring glands. An enlarged view of the boxed region in (a) and (i) respectively, illustrates (e,m) segmented glands, (f,n) gland tensors, and (g,o) gland network subgraphs. (f,n) Arrows denote the directionality of each gland. Boundary colors (blue to red) correspond to angles $\theta \in [0^\circ\ 180^\circ]$. (g,o) Localized gland networks define the region of each tensor co-occurrence matrix. (d,l) Summed tensor co-occurrence matrices denote the frequency in which two glands of two directionalities co-occur across all neighborhoods (white is greater co-occurrence). Diagonal co-occurrence values omitted to provide better contrast in the off-diagonal components. (h,p) Colormap of the gland subgraphs correspond to the intensity average in each neighborhood.

Table 5.4: Representative CGT features

CGT Feature (Θ)	**Description**
Entropy	$\sum_{a,b} -\mathcal{M}(a,b)\ \log(\mathcal{M}(a,b)))$
Energy	$\sum_{a,b} \mathcal{M}(a,b)^2$
Correlation	$\sum_{a,b} \frac{(a-\mu a)(b-\mu b)\mathcal{M}(a,b)}{\sigma_a \sigma_b}$
Contrast (variance)	$\sum_{a,b} \lvert a-b\rvert^2 \mathcal{M}(a,b)$

formulations for Θ are described in Table 5.4 to characterize information from each gland network $\mathcal{N}_i$ and a visualization of the mean intensity measure of each $\mathcal{N}_i$ on histology is shown in Figure 5.5.

5.10.6 Differentiation of BCR and NR cases via CGT

Figure 5.5 shows two representative studies: a BCR and NR case. For the BCR case, we can see greater disorder in the gland orientation via the tensor plot in Figure 5.5(f). The tensor-based colormap for BCR characterizes the disorder in BCR cases, as the glands appear as a large spectrum of colors, denoting different directionalities. Conversely, for the NR case, (Figure 5.5(n)), the colormap is more consistent, suggesting less variance in the gland directionality.

We can confirm these differences in Figures 5.5(d,l) via the contrasted tensor co-occurrence matrix. The brightness of the off-diagonal elements of the matrix demonstrate greater co-occurrences of differentially oriented gland tensors for the BCR case (Figure 5.5(d)) compared to the NR case (Figure 5.5(l)).

These differences in the tensor co-occurrence matrices are detected by the second order statistics, as Figure 5.5(h,p) demonstrates different color patterns based on the value of the statistics in each subgraph. Figure 5.5(h) shows demonstrates a higher mean intensity value reflected by the brighter co-occurrence matrix, while Figure 5.5(p) shows a lower blue color pattern across the glands.

It can be observed in Figures 5.5(c,k) that subgraphs capture local gland neighborhoods, by eliminating noise from stromal areas which separate the glandular areas. Furthermore, CGTs introduce more biological information (gland tensors) compared to texture features, which only focus on grayscale pixel intensity.

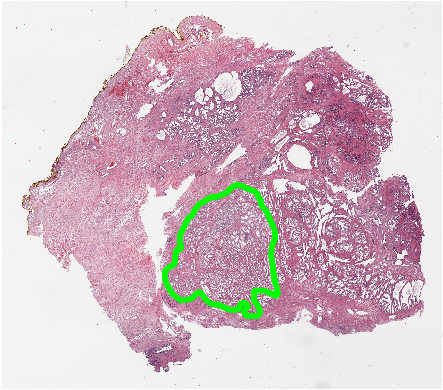


Figure 5.6: Annotation of a region of interest (shown in green) on prostate histopathology is performed by a pathologist. QH analysis is performed only in these regions.

Table 5.5: Overview of Clinical Datasets

Clinical Variables	$\mathcal{P}^A$ **Cohort A** ($n_p = 20$)	$\mathcal{P}^B$ **Cohort B** ($n_p = 20$)	$\mathcal{P}$ **Total Cohort** ($n_p = 40$)
Pathological Gleason Score			
3+3	4 (20%)	1 (5%)	5 (12.5%)
3+4	7 (35%)	17 (85%)	24 (60%)
4+3	7 (35%)	2 (10%)	9 (22.5%)
3+5	1 (5%)	- (-)	1 (2.5%)
4+4	1 (5%)	- (-)	1 (2.5%)
Pathologic Stage			
pT2	8 (40%)	12 (60%)	20 (50%)
pT3a	9 (45%)	6 (30%)	15 (37.5%)
pT3b	3 (15%)	2 (10%)	5 (12.5%)

5.11 Experimental Design

5.11.1 Data Acquisition and Data Description

The datasets (obtained from the Hospital at the University of Pennsylvania) were comprised of 40 CaP patients who had undergone RP treatment, selected for identified as having Gleason scores 6-8 and pathologic stage pT2-pT3. Of these patients, 20 were diagnosed with BCR (BCR) within 5 years of RP, and 20 had no recurrence (NR).

The data was collected from two independent sources: Cohort A ($\mathcal{P}^A$) from the department of Pathology and Laboratory Medicine at the University of Pennsylvania ($n = 20$) and Cohort B ($\mathcal{P}^B$) from the department of Clinical Epidemiology and Biostatistics at the University of Pennsylvania ($n = 20$). A further breakdown of the data is summarized in Table 5.5.

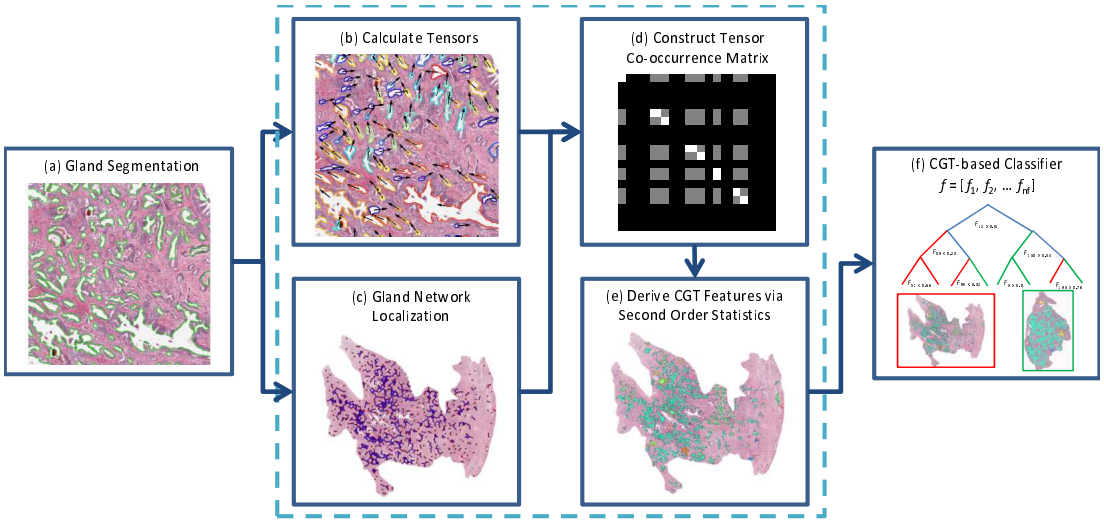


Figure 5.7: Workflow for building a CGT-based classifier. (a) Gland segmentation is performed on a region of interest. CGT methodology (highlighted within the dashed lines) leverages the gland segmentation to compute CGT features. (b) Tensor calculation and (c) subgraph computation is performed on the segmented image. (d) tensor co-occurrence matrix aggregates co-occurring gland tensors in localized gland networks. (e) mean, standard deviation and range of second order statistics (shown as different colored gland networks) create a set of CGT features for the region. (f) A CGT-based classifier can then be built using the features obtained from (e). Alternatively, another QH-based classifier can be built via the extraction of a different set of QH features.

For all patients, following RP, the excised prostate was sliced, stained with hematoxylin and eosin (H&E), and digitized at a resolution of 0.5 μm per pixel or 20x magnification using an Aperio® Whole Slide scanner. For each digitized image, CaP regions were annotated by a pathologist as shown in Figure 5.6. 56 cancer regions were annotated across 40 patients, 28 from BCR patients and 28 from NR patients. 24 regions pertaining to non-cancerous regions from 20 CaP patients in Cohort B were annotated as controls.

5.11.2 CGT Extraction Workflow

The CGT feature extraction workflow provides a template for building the classifier $\mathcal{C}^{CGT}$ to predict on a patient $\mathcal{P}_i$ or region $\mathcal{R}_i$. The workflow begins with the identification and segmentation of glandular boundaries from the image show in Section 5.11.2. Next is the calculation of features f from the segmentations. Finally, a classifier $\mathcal{C}^{CGT}$ is constructed from CGT features f.

Identification of Glandular Boundaries

An automatic region-growing based prostate gland segmentation algorithm [170] is used to detect and segment glandular boundaries on the histological image as illustrated in Figure 5.8. Segmentation is performed using the luminance channel in CIELAB color space. Using the luminance channel, gland lumens appear as contiguous, high intensity pixel regions bound by sharp, well-defined edges. To identify glands, the luminance image is convolved with a Gaussian kernel at multiple scales $\sigma_g \in \{0.025, 0.05, 0.1, 0.2\}$ mm to account for multiple gland sizes. The peaks of the resulting smoothed luminance images are used as seeds for a region growing procedure briefly outlined below.

1. A $12\sigma_g \times 12\sigma_g$ bounding box is initialized around each initial seed pixel, which represents the current region (CR), with 8-connected pixels surrounding CR, denoted as the current boundary (CB).

2. Next, the pixel in CB with the highest intensity is removed from CB and incorporated into CR. The 8 surrounding pixels of this new CR pixel, which are not

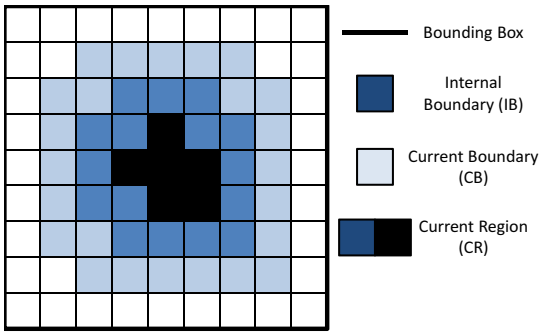


Figure 5.8: Schematic for region growing.

already in CR, are incorporated into CB.

3. The boundary strength is identified at each iteration as shown in Figure 5.8. We define the internal boundary (IB) as all CR pixels adjacent to CB. Boundary strength is defined as the mean intensity of the pixels in IB minus the mean intensity of the pixels in CB.

4. Steps 2 and 3 are repeated until the algorithm attempts to add a pixel outside the bounding box.

5. The optimal region is defined as region CR at the iteration where maximum boundary strength was achieved.

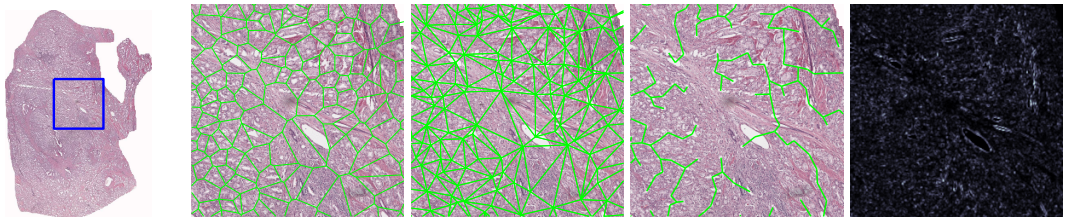
Overlapping regions are subsequently resolved by removing the region with the lowest boundary strength. An example of our results can be seen in Figure 5.7(b).

CGT Feature Extraction

Based on the segmentation in Section 5.11.2, CGT features are calculated as described in Section 5.10. Mean, standard deviation, and range of Θ across all $\mathcal{N}_i$ constitute the set of 39 CGT features $f \in \{f_1, f_2, \ldots, f_{39}\}$. A list of other potential QH features f which can be extracted either directly from the image region or segmentation of the image region is discussed in Section 5.11.3.

Table 5.6: Summary of Comparative Quantitative Histomorphometric (QH) features

Feature Type (QH)	**Description**	n_f
Gland Morphology (M)	Area Ratio, Distance Ratio, Standard Deviation of Distance, Variance of Distance, Distance Ratio, Perimeter Ratio, Smoothness, Invariant Moment 1-7, Fractal Dimension, Fourier Descriptor 1-10 (Mean, Std. Dev, Median, Min / Max of each)	100
Voronoi Diagram (V)	Polygon area, perimeter, chord length: mean, std. dev., min/max ratio, disorder	12
Delaunay Triangulation (D)	Triangle side length, area: mean, std. dev., min/max ratio, disorder	8
Minimum Spanning Tree (MST)	Edge length: mean, std. dev., min/max ratio, disorder	4
Glandular Density (GD)	Density of glands, distance to nearest gland	27
Co-occurrence Texture (T)	Contrast energy, Contrast inverse moment, Contrast average, Contrast variance, Contrast entropy, Intensity average, Intensity variance, Intensity entropy, Entropy, Energy, Correlation, 2 measures of information: mean, std. dev.	26



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

Figure 5.9: (a) QH features are extracted from an annotated region on a digitized prostate histology slide following radical prostatectomy. Quantitative histopathology feature extraction is performed on (a) the annotated region. Graphs for (b) Voronoi, (c) Delaunay, and (d) Minimum Spanning Trees as well as (e) a texture image feature are shown from the area denoted by a blue box in (a).

Building a CGT-based classifier

A subset of $\mathcal{P}$ or $\mathcal{R}$ described by f is used for training a Random Forest classifier $\mathcal{C}^{CGT}$. This process is described in Section 5.11.4 and can be used to create other classifiers $\mathcal{C}^{QH}$ from other QH features f.

5.11.3 Comparative Methodologies

Quantitative Histomorphometry ($\mathcal{C}^{QH}$)

We provide a brief summary of the comparative QH methodologies listed in Table 5.6.

Gland Morphology (M): Morphological descriptors [146] are extracted from the segmented glandular boundaries obtained in Section 5.11.2. Statistics such as the area ratio, perimeter ratio, and distance ratio are derived from the gland boundary information and the mean, standard deviation, median, and minimum to maximum ratio are calculated across all glands. These features are summarized in Table 5.6.

Voronoi Diagram (V): Glandular architecture [146] can be characterized via the construction of graphs $G = \{V, E\}$, where V represents the set of vertices and E represents the set of edges which connect them. Voronoi diagrams divide each $\mathcal{R}_i$ into non-overlapping polygons, each associated with a gland γ_i, where each edge bisects 2 neighboring gland centroids γ_i^c and γ_j^c, where $i, j \in \{1, 2, \ldots, n_g\}$ and n_g is the number of glands in image $\mathcal{R}_i$. An example is shown in Figure 5.9(b). Statistics such as the area, perimeter and chord length are recorded for each polygon and the average, standard deviation, disorder, and minimum to maximum ratio are calculated across all polygons in the image.

Delaunay Triangulation (D): Delaunay triangulation divides the image $\mathcal{R}_i$ into triangles whose edges connect the gland centroids γ_i^c and γ_j^c as vertices. An example is shown in Figure 5.9(c). Delaunay triangulation is related to the Voronoi diagram in that for each polygon in the Voronoi diagram, there is an accompanying edge which connects its gland centroid γ_i^c with γ_j^c of an adjacent polygon. Edge length and area are computed for each triangle and the mean, standard deviation, minimum to maximum ratio, and disorder statistics are calculated across all triangles.

Minimum Spanning Tree (MST): Minimum Spanning Trees (MST) are another graph representation where all gland centroids γ_i^c are connected with a minimum total edge weight defined as $\underset{E}{\text{argmin}} \sum_{i,j} E_{ij} \times d(i,j)$, where $d(i,j)$ is the Euclidean distance between γ_i^c and γ_j^c. An example is shown in Figure 5.9(d). The average, standard deviation, disorder, and minimum to maximum ratio statistics are calculated across all edges in the graph.

Gland Density (GD): Gland density features encompass two types of features. The first set of features denotes the number of γ_j^c which lie within a 10, 20, 30, 40, and 50 pixel radius of each γ_i^c. The second set of feature denotes the distance between each γ_i^c and its 3, 5, and 7 nearest centroids γ_j^c. The average, standard deviation, and disorder of each of these features are computed across all glands.

Image Co-occurrence Textures (T): Second order co-occurrence features [161] are calculated from a symmetric co-occurrence matrix which aggregates the frequency in which two pixel intensities co-occur within a pre-determined window distance around each pixel $\dot{c}$. The size of the co-occurrence matrix is determined by the maximum possible intensity value in the image, which for 8-bit images is $2^8 = 256$. A window distance of 1 pixel was chosen. For each $\dot{c}$, contrast energy, contrast inverse moment, contrast average, contrast variance, contrast entropy, intensity average, intensity variance, intensity entropy, entropy, energy, correlation, and two information measures are computed from the co-occurrence matrix. The mean and standard deviation across all $\dot{c}$ are used to build a single set of texture features f for each image $\mathcal{R}_i$.

Prostate Cancer Prediction Tools ($\mathcal{C}^{PT}$)

We provide a brief summary of the comparative CaP prediction tools listed in Table 5.7.

Kattan Nomogram (K): The Kattan nomogram [23] was one of the earliest prostate cancer postoperative prediction tools to be developed for predicting biochemical failure following radical prostatectomy. Clinical predictors for the Kattan nomogram include 1) Pre-operative PSA, 2) Gleason Sum, 3) Primary Gleason score, 4) Surgical Margins, 5) Prostate Capsular Invasion, 6) Seminal Vesicle Invasion (SVI), and 7) Lymph Node Involvement. A raw score s, $0 \leq s \leq 300$ (higher score pertains to higher risk of BCR) is derived from these predictors and risk for each $\mathcal{P}_i$ is assessed in terms of an 84 month BCR-free probability.

Table 5.7: Summary of Postoperative CaP Prediction Tools

Prediction Tool (PT)	**Clinical Variables**
Kattan (K)	1) Pre-operative PSA, 2) Gleason Sum, 3) Primary Gleason score, 4) Surgical Margins, 5) Prostate Capsular Invasion, 6) Seminal Vesicle Invasion (SVI), and 7) Lymph Node Involvement
Stephenson (S)	1) Year of Radical Prostatectomy, 2) Surgical Margins, 3) Extraprostatic Extension, 4) Seminal Vesicle Invasion (SVI), 5) Lymph Node Involvement, 6) Primary Gleason, 7) Secondary Gleason, and 8) Pre-operative PSA
UCSF-CAPRA ($CAPRA$)	1) Age, 2) Pre-operative PSA, 3) Primary Gleason, 4) Secondary Gleason, 5) Tumor Stage, and 6) Percent Positive Biopsy Cores
MS-KCC (MSK)	1) Pre-Treatment PSA, 2) Age, 3) Primary Gleason Grade, 4) Secondary Gleason Grade, 5) Gleason Sum, 6) Year of Prostatectomy, 7) Months Free of Cancer, 8) Surgical Margins, 9) Extra Capsular Extension, 10) Seminal Vesicle Involvement, and 11) Lymph Node Involvement

Stephenson Nomogram (S): The Stephenson nomogram [24] was developed along with Michael Kattan to incorporate the year of surgical intervention for predicting BCR. Based on 1) Year of Radical Prostatectomy, 2) Surgical Margins, 3) Extraprostatic Extension (EPE), 4) Seminal Vesicle Invasion (SVI), 5) Lymph Node Involvement, 6) Primary Gleason, 7) Secondary Gleason Scores, and 8) Pre-operative PSA. A raw score s, $0 \leq s \leq 240$, (higher score pertains to higher risk of BCR) is derived from these clinical features. Risk for each $\mathcal{P}_i$ is assessed in terms of an 80 month BCR-free probability based on the raw score.

Cancer of the Prostate Risk Assessment (UCSF-CAPRA) ($CAPRA$): The University of California in San Francisco (UCSF) developed a risk assessment tool for predicting BCR in CaP patients following radical prostatectomy. Their scoring system, Cancer of the Prostate Risk Assessment test (CAPRA) [26, 27], is based on overall score $s \in \{0, 1, \ldots, 10\}$, where 10 represents the highest risk of BCR. Clinical predictors for CAPRA include 1) Age, 2) Pre-operative PSA, 3) Primary Gleason, 4) Secondary Gleason scores, 5) Tumor Stage, and 6) Percent Positive Biopsy Cores.

Memorial Sloan-Kettering Cancer Center (MS-KCC) Nomogram (MSK): One of the

most popular nomograms with contributions from Kattan and Stephenson is the MS-KCC nomogram [23,24]. The MS-KCC nomogram incorporates 1) Pre-Treatment PSA, 2) Age, 3) Primary Gleason Grade, 4) Secondary Gleason Grade, 5) Gleason Sum, 6) Year of Prostatectomy, 7) Months Free of Cancer, 8) Surgical Margins, 9) Extra Capsular Extension, 10) Seminal Vesicle Involvement, and 11) Lymph Node Involvement. Risk score s, $0 \leq s \leq 1$, for each $\mathcal{P}_i$, is assessed in terms of its 10-year BCR-free probabilities.

5.11.4 Evaluation Measures

Random Forest Classifier

For all experiments, Random Forests (Bagged Decision Tree classifiers) [180] were used to train a predictor $\mathcal{C}$ from each feature set f or score s. The resulting $\mathcal{C}$ is subsequently used to classify each patient $\mathcal{P}_i$ as either $\ell(\mathcal{P}_i) = +1$ or $\ell(\mathcal{P}_i) = -1$ or each region as either $\ell(\mathcal{R}_i) = +1$ or $\ell(\mathcal{R}_i) = -1$, where +1 and -1 refer to the positive and negative classes for each classification task, further described in Section 5.12.

Classification Accuracy

The predictive value of each classifier is evaluated via classification accuracy ϕ^{Acc} and via the area under the receiver operating characteristic curve (AUC) ϕ^{AUC}. If the true label value of $\ell(\mathcal{R}_i)$ is of the positive class $\ell(\mathcal{R}_i) = +1$ and the classifier predicts correctly, this result is a true positive (TP) classification. If the classifier predicts this result incorrectly, the result is a true negative (FP). Similarly, if $\ell(\mathcal{R}_i) = -1$ and the classifier predicts correctly, the result is a true negative (TN). Otherwise, the result is a false negative (FN). Classification Accuracy (ϕ^{Acc}) measures the ability of a classifier to predict on a new set of testing data provided by the features f or scores s and is calculated as

$$\phi^{Acc} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{5.6}$$

Receiver Operating Characteristic

ϕ^{AUC} represents an overall measure of a classifier's predictive value, independent of the decision threshold, based on the receiver operative characteristic (ROC) curve. The ROC curve is constructed by computing $1 - \phi^{Spec}$ and ϕ^{Sens} at each decision threshold where ϕ^{Spec} and ϕ^{Sens} are defined as

$$\phi^{Sens} = \frac{TP}{TP + FN}, \tag{5.7}$$

and

$$\phi^{Spec} = \frac{TN}{TN + FP}. \tag{5.8}$$

The AUC represents the area under the ROC curve, where an AUC of 1 pertains to a perfect classifier, and an AUC of 0.5 pertains to a classifier which is no better than random guessing.

The Wilcoxon Rank Sum Test [181] was subsequently utilized to determine statistical significance between classifications for each $\mathcal{C}$ compared to $\mathcal{C}^{CGT}$.

Kaplan-Meier Analysis

Kaplan-Meier analysis [182] is used to compare the BCR-free survival time between two groups. In this study, the two groups are determined by a predictor $\mathcal{C}$, which predicts that a patient $\ell(\mathcal{P}_i) \in \{-1, +1\}$, will either experience biochemical recurrence (BCR) or not (NR). When plotted onto time versus BCR-free survival rate, the BCR free survival rate of the group will decrease at the time when each $\mathcal{P}_i$ develops BCR. Thus, we expect the curve for the set of patients $\mathcal{P}$ predicted to have BCR to drop quickly while the set of patients predicted to have the label NR to remain BCR-free with no drop in the curve. The quantitative difference between the survival outcome can be determined via the logrank test [183]. The non-parametric test yields a p-value, where lower p-values denote greater significance between the survival distributions.

5.12 Experimental Results

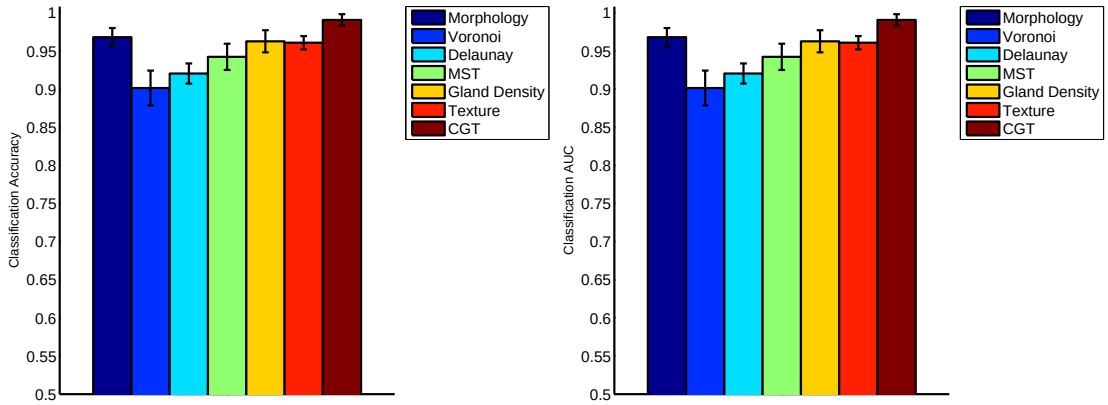
We compared the ability of predictors built using CGT features $\mathcal{C}^{CGT}$ against other predictors $\mathcal{C}^{QH}$ and $\mathcal{C}^{PT}$ to differentiate different categories of CaP patients $\mathcal{P}$ and regions $\mathcal{R}$ via the following experiments:

1. Identifying Cancerous (+1) versus Non-Cancerous (-1) Regions ($\mathcal{R}$)
2. Distinguishing Biochemical Recurrence (+1) and Non-recurrence (-1) using Cancerous Regions ($\mathcal{R}$)
3. Predicting Biochemical Recurrence versus Non-Recurrence in CaP Patients following RP ($\mathcal{P}$)
4. Comparing BCR prediction of Patients CaP patients following RP ($\mathcal{P}_B$) via CGTs versus CaP Prediction Tools
5. Receiver Operating Characteristics (ROC) analysis on comparing BCR prediction of CaP patients following RP ($\mathcal{P}_B$) via CGTs versus CaP Prediction Tools
6. Kaplan-Meier Analysis comparing BCR prediction of CaP patients following RP ($\mathcal{P}_B$) via CGTs versus CaP Prediction Tools

Each experiment and accompanying results are described in detail below.

5.12.1 Experiment 1: Identifying Cancerous versus Non-Cancerous Regions $\mathcal{R}$

Design: 80 regions $\mathcal{R}_i$ were annotated by expert pathologists pertaining to 56 cancerous regions and 24 non-cancer regions. All features f were extracted across the entire annotated region. We compare the efficacy of CGT features with previously studied QH features for the purpose of differentiating cancerous regions (+1) from non-cancerous regions (-1) on prostate histology. Comparison between QH features is done by creating classifiers $\mathcal{C}^{QH}$ from each set of QH features, where $QH \in \{M, V, D, MST, GD, T, CGT\}$ For evaluation, 100 iterations of randomized 3-fold cross validation was performed using



(a) (b)

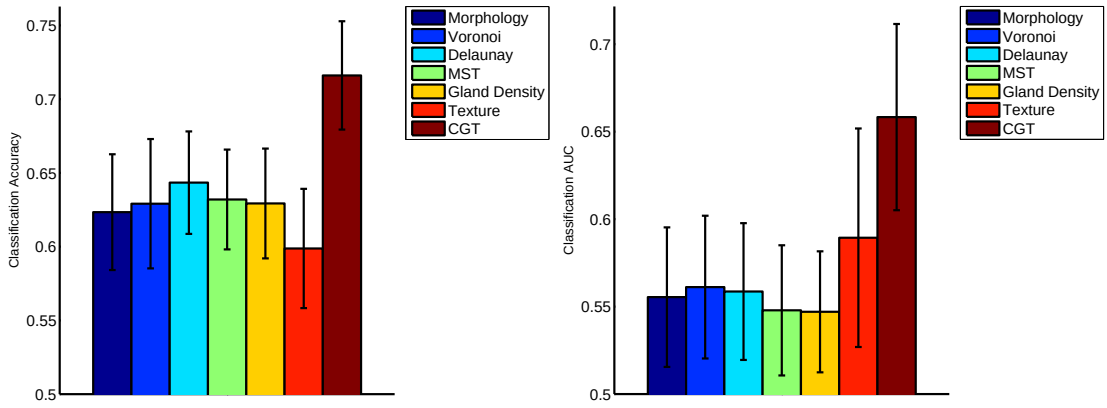
Figure 5.10: Comparison of $\mathcal{C}^{QH}$ in terms of mean (a) ϕ^{Acc} and (b) ϕ^{AUC} (with error bounds denoting standard deviation) for distinguishing cancerous and non-cancerous tissue in 80 regions $\mathcal{R}_i$ over 100 runs of Random Forest with randomized 3-fold cross validation

a Random Forest classifier to classify each $\mathcal{R}_i$ as either Cancerous or Non-cancerous for each set of QH features f.

Results: As shown in Figure 5.10, CGT features improve in distinguishing cancer versus non-cancer regions $\mathcal{R}$ compared to 6 other QH features (previously described). Mean and standard deviation for ϕ^{Acc} and ϕ^{AUC} are shown in Table 5.8. CGT shows statistically significant ($p < 0.05$) improvement in terms of ϕ^{Acc} and ϕ^{AUC} compared to all QH features as shown in Table 5.8. Our results suggest CGTs as an improvement over existing QH methodologies for differentiating cancerous regions from non-cancer regions on CaP histology.

5.12.2 Experiment 2: Identifying Regions $\mathcal{R}$ associated with Biochemical Recurrence

Design: 56 cancer regions $\mathcal{R}_i$ were annotated, 28 from BCR patients and 28 from NR patients. All features f were extracted across the entire annotated cancer region. We compare the efficacy of CGT features with previously studied QH features (Table 5.6) for the purpose of differentiating cancerous regions $\mathcal{R}$ extracted from patients who will develop BCR (+1) from patients who will not (-1), following RP. Classifiers $\mathcal{C}^{QH}$, where

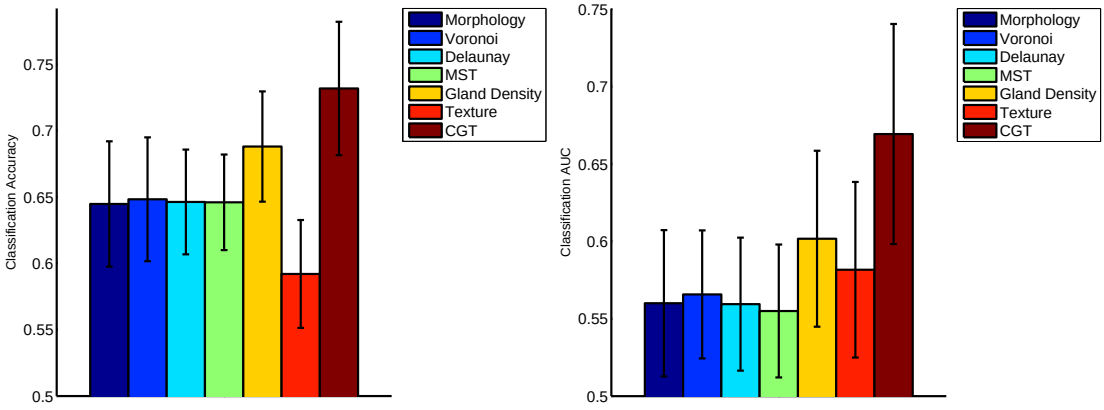


(a) (b)

Figure 5.11: Comparison of $\mathcal{C}^{QH}$ in terms of mean (a) ϕ^{Acc} and (b) ϕ^{AUC} (with error bounds denoting standard deviation) for identifying BCR from 56 cancer regions corresponding to 40 patients over 100 runs of Random Forest with randomized 3-fold cross validation

$QH \in \{M, V, D, MST, GD, T, CGT\}$ were built for each QH feature. To identify regions associated with BCR, patient controlled classification was performed such that multiple regions $\mathcal{R}_i$ from a single patient $\mathcal{P}_i$ were either all in the training set or all in the testing set for all classifications. For evaluation, 100 iterations of randomized 3-fold cross validation was performed using Random Forest to classify each $\mathcal{R}_i$ as either belonging to BCR or NR patients.

Results: In Figure 5.11, CGT is shown to outperform 6 QH features for the task of identifying regions $\mathcal{R}$ belonging to BCR patients for both ϕ^{Acc} and ϕ^{AUC}. Statistically significant ($p < 0.05$) improvement in both ϕ^{Acc} and ϕ^{AUC} was shown for CGT over all QH features as shown in Table 5.9. As demonstrated in Table 5.9, many of the QH features perform only slightly better than guessing, with AUC values near 0.5.This is not surprising given that many of these QH features are modeled after characteristics of Gleason grade, which is often unable to distinguish BCR in CaP patients with GS 6-8. CGTs represent a new type of feature, using gland tensors to describe glandular disorganization, that is not explicitly captured by any previous QH feature.



(a) (b)

Figure 5.12: Comparison of $\mathcal{C}^{QH}$ in terms of mean (a) ϕ^{Acc} and (b) ϕ^{AUC} (with error bounds denoting standard deviation) for predicting BCR in 40 patients $\mathcal{P}$ over 100 runs of Random Forest with randomized 3-fold cross validation

5.12.3 Experiment 3: Identifying CaP Patients $\mathcal{P}$ with Biochemical Recurrence

Design: 40 men with CaP who had undergone RP were selected, 20 of whom had BCR and 20 of whom did not. For each patient, the largest annotated cancer region for each of the 40 patients $\mathcal{P}_i$ was selected to represent each $\mathcal{P}_i$. We compare the efficacy of CGT features with previously studied QH features (Table 5.6) for the purpose of differentiating patients $\mathcal{P}$ who will develop BCR (+1) from patients who will not (-1), following RP. Classifiers $\mathcal{C}^{QH}$, where $QH \in \{M, V, D, MST, GD, T, CGT\}$, were built for each QH feature. For evaluation, 100 iterations of randomized 3-fold cross validation was performed using Random Forest to predict BCR for each $\mathcal{P}_i$.

Results: In Figure 5.12, CGTs are shown to outperform each of the 6 other QH features in predicting BCR in 40 CaP patients $\mathcal{P}$ in terms of ϕ^{AUC} and ϕ^{Acc}. These results were statistically significant ($p < 0.05$) as shown in Table 5.10. Furthermore, the results in Table 5.10 are consistent with the results from Table 5.9, where regions $\mathcal{R}_i$ are classified as belonging to a patient $\mathcal{P}$ with BCR (+1) or NR (-1). These results suggest the use of CGTs over other QH features for identifying BCR in CaP patients from CaP histology.

Table 5.8: Mean and Standard Deviation of (a) ϕ^{Acc} and (b) ϕ^{AUC} of $\mathcal{C}^{QH}$ for distinguishing cancer from non-cancer in 80 regions $\mathcal{R}_i$ over 100 runs of Random Forest with randomized 3-fold cross validation. Associated Wilcoxon Rank Sum Test p-values for ϕ^{Acc} and ϕ^{AUC} of $\mathcal{C}^{QH}$ compared to $\mathcal{C}^{CGT}$ is provided.

	Gland Morphology	Voronoi	Delaunay	Minimum Spanning Tree	Gland Density	Texture	CGT
ϕ^{Acc}	$96.80 \pm 1.21\%$	$90.14 \pm 2.28\%$	$92.04 \pm 1.33\%$	$94.23 \pm 1.72\%$	$96.26 \pm 1.45\%$	$96.08 \pm 0.87\%$	$\mathbf{99.06 \pm 0.76}\%$
p-value	3.2183e-28	2.1923e-35	1.0692e-35	7.0137e-35	8.3012e-31	4.6573e-35	-
ϕ^{AUC}	0.9816 ± 0.0087	0.9126 ± 0.0239	0.9252 ± 0.0192	0.9488 ± 0.0195	0.9629 ± 0.0152	0.9459 ± 0.0058	$\mathbf{0.9951 \pm 0.0077}$
p-value	7.1411e-23	2.2052e-34	2.2058e-34	3.9028e-34	7.4428e-31	2.1993e-34	-

Table 5.9: Mean and Standard Deviation of (a) ϕ^{Acc} and (b) ϕ^{AUC} of $\mathcal{C}^{QH}$ for predicting 56 cancerous regions $\mathcal{R}_i$ corresponding to BCR patients and NR over 100 runs of Random Forest with randomized 3-fold cross validation. Associated Wilcoxon Rank Sum Test p-values for ϕ^{Acc} and ϕ^{AUC} of $\mathcal{C}^{QH}$ compared to $\mathcal{C}^{CGT}$ are shown below.

	Gland Morphology	Voronoi	Delaunay	Minimum Spanning Tree	Gland Density	Texture	CGT
ϕ^{Acc}	$62.34 \pm 3.92\%$	$62.91 \pm 4.39\%$	$64.34 \pm 3.48\%$	$63.20 \pm 3.39\%$	$62.93 \pm 3.72\%$	$59.87 \pm 4.04\%$	$\mathbf{71.61 \pm 3.67}\%$
p-value	2.1162e-29	3.9107e-26	6.5838e-26	4.7951e-29	7.8734e-29	4.5267e-32	-
ϕ^{AUC}	0.5554 ± 0.0399	0.5612 ± 0.0408	0.5586 ± 0.0391	0.5479 ± 0.0372	0.5470 ± 0.0346	0.5894 ± 0.0624	$\mathbf{0.6583 \pm 0.0532}$
p-value	3.4576e-26	1.4102e-24	1.217e-25	3.1953e-28	5.4087e-29	2.1755e-13	-

Table 5.10: Mean and Standard Deviation of (a) ϕ^{Acc} and (b) ϕ^{AUC} of $\mathcal{C}^{QH}$ for predicting BCR in 40 CaP patients $\mathcal{P}$ following RP over 100 runs of Random Forest with randomized 3-fold cross validation. Associated Wilcoxon Rank Sum Test p-values for ϕ^{Acc} and ϕ^{AUC} of $\mathcal{C}^{QH}$ is shown.

	Gland Morphology	Voronoi	Delaunay	Minimum Spanning Tree	Gland Density	Texture	CGT
ϕ^{Acc}	$59.20 \pm 4.07\%$	$64.47 \pm 4.72\%$	$64.82 \pm 4.66\%$	$64.62 \pm 3.95\%$	$64.60 \pm 3.60\%$	$68.80 \pm 4.15\%$	$\mathbf{73.18 \pm 5.02}\%$
p-value	5.418e-10	1.2662e-08	1.1547e-10	1.7115e-11	8.4514e-10	6.9014e-28	-
ϕ^{AUC}	0.5817 ± 0.0567	0.5600 ± 0.0472	0.5657 ± 0.0414	0.5594 ± 0.0430	0.5550 ± 0.0429	0.6017 ± 0.0568	$\mathbf{0.6693 \pm 0.0711}$
p-value	8.4424e-08	3.6478e-06	1.453e-07	4.0656e-09	4.9759e-11	0.012387	-

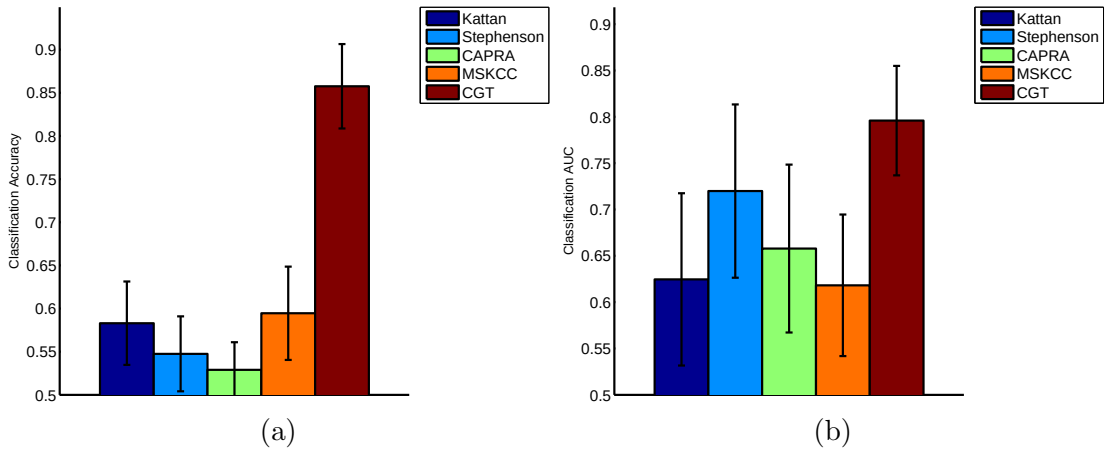


Figure 5.13: Comparison of $\mathcal{C}^{CGT}$ with $\mathcal{C}^{PT}$ in terms of mean (a) ϕ^{Acc} and (b) ϕ^{AUC} (with error bounds denoting standard deviation) for predicting BCR in 20 patients in $\mathcal{P}^B$ over 100 runs of Random Forest with randomized 3-fold cross validation

5.12.4 Experiment 4: Cross-validation within patient cohort $\mathcal{P}^B$:

Design: Of the 40 patients $\mathcal{P}_i$, only 20 patients $\mathcal{P}_i^B$ had associated clinical variables for nomogram prediction. We defined the group with additional clinical variables as Cohort B ($\mathcal{P}^B$), and the original group as Cohort A ($\mathcal{P}^A$). The breakdown for both cohorts are described in Table 5.5. For calculating QH features, the largest annotated cancer region was selected to represent each $\mathcal{P}_i$. f extracted from $\mathcal{R}_i$ is used to train a classifier $\mathcal{C}^{QH}$, which is used to predict the outcome $\ell(\mathcal{P}_i) \in \{-1, +1\}$ of patients $\mathcal{P}_i$. Comparatively, prostate cancer prediction tools (e.g. nomograms) also predict patient outcome $\ell(\mathcal{P}_i)$, in terms of BCR.We calculated scores s for each of the nomograms based on the clinical variables discussed in Table 5.7. These nomogram scores are used to represent each of the patients $\mathcal{P}_i^B$ in the same way that f is used to represent each $\mathcal{P}_i^B$. A classifier built from these prediction tools is denoted as $\mathcal{C}^{PT}$. Classifiers $\mathcal{C}^{PT}$ and $\mathcal{C}^{CGT}$ are built from s and f respectively. For evaluation, 100 iterations of randomized 3-fold cross validation was performed using Random Forest to predict BCR for each $\mathcal{P}_i^B$.

Results: In Figure 5.13(a), the performance of $\mathcal{C}^{CGT}$ was evaluated against CaP prediction tools $\mathcal{C}^{PT}$, where $PT \in \{K, S, CAPRA, MSK\}$. CGTs show improvement over 4 state-of-the-art prostate cancer nomograms, demonstrating 85% classification accuracy compared to 59% accuracy for the MS-KCC nomogram, which had the second highest ϕ^{Acc} (Table 5.11). In Table 5.11, CGTs show statistically significant improvement in ϕ^{AUC} over all nomograms, and outperforms the Stephenson and CAPRA nomograms ($\phi^{AUC} = 0.79$ for CGTs compared to $\phi^{AUC} = 0.72$ for Stephenson and $\phi^{AUC} = 0.66$ for CAPRA). These results suggest that CGTs have a greater predictive value than the state-of-the-art post-operative nomograms. Furthermore, this improvement can be obtained using only features present on CaP histology, whereas nomograms require additional clinical information such as pre-operative PSA, surgical margins, seminal vesicle invasion, and lymph node involvement.

Table 5.11: Mean and Standard Deviation of (a) ϕ^{Acc} and (b) ϕ^{AUC} for predicting BCR in $\mathcal{P}^B$. over 100 runs of Random Forest with randomized 3-fold cross validation. Associated Wilcoxon Rank Sum Test p-values for ϕ^{Acc} and ϕ^{AUC} of $\mathcal{C}^{PT}$ compared to $\mathcal{C}^{CGT}$ for predicting BCR are shown.

CaP Predictor	Kattan	Stephenson	CAPRA	MS-KCC	CGT
ϕ^{Acc}	$58.30 \pm 4.83\%$	$54.75 \pm 4.34\%$	$52.90 \pm 3.19\%$	$59.45 \pm 5.41\%$	$\mathbf{85.75 \pm 4.89}\%$
p-value	4.2558e-35	3.3989e-35	1.3793e-35	5.2764e-35	-
ϕ^{AUC}	0.6246 ± 0.0929	0.7199 ± 0.0935	0.6579 ± 0.0906	0.6182 ± 0.0764	$\mathbf{0.7959 \pm 0.0591}$
p-value	5.1173e-25	6.0125e-10	2.1506e-21	7.2366e-29	-

5.12.5 Experiment 5: Receiver Operating Characteristic (ROC) analysis of $\mathcal{P}^B$:

Design: To increase the size of the testing set to 20 patients, we performed a study using two independent cohorts $\mathcal{P}^A$ and $\mathcal{P}^B$. Analysis via the calculation of the Receiver Operating Characteristic (ROC) curve is used to determine the overall performance across all classification thresholds of each classifier $\mathcal{C}$.

For the CGT features, we perform the classification on $\mathcal{P}^B$ by creating a classifier $\mathcal{C}^{CGT}$ trained from $\mathcal{P}^A$. For the Random Forest classifier, each prediction $\ell(\mathcal{P}_i^B)$ is given a fuzzy decision value $\hat{p}$ between -1 and 1. Nomograms do not require further training beyond the original fitting done in the original study from which the nomogram was developed. Each nomogram is designed to predict BCR risk based on a score s. By setting decision thresholds at different $\hat{p}$ and s for $\mathcal{C}^{QH}$ and $\mathcal{C}^{PT}$ respectively, we obtained sensitivity and specificity scores at each threshold. The area under the ROC curve (ϕ^{AUC}) is subsequently calculated for each $\mathcal{C}^{QH}$ and $\mathcal{C}^{PT}$ to compare their performance on $\mathcal{P}^{\mathcal{B}}$.

Results: In Figure 5.14, we show reciever operating characteristic (ROC) curves for an independent cohort $\mathcal{P}^B$ for $\mathcal{C}^{PT}$, $PT \in \{K, S, CAPRA, MSK\}$ and $\mathcal{C}^{CGT}$. $\mathcal{C}^{CGT}$ demonstrates a clear improvement over 4 state-of-the-art nomograms $\mathcal{C}^{PT}$, showing an AUC of 0.76 compared to AUCs near 0.5 for all $\mathcal{C}^{PT}$. The weak performance on this independent cohort highlights the difficulty of modern nomograms to predict BCR for men with intermediate-risk GS scores and suggests that the addition of QH features could improve upon the current nomogram standards.

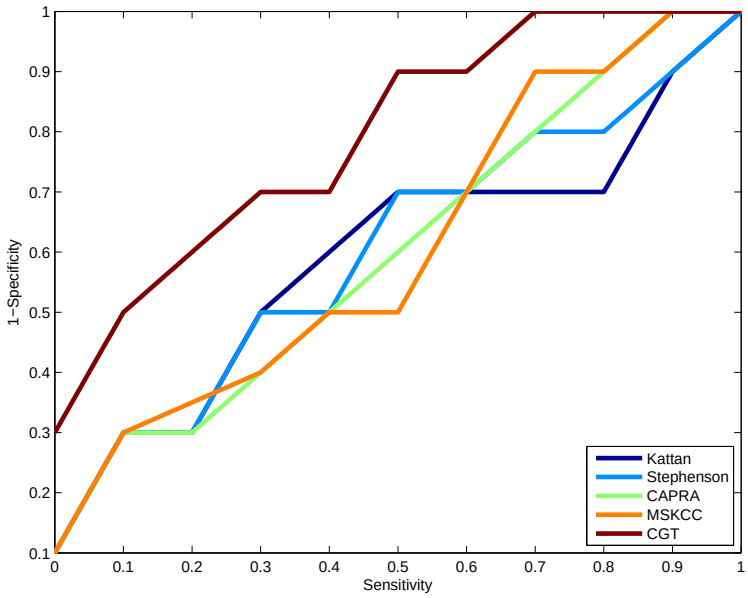


Figure 5.14: Comparison of Receiver Operating Characteristic (ROC) Curves for $\mathcal{C}^{CGT}$ versus 4 postoperative CaP nomograms $\mathcal{C}^{PT}$ in an independent 20 patient cohort $\mathcal{P}^B$.

5.12.6 Experiment 6: Kaplan-Meier analysis of $\mathcal{P}^B$:

*Design:*Kaplan-Meier analysis of 20 patients in $\mathcal{P}^B$ demonstrates the difference in BCR free survival time associated with each predictor $\mathcal{C}^{PT}$, $PT \in \{K, S, CAPRA, MSK\}$ and $\mathcal{C}^{CGT}$. Similar to the previous experiment in Section 5.12.5, $\mathcal{C}^{CGT}$ is trained from $\mathcal{P}^A$ and $\mathcal{C}^{PT}$ does not require additional training. Each $\mathcal{C}$ predicts labels $\ell(\mathcal{P}_i^B) \in \{-1, +1\}$ (BCR or NR) for each $\mathcal{P}_i^B$, and a BCR-free survival curve based on the time to recurrence information of each $\mathcal{P}_i^B$ is generated from each of the resulting predicted BCR and NR groups. The logrank test was used to determine a p-value associated with the difference between the survival curves. A lower p-value indicates greater differences in the BCR-free survival between the predicted BCR and NR groups.

Results: In Figure 5.15, we show Kaplan-Meier survival curves based on the predicted BCR and NR groups of each $\mathcal{C}$. Based on the logrank test (shown in Table 5.12), patients were best differentiated via $\mathcal{C}^{CGT}$, with a p-value of 0.0016 compared to 0.0596 for the MS-KCC nomogram, which had the next lowest p-value. The superior differentiation in the survival curves afforded by $\mathcal{C}^{CGT}$ is indicative of its value over current nomograms.

Table 5.12: Logrank test p-values for comparison of Kaplan-Meier survival curves of $\mathcal{P}^B$ stratified into BCR and NR groups by each $\mathcal{C}$

CaP Predictor	Kattan	Stephenson	CAPRA	MS-KCC	CGT
p-value	0.18458	0.29343	0.42146	0.059585	0.0016624

$\mathcal{C}^{CGT}$ represents the only predictor which show statistically significant differentiation ($p < 0.05$) in the survival outcomes of its predicted patient cohorts.

5.13 Summary

We presented a novel set of QH features using co-occurring gland tensors (CGT) calculated on local subgraphs. CGTs represent a novel combination of subgraphs, gland tensors, and tensor co-occurrence matrices to quantify the local disorder in the gland tensors on prostate cancer (CaP) histopathology. Following 4 sets of experiments on 40 Gleason score 6-8, CaP patients following RP, we found CGT features demonstrated a statistically significant ($p < 0.05$) improvement in classification accuracy compared to 6 comparison QH features. Furthermore, we found CGTs to outperform 4 state-of-the-art postoperative nomograms for predicting BCR in CaP patients. Complementary Kaplan-Meier analysis of the independent cohort of prostate cancer patients with intermediate-risk pathological Gleason scores (via the logrank test) demonstrated that only the CGTs showed a statistically significant ($p < 0.05$) difference in the survival distributions of the predicted cohorts. While we attempted to account for bias from clinical variables by comparing our work with state-of-the-art nomograms which use these variables, we acknowledge the need to validate our results on additional data on even more constrained cohorts to control for the clinical variables.

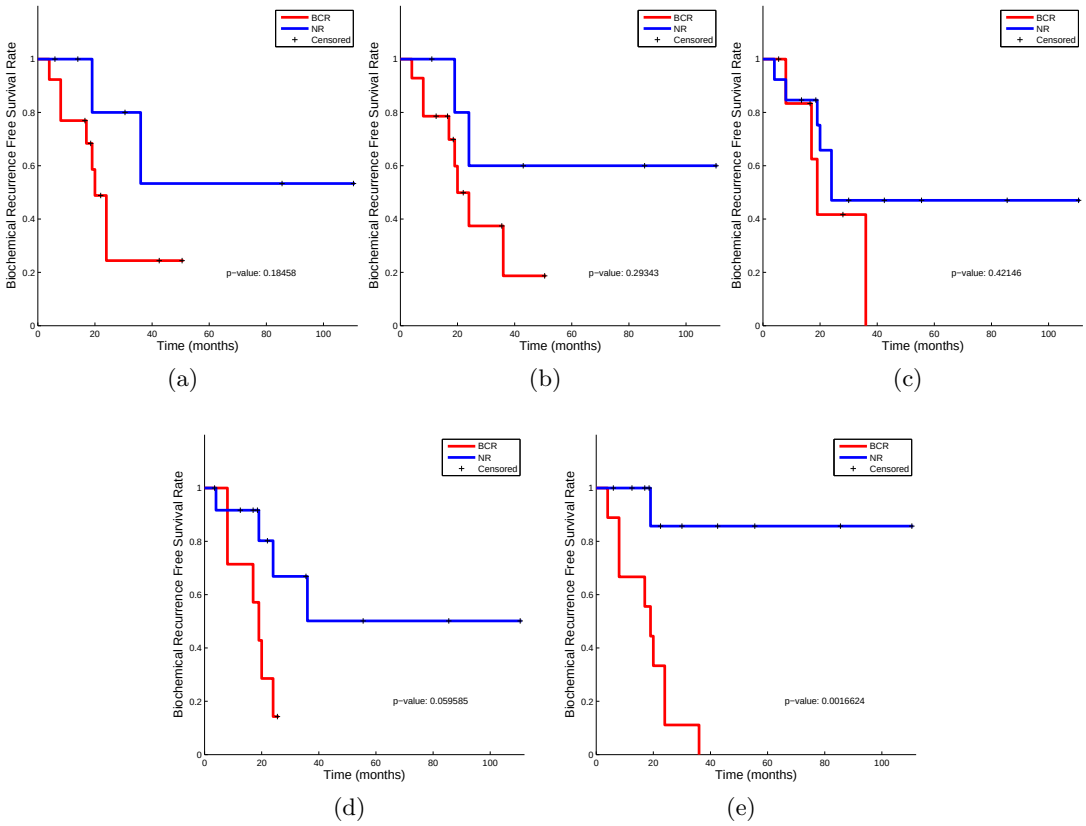


Figure 5.15: Comparison of Kaplan-Meier BCR-free survival curves differentiated via (a) Kattan nomogram, (b) Stephenson nomogram, (c) UCSF-CAPRA, (d) MS-KCC nomogram, and (e) CGT classifiers on an independent 20 patient cohort $\mathcal{P}^B$. Lower p-values are indicative of better predictors of BCR.

Chapter 6

Evaluation of Supervised Multi-view Canonical Correlation Analysis for an integrated histologic and proteomic biomarker

In this work, we leverage the sMVCCA framework (described in Section 4) to build an integrated biomarker which combines quantitative image features derived from tissue histopathology along with proteomic features obtained via mass spectrometry. A classifier built from the integrated biomarker is used to identify patients at risk for 5 year biochemical recurrence following radical prostatectomy. The main steps involved in building this classifier include (1) feature extraction of histological image and proteomic features (Figs. 6.1(a)-(d)), (2) sMVCCA for an integrated biomarker representation (Fig. 6.1(e)), and (3) building a classifier from the integrated biomarker to distinguish patients at risk for BCR within 5 years from those who are not (Fig. 6.1(f)). Results were compared with other QH measures and CaP prediction models.

Some content of this chapter is taken from [88], on which the author of the dissertation is the second author. This content describes the acquisition of the proteomic and histologic data as well as the extraction of histologic features, which was subsequently analyzed by the author as part of his thesis research.

6.1 Data Acquisition and Data Description

40 patients with biopsy confirmed CaP underwent RP at the Hospital at the University of Pennsylvania (HUP). Following radical prostatectomy, the resected prostate was sectioned with a meat cutter into histological slices for analysis. Pathological Gleason score and detailed cancer annotations were provided by pathologists at HUP.

Cases	Classes	Features
40	Biochemical recurrence (BCR) vs. Non-biochemical recurrence (NR)	Proteomic: 650 expression values of proteins such as heat shock protein and Ras regulated proteins; Histomorphometric: 242 features extracted from high resolution images of the dominant tumor nodule

Table 6.1: Brief description of the UPENN prostate cancer dataset and features extracted from each modality.

Numerous works have demonstrated the utility of QH features for Gleason scoring [31, 173] and prostate cancer prognosis [171, 175].

A representative slice containing the most dominant tumor nodule in each specimen was digitized at 20x magnification (0.5μm per pixel) using a whole slide digital scanner. Mass spectrometry was performed on the same dominant tumor nodule to identify a set of proteins corresponding to the annotated tumor region in the digital image. BCR was indicated by a PSA of at least 0.2 ng/mL. Among all the patients, 21 experienced biochemical recurrence (BCR) within 5 years of surgery while the other 19 did not have BCR.

6.1.1 Proteomic Feature Extraction and Selection

Prostate slides were deparaffinized, and rehydrated as described in [184]. Tumor areas previously defined on a serial H&E section were collected by needle dissection, and formalin cross-links were removed by heating at 99 degrees Celsius. The FASP (Filter-Aided Sample Preparation) method [185] was then used for buffer exchange and tryptic digest. After peptide purification on C-18 StageTips [186], samples were analyzed using nanoflow C-18 reverse phase liquid chromatography/tandem mass spectrometry (nLC-MS/MS) on a LTQ Orbitrap mass spectrometer. A top-5 data-dependent methodology was used for MS/MS acquisition, and data files were processed using a label free MaxQuant peptide identification package [187] that uses extracted ion chromatograms to calculate protein abundance. The resulting 650 dimensional feature vector was obtained consisting of quantifiable proteins found across at least 50% of the studies and is used to characterize each patient's protein expression profile following surgery. Missing values were replaced by data imputation as described in [188].

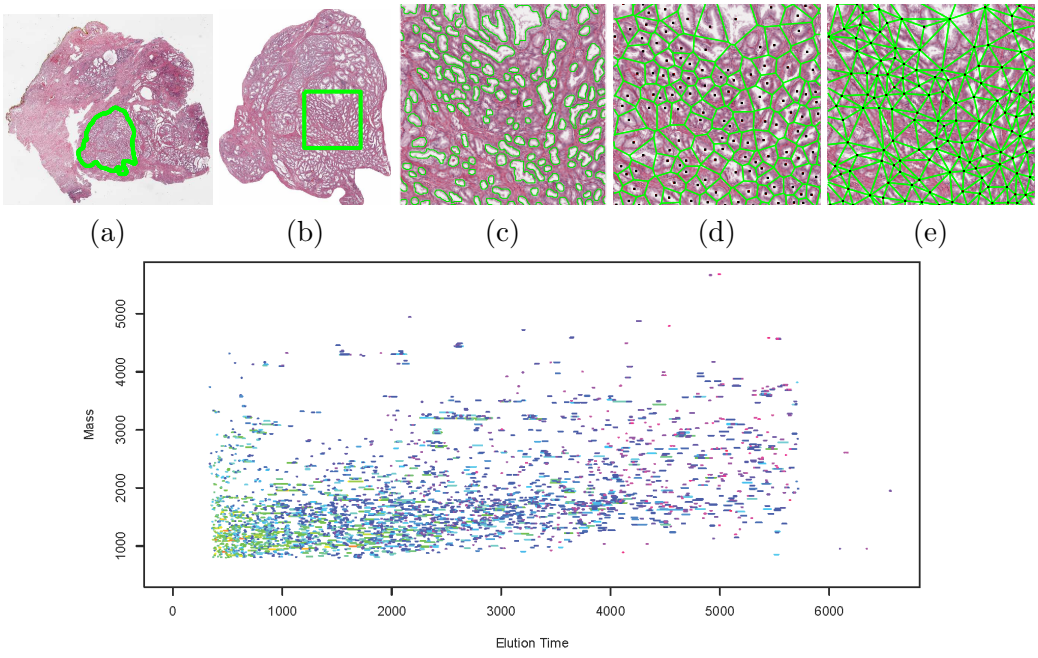


Figure 6.1: (a) Prostate histology with cancerous region annotated by a pathologist. (b) Area of QH feature extraction (with zoom window to demonstrate QH features) (c) Gland morphology captured by automated segmentation of the interior lumen boundary. (d) Voronoi diagram and (e) Delaunay triangulation of gland centroids (shown in red) describe architecture. (f) 2D liquid chromatography and mass spectrometry profile allows for a high sensitivity detection and quantification of proteins.

6.1.2 Histomorphometric Feature Extraction

For the assessment of prostate whole-mount histology, analysis of glands are of particular interest as their shape and arrangement have been found to be highly correlated with cancer aggressiveness [189,190]. Based on the segmentation and the annotated regions on the digital images (Figure 6.1), we extracted a total of 242 histomorphometric features from gland morphology, architecture, to distinguish aggressive CaP. We briefly describe the image segmentation and feature extraction process below.

Gland Segmentation

Prior to extracting image features, we employ an automatic region-growing gland segmentation algorithm presented by [191]. The boundaries of the interior gland lumen and the centroids of each gland, allow for extraction of 1) morphological and 2) architectural features from histology as described briefly below. More extensive details on these methods are available in our other publications [170].

Glandular Morphology

The set of 100 morphological features [51] consists of the average, median, standard deviation, and min/max ratio for gland area, maximum area, area ratio, estimated boundary length, standard deviation of distance, variance of distance, distance ratio, perimeter ratio, smoothness, fractal dimension, as well as descriptors of invariant moments and Fourier transforms. (See Table 6.1). These features have been shown previously to distinguish Gleason grades on H&E stained prostate histopathology [174].

Architectural Feature Extraction

51 architectural image features were extracted in order to quantify the arrangement of glands present in the prostate section (See Table 6.1). Previously, these features were shown to be useful in discriminating between different Gleason grades of CaP histopathology [174]. Voronoi diagrams, Delaunay triangulation [175] and minimum spanning trees were constructed on the digital histologic image using the gland centroids

as vertices, the gland centroids having previously been identified via the scheme in [170].

Co-occurring Gland Tensors

Co-occurring gland tensors (CGTs) [8] represent a novel type of feature which computes an approximation of disorder in glandular orientation on the histology. Following gland segmentation, graph orientations for each gland are computed from the principal axis of the segmented gland boundaries. Angles are subsequently obtained from the principal axes and quantized across 0 to 180 degrees. To capture local distributions in gland directionality, we apply a spatial constraint defined by a probabilistic decaying function. 13 second order statistical features are subsequently extracted from tensor co-occurrence matrices which aggregate orientation information from each neighborhood. The mean, standard deviation and range of these second order features result in 39 CGT features.

Gland Subgraphs

26 features [177] used to calculate cell cluster graphs in orophareangyal and prostate cancers as well as breast cancer [176] were used to characterize the glands in prostate cancer. Based on the probabilistic decay function to generate local subgraphs between the glandular centroids, features such as eccentricity and connected component coefficients are extracted from the resulting local subgraphs.

Intensity Texture

Second order co-occurrence features [31, 161] are calculated from a symmetric co-occurrence matrix which aggregates the frequency in which two pixel intensities co-occur within a pre-determined window distance around each pixel. The size of the co-occurrence matrix is determined by the maximum possible intensity value in the image, which for 8-bit images is $2^8 = 256$. A window distance of 1 pixel was chosen. For each pixel, contrast energy, contrast inverse moment, contrast average, contrast variance, contrast entropy, intensity average, intensity variance, intensity entropy, entropy, energy, correlation, and two information measures are computed from the co-occurrence matrix. The mean and standard deviation of these features across all pixels are used

to build a set of 26 intensity texture features f for each image.

6.2 Methods of Evaluation

6.2.1 Feature Selection via Wilcoxan Rank Sum Test

Features extracted from the various modalities are summarized in Table 6.1. An nested cross-validated feature selection scheme is used independent of the classification folds. The cross-validated classification folds are used as an 'outer loop' to eliminate overfitting and bias in the feature selection process [192, 193]. Wilcoxon rank sum test (WRST) was then used to select features in each modality [93]. Features with p-value ≤ 0.05 were considered to be statistically significant in differentiating the object classes and were selected as discriminatory. This process was repeated n-1 times using n-2 samples for a leave-one-out consensus selection of features to compute each WRST. The intersection of features found to be $p \leq 0.05$ across at least 2/3 of the n-1 sets were used as the feature subset for each classification fold. Through this process, n sets of features were obtained for each classification fold.

6.2.2 Embedding Construction

For each dimensionality $d \in \{1, 2, \ldots, 10\}$, n embeddings were constructed from n feature subsets identified in Section 6.2.1. Labels used for training the embedding was limited to the training label set available for the classifier and feature selection process $\mathcal{S}^{tr} \notin \mathcal{S}^{ts}$. No testing labels contributed to the construction of the $n * d$ embeddings.

For RCCA and SRCCA, regularization parameters γ_1 and γ_2 were selected via grid search optimization intervals $\theta_1 = 0.005$ to $\theta_2 = 0.2$ with 80 evenly spaced intervals.

6.2.3 Classification via Random Forest

n-fold cross validation of Random Forest classifiers [180] were used to evaluate the performance of sMVCCA with the following comparative strategies: (i) PCA, (ii) GEC,

(iii) CCA (iv) RCCA, (v) MVCCA, (vi) SRCCA, (vii) selected histomorphometric features, and (viii) selected proteomic features.

For each fold, feature selection was performed as described in Section 6.2.1. Subsequently, for each fold, an embedding $\mathcal{E}_i$ is created as described in Section 6.2.2. Lastly, we can evaluate the embedding by training n classifiers $\mathcal{C}_i$ on each of n embeddings $\mathcal{E}_i$ and $\mathcal{S}_i^{tr}$ for the purpose of predicting the label $\mathbb{Y}$ of each corresponding $\mathcal{S}_i^{ts}$. The result of the n classifications is used to create the area under the receiver operating characteristic curve (AUC) [194] is used to evaluate the overall performance of each data fusion method for generating embeddings which can be used as an integrated biomarker for BCR.

6.2.4 Kaplan-Meier Analysis of biochemical recurrence free survival rates

Additional Kaplan-Meier analysis [182] was performed on 30 samples (15 BCR, 15 NR) where time to recurrence information was available. Kaplan-Meier curves demonstrate the resulting survival outcome of the predicted object class groups. In this study, the object class groups $\mathbb{Y}$ are predicted via the classifiers constructed from each of eigenvectors $\mathcal{E}$ or features as described previously. An optimal prediction would demonstrate a large difference in the survival outcome between the predicted groups. When plotted onto time versus BCR-free survival rate, the BCR free survival rate of the group will decrease at the time after surgery when each patient develops BCR. Thus, we expect the curve for the set of patients predicted to have BCR to trend towards 0 quickly while curve pertaining to the set of patients predicted to have the label NR to maintain 100% BCR-free survival. This difference in time and outcome between the two predicted subject groups can be quantified via the log-rank test [183], where $p \leq 0.05$ qualifies as a statistically significant difference between outcomes of the two predicted cohorts.

6.2.5 Computational Run Time

We have also examined the computational run times for generating the embeddings from the features. For each data fusion method, mean and standard deviation across

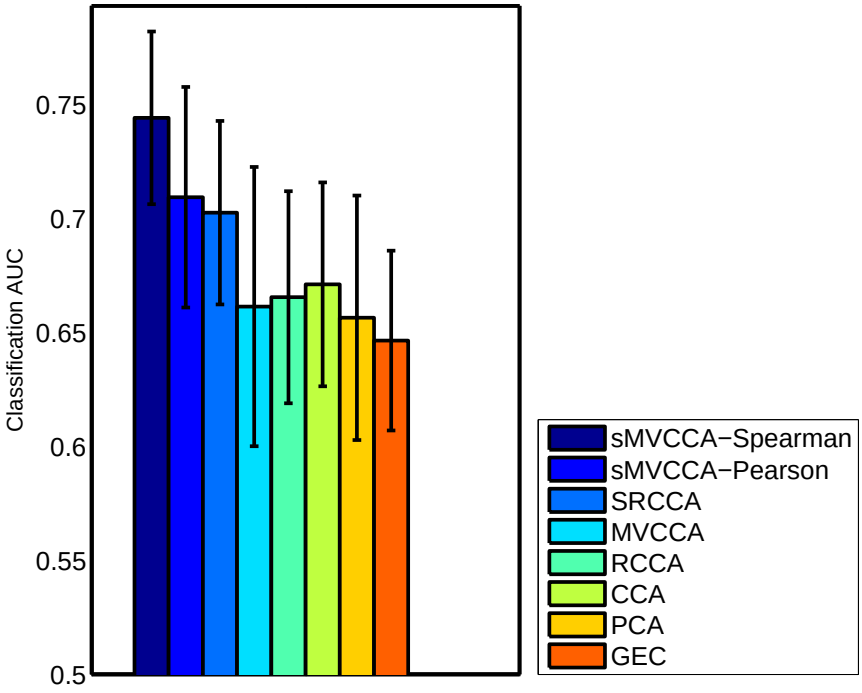


Figure 6.2: Mean and standard deviation of classification AUC achieved by 40 Random Forest classifiers in sMVCCA, SRCCA, MVCCA, RCCA, CCA, PCA and GEC reduced space via n-fold cross validation across dimensionality $d \in \{1, 2, \ldots, 10\}$.

Method	SRCCA	MVCCA	RCCA	CCA	PCA	GEC	Histomorphometric	Proteomic
sMVCCA-Pearson	0.3689	**0.0338**	**0.0267**	**0.0415**	**0.0163**	**0.0025**	**3.3794e-06**	0.7599
sMVCCA-Spearman	**0.0143**	**0.0009**	**0.0003**	**0.0005**	**0.0003**	**1.1552e-05**	**1.1811e-09**	**0.0311**

Table 6.2: p-values comparing AUC values $d \in \{1, 2, \ldots, 10\}$ for sMVCCA-Pearson and sMVCCA-Spearman with comparative data fusion methodologies and Imaging and Proteomic features alone via Student t-test. Significant p-values ($p < 0.05$) are shown in **bold**.

40 embeddings and 10 dimensionalities. Experiments were run on a quad-core i7-3770 CPU with a clock speed of 3.4 GHz and programs were written on MATLAB®.

6.3 Results and Discussion

6.4 Classification of BCR and NR CaP patients

In terms of classification AUC, sMVCCA consistently outperformed other unsupervised and supervised methods across dimensions $d \in \{1, 2, \ldots, 10\}$ as shown in Figure 6.2. The breakdown of mean classification performance by dimensionality is shown in Table 6.1. We can make several observations upon inspection of this information. Firstly,

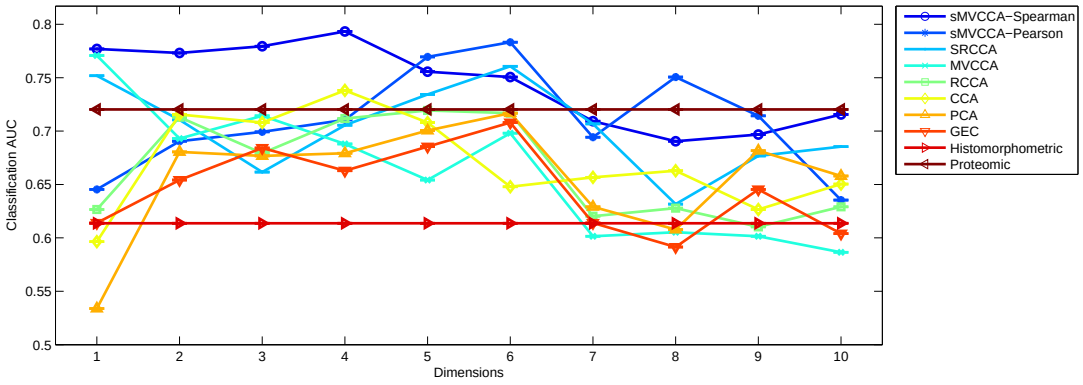


Figure 6.3: Histomorphometric and Proteomic Data Fusion via sMVCCA, SRCCA, MVCCA, RCCA, CCA, PCA and GEC: Mean AUC for each dimensionality $d \in \{1, 2, \ldots, 10\}$ in predicting BCR. Mean classification of the selected histomorphometric and proteomic features are shown for reference.

we can recognize a clear difference in sMVCCA-Spearman compared to most other data fusion methods. While sMVCCA-Pearson, SRCCA, and MVCCA are able to outperform sMVCCA-Spearman for specific dimensionalities, sMVCCA-Spearman is shown to be the most consistent. Secondly, There is a noticable decline in overall classification AUC across all data fusion methods for $d > 6$. Given the noticable consensus across data fusion methods, we can deduce that there is an optimal dimensionality for constructing a classifier. Due to the use of ranked features combined with labeled information, sMVCCA-Spearman is able to uncover discriminatory features with fewer dimensions, as maximum AUC is achieved at $d = 4$. sMVCCA-Pearson and SRCCA also illustrate strong performance but with greater dimensions, where maximum AUC is achieved at $d = 6$ for both these methods.

Statistical significance via the Wilcoxon Rank Sum Test is shown in Table 6.2. sMVCCA-Spearman showed statistically significantly ($p < 0.05$) better classification AUC compared to all other data fusion methods and individual modalities alone. sMVCCA-Pearson also displayed significantly better classification performance against all other unsupervised methods, but was not statistically significantly better than SRCCA and proteomic data alone.

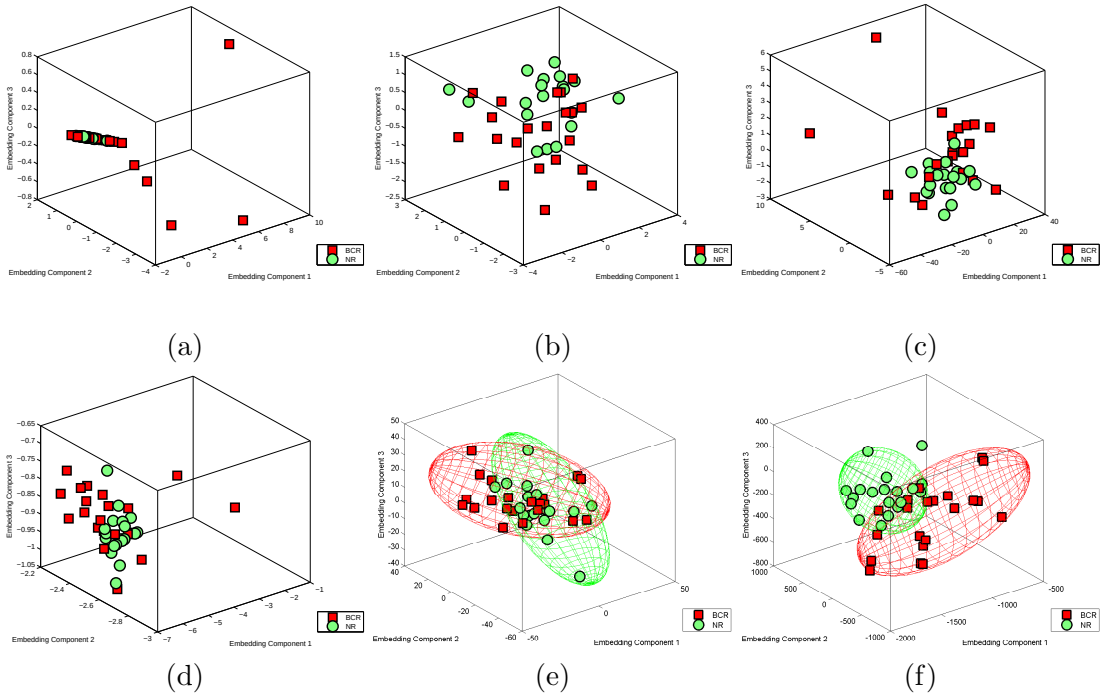


Figure 6.4: 3-D Embedding plots by Principal Component Analysis of the best performing (a) Histomorphometric Features and (b) PCA of Proteomic Features show the distribution of CaP patients with BCR (red squares) and NR (green circles). 3-D embedding plots pertaining to the highest classification AUC of the integrated histomorphometric and proteomic features via (c) MVCCA, (d) SRCCA, (e) sMVCCA-Pearson, and (f) sMVCCA-Spearman reveal potential manifestations of an integrated biomarker for BCR.

We hypothesize that sMVCCA-Pearson and SRCCA do not account for the nonlinearity in correlations between histology and proteomic features, while sMVCCA-Spearman's use of ranked features is able to better account for the outliers seen in the 3D representation of histomorphometric features via PCA in Figure 6.4(a). 3-D embedding plots are shown for visualization purposes in Figure 6.4. The separability illustrated in these features are consistent with the classification results reported in Figure 6.2, as MVCCA (Figure 6.4(c))and SRCCA (Figure 6.4(d)) do not demonstrate significantly better separation compared to proteomic features alone (Figure 6.4(b)). However, Figure 6.4(f) is reflective of sMVCCA-Spearman's superior classification performance, illustrating the separation of CaP patients with BCR and NR.

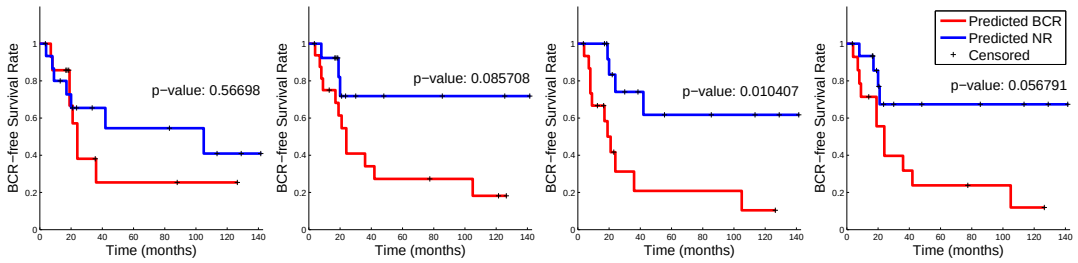


Figure 6.5: Kaplan-Meier Analysis of Prostate Cancer Patient Outcome as determined via the classification dictated by (a) Histomorphometric Features, (b) Proteomic Features, (c) sMVCCA-Pearson, and (d) sMVCCA-Spearman

6.5 Comparing biochemical recurrence free surival rates via Kaplan-Meier Analysis

Figure 6.5 shows the BCR-free survival outcome of the patients predicted via Random Forest classifiers built on different sets of features. Histomorphometric and proteomic features do not show statistically significant separation in the predicted BCR and predicted NR outcomes shown via the log rank p-values of 0.567 and 0.086 respectively However, sMVCCA shows better stratification of outcome via lower p-values, with the integrated features produced via sMVCCA-Pearson showing $p = 0.010$ and sMVCCA-Spearman showing $p = 0.057$. While sMVCCA-Pearson produced a statistically significant stratification, sMVCCA-Spearman did not, with a p-value just above the 0.05 significance threshold. Although sMVCCA-Spearman was not able to stratify patients early on ($t < 20$ months) following RP, it has the same number of false positives as sMVCCA-Pearson after 20 months following RP, demonstrating better predictive value when compared to histmorphometric and proteomic features alone. As the p-value is just above the 0.05 significance threshold, it is possible that additional samples in the future could improve the p-value here.

6.6 Computational Run Time

Figure 6.6 shows the comparative computation run times of sMVCCA, SRCCA, MVCCA, RCCA, and CCA for generating the embeddings across all experiments in this study.

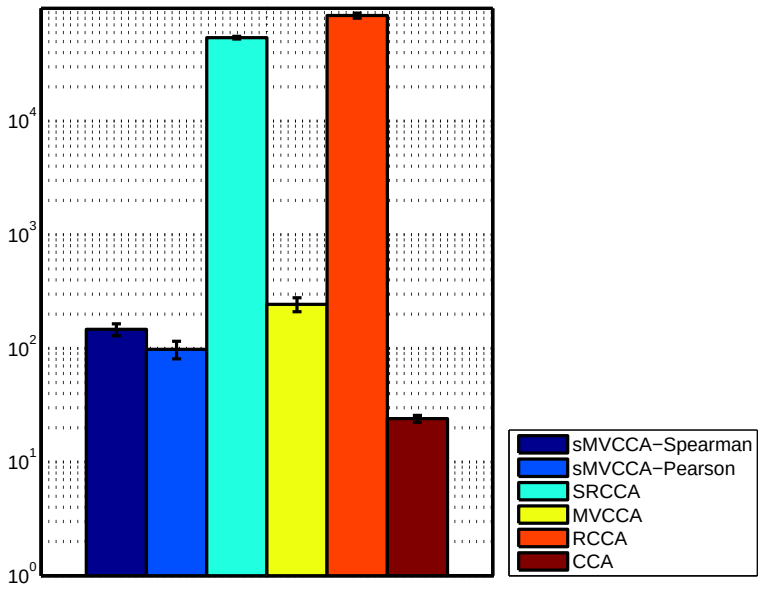


Figure 6.6: Comparison of Mean and Standard Deviation of Computational Run Times for generating embeddings via data fusion methods sMVCCA, SRCCA, MVCCA, RCCA, and CCA across all dimensionalities $d \in \{1, 2, \ldots, 10\}$

SRCCA and RCCA have the longest run times due to its expensive regularization procedure. Our experiments with SRCCA and RCCA took at least 2 orders of magnitude (100 times) longer compared to the other methods. Both sMVCCA-Pearson and sMVCCA-Spearman finish in less than an order of magnitude longer compared to MVCCA and CCA, but offer superior classification performance. These differences were found to be statistically significant across all runs.

6.7 Summary

We presented a novel supervised multi-view canonical correlation analysis method that is able to construct an integrated biomarker for identifying patients at risk for biochemical recurrence. Via sMVCCA, we were able to consistently predict BCR (mean of 0.74 AUC) at a statistically significantly ($p < 0.05$) higher rate compared to previous data fusion methods and compared to using histomorphometric and proteomic features alone. We demonstrate improved computational run times compared to previous supervised data fusion methods such as SRCCA. Additionally, our method is able to integrate data from any number of modalities to a joint subspace that is robust

to modality specific noise. We acknowledge some limitations in this study including the lack of additional views which MVCCA and sMVCCA are able to handle. However, we have demonstrated improvement in terms of classification and computational performance against the state-of-the-art SRCCA which has only been able to utilize 2 views.

Chapter 7

Concluding Remarks and Future Work

The motivation for an integrated diagnostics system stems from the deficiencies of single modality markers to adequately predict aggressive CaP. This work aimed to address the various challenges involved in building an improved predictor of biochemical recurrence following radical prostatectomy. The aims addressed include the following:

1. Challenges in data representation can be alleviated via dimensionality reduction. We developed a novel methodology (AdDReSS), which provides a superior low dimensional representation using active learning compared to random sampling of labeled training instances. AdDReSS was validated via classification, Silhouette cluster index, and learning efficiency.

2. Quantitative histomorphometry can improve upon Gleason scoring, the current gold standard for predicting aggressive prostate cancer. We contributed 2 QH methodologies, COrE and CGTs, which represent a new way of measuring the disorder of cells and glands, respectively, within the prostate tissue. These methods were validated on multiple cohorts of radical prostatectomy patients from different medical institutions and compared with state-of-the-art predictors of biochemical recurrence.

3. Integrated diagnostics represent the future for personalized medicine, and developing methods for combining useful prognostic markers is necessary to realize better prediction rates. sMVCCA was developed to take into account both the inter-modal dependencies (as measured by correlation) as well as their predictive value to create a low dimensional integrated biomarker. This was tested on a histological and proteomic data cohort which demonstrated the value of combining

two predictive modalities to produce an even stronger predictor. sMVCCA was shown to outperform previous data integration strategies in this task as shown via superior classification AUC and Kaplan-Meier p-values.

These methods represent significant contributions towards creating integrated diagnostics. Data representation, integration, and automated analysis of histological images are all vital components for future diagnosis of disease and prognosis of treatment outcomes. Although, we demonstrated many of these algorithms in the context of an integrated biomarker for prostate cancer, although the machine learning methods developed are extensible towards many potential applications beyond disease prediction.

Despite promising initial results, improved quantitation of proteins on an expanded QH-protein prostate cancer dataset has the potential to greatly improve the quality of the current study. With regards to biochemical recurrence prediction, patients in this dissertation were classified based on whether or not biochemical recurrence was found, but did not heavily take into consideration censored data for classification. Time to recurrence is an important attribute of aggressive CaP and modeling the distinction between censored patient information and non-recurrence could lead to more accurate predictions of treatment failure [195]. For example, Cordon-Cardo et al. utilized a modified support vector machine called SVRc which takes into account censored data, [28, 196]. Future work will aim to take advantage of time to recurrence information beyond post-classification survival analysis.

We had also begun to explore race-specific biomarkers. However, our current cohort was limited to only 3 African American patients with radical prostatectomy. Therefore, expanding on this cohort would be imperative towards validation of race specific biomarkers. Additionally, future work in quantitative histomorphometry would involve expansion of the validation study using a radical prostatectomy cohort from an independent institution, and in general, increasing the overall number of patients involved in the study.

Furthermore, correlation of QH-protein pairs may allow for the development of QH-based protein surrogates. QH-based protein surrogates would allow for the identification of protein concentrations from digital pathology without tissue extraction. Current methods for detecting proteins spatially are expensive may be difficult to quantify. Immunohistochemistry requires the development of specific antibodies for the staining of each identified protein. Hyperspectral imaging is capable of examining every spectra for protein signatures, but remains under development. Imaging mass spectrometry [197] captures an exact spatial location of molecules but is expensive. Once highly correlated QH-protein pairs are identified, classifiers built on correlated QH features can be used to predict protein concentrations in prostate tissue.

The proposed avenues of research in disease prognosis stemming from this dissertation are very exciting and will provide us with the knowledge to understand and properly treat the most malignant diseases of today. Fulfillment of these goals will realize the future of integrated diagnostics and personalized healthcare.

Chapter 8

Appendices

Appendix A: Overview of Feature Selection Methods

A feature selection or pruning step can be used to identify a set of informative features $\hat{F}(x_i) = [f_{\hat{u}}(x_i)|\hat{u} \in \left\{1, 2, ..., \hat{M}\right\}]$ where $\hat{M} < M$ for each sample $x_i \in D$, given the labels $Y \in \{-1, 1\}$ of n samples across the two paired object classes 1 and 2.

Student's t-test

The feature pruning method described in [60,74,92] is based on t-statistics used to model Gaussian distributions. For all $x_i \in D$ and for a specific feature $u \in \{1, 2, ..., M\}$, the mean $f_u^{\mu+}$, $f_u^{\mu-}$ and variance $f_u^{\sigma^2+}$, $f_u^{\sigma^2-}$ of the features for the +1 or -1 class were computed. Hence

$$f_u^{\mu+} = \frac{1}{n_+} \sum_{\substack{x_a \in D_j \\ Y(x_a)=+1}} f_u(x_a), \tag{8.1}$$

$$f_u^{\sigma^2-} = \frac{1}{n_-} \sum_{\substack{x_b \in D_j \\ Y(x_b)=-1}} (f_u(x_b) - f_u^{\mu-})^2. \tag{8.2}$$

The values of $f_u^{\mu+}$, $f_u^{\mu-}$, $f_u^{\sigma^2+}$, $f_u^{\sigma^2-}$ were then used to calculate the information content of each feature as

$$\mathcal{T}(f_u) = \frac{f_u^{\mu+} - f_u^{\mu-}}{\sqrt{\frac{f_u^{\sigma^2+}}{n_+} + \frac{f_u^{\sigma^2-}}{n_-}}}. \tag{8.3}$$

We can select the features with the greatest significance via the p-value associated with the T-statistic given for each feature. An arbitrary significance threshold $p < 0.05$

is commonly used to prune features f_u shown to provide statistically significant $p < 0.05$ differences between the Gaussian distributions of the labels modeled via f_u.

Wilcoxon Rank Sum Test

Wilcoxon rank-sum test (also known as the Mann-Whitney U test) [181] is a non-parametric statistical hypothesis test often used as an alternative to the paired Student's t-test.

The data, f_u, can be split using its labels into the n_1 samples that belong to $Y(x_a) = 1$ and the n_2 samples that belong to class $Y(x_b) = -1$, where $n_1 + n_2 = n$. These two partitions can then be used to calculate the discrimination level between the samples of the two classes

$$U = \min\left\{\Big(\sum_{i=1}^{n_2} b_i - \frac{n_2(n_2+1)}{2}\Big), \Big(n_1 n_2 - \sum_{i=1}^{n_2} b_i - \frac{n_2(n_2+1)}{2}\Big)\right\}, \tag{8.4}$$

where b_i represents the rank of the sample $i \in Y(x_i)$. U can subsequently be used to consult significance tables and obtain a p-value where $p < 0.05$ can be interpreted as a significant difference between the medians of $Y(x_a) = -1$ and $Y(x_a) = +1$ for f_u [181].

Appendix B: Overview of Machine Learning Classifiers

Support Vector Machines (SVMs)

Support vector machines (SVMs) were first introduced by Cortes and Vapnik [151] and are based on the structural risk minimization (SRM) principle from statistical learning theory. The SVM attempts to minimize a bound on the generalization error (error made on test data). SVM-based techniques focus on "borderline" training examples (or support vectors) that are most difficult to classify. The SVM projects the input training data $G^{\phi}(x_i)$, for $x_b \in S_j^{Tr}$, onto a higher-dimensional space using the linear kernel defined in Equation 8.5 as

$$\Pi(G^{\phi}(x_a), G^{\phi}(x_b)) = [G^{\phi}(x_a)]^T G^{\phi}(x_b) + \mathbf{b}, \tag{8.5}$$

where $\mathbf{b}$ is the bias estimated on the training set $S_j^{Tr} \subset D$. The general form of the SVM is given by

$$\mathcal{C}^{SVM} = \sum_{\beta=1}^{n_s} \xi_\beta Y(x_b) \Pi(G^\phi(x_a), G^\phi(x_b)), \tag{8.6}$$

where x_β, for $\beta \in \{1, 2, ..., n_s\}$ denotes the number of support vectors and the model parameter ξ is obtained by maximizing the following objective function.

$$\Lambda(\xi) = \sum_{\beta=1}^{n_s} \xi_\beta - \frac{1}{2} \sum_{\beta,\gamma=1}^{n_s} \xi_\beta \xi_\gamma Y(x_\beta) Y(x_\gamma) \Pi, \tag{8.7}$$

subject to the constraint $\sum_{\beta=1}^{n_s} \xi_\beta Y(x_\beta) = 0$ and $0 \leq \xi_\beta \leq \omega$, where $\beta, \gamma \in \{1, 2, ..., n_s\}$, and where the parameter ω controls the trade-off between the empirical risk (training errors) and model complexity.

C4.5 Decision Trees (C4.5)

A special type of classifier is the decision tree, which is trained using an iterative selection of individual features $f_u(x_a)$ that are the most salient at the each node in the tree [198]. One of the most commonly used algorithms for generating decision trees is the C4.5 rules proposed by Quinlan [198]. The rules generated by this approach are in conjunctive form such as "if A and B then C" where both A and B are the rule antecedents, while C is the rule consequence. Every path from the root to the leaf is converted to an initial rule by regarding all the conditions appearing in the path as the conjunctive rule antecedents while regarding the class label $Y(x_a)$ $x_a \in D$, held by the leaf as a rule consequence. Tree pruning is then done by using a greedy elimination rule which removes antecedents that are not sufficiently discriminatory. The rule set is then further refined by the way of the minimum description length (MDL) principle [199] to remove those rules that do not contribute to the accuracy of the tree.

Appendix C: Correlation Study between histologic and proteomic biomarkers

Recently, studies to connect morphometric features in pathology with molecular data has become of particular interest for our understanding of disease [200]. Furthermore, discovery of correlated histomorphomic and proteomic features could allow for the identification of proteins via an image-based histologic signature.

Following removal of the prostate via radical prostatectomy, QH features extracted from cancer regions annotated by a pathologist can be correlated with corresponding protein expression levels taken from the same regions.

Investigating correlated QH-protein pairs in biochemical recurrence cohorts

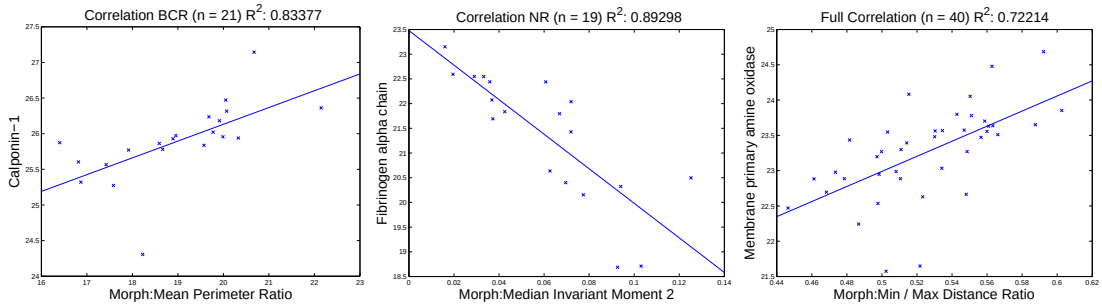
We examined Spearman's rank correlation between 242 QH and 650 protein expression pairs under three cohorts:

1. linked CaP patients ($n = 21$) who have experienced BCR following RP
2. linked CaP patients ($n = 19$) who have not experienced BCR following RP.
3. all linked CaP patients ($n = 40$) who have undergone RP

Investigation into QH-protein correlations in these cohorts may provide valuable biological insight towards the phenotypical and molecular mechanisms involved in biochemical recurrence and progression of prostate cancer.

The breakdown in distribution of patients for this study is shown in Figure 5.5.

In Figure 8.1, we show the top correlated QH-protein expression pairs for each of the cohorts investigated. High correlation was discovered across all three cohorts and the overall results are shown in Chapter 8.



(a) (b) (c)

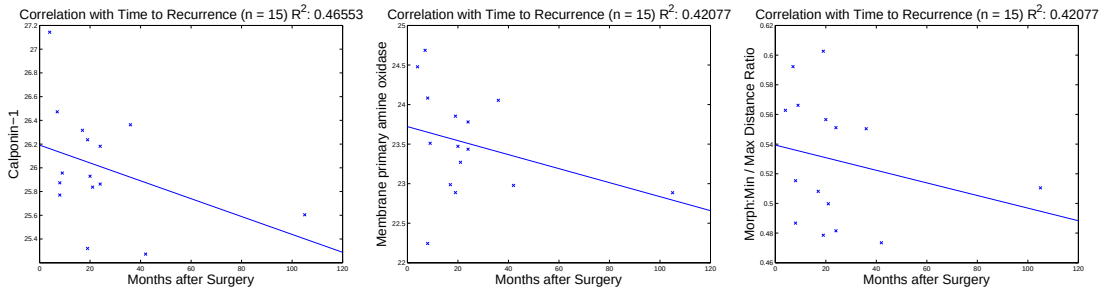
Figure 8.1: Top correlated QH-protein expression pair for (a) patients who experienced BCR, (b) patients who did not experience BCR, (c) all patients undergoing radical prostatectomy

Discussion of top correlated QH-protein pairs across CaP patients with failed radical prostatectomies

Calponin-1 is a thin flament-associate protein involved in the regulation of smooth muscle contraction[1]. Tuxhorn et al. [201] found that while calponin staining was found in normal tissue, there was decreased staining of calponin found in Gleason 3 prostate cancer tissue ($p < 0.001$). Ramaswamy et al [202] noted in a study of molecular signatures associated with metastasis, that many of the gene-expression signatures were derived from non-epithelial components of the tumor, as is the case with calponin.

Its positive correlation with the perimeter to area ratio of the gland is related to a breakdown in the glands. Smaller glands have a greater perimeter to area ratio, suggesting that loss of calponin is related to the reduced structural integrity of the glands. Involvement in the BCR cohort further suggests that calponin may be essential in preventing extracapsular spread of CaP. Figure 8.2 demonstrates the ability of Calponin-1 to predict time to recurrence via a $R^2 = 0.4655$ Spearman correlation.

[1]Information found via the UniProt (Universal Protein Resource) repository http://www.uniprot.org/uniprot/P51911



(a) (b) (c)

Figure 8.2: Predictive value of (a) Calponin-1, (b) membrane primary amine oxidase, and (c) Min/Max distance ratio of glands is shown via correlation with time to recurrence following radical prostatectomy.

Discussion of top correlated QH-protein pairs across CaP patients with successful radical prostatectomies

Fibrinogen has a dual functionality in yielding monomers that become fibrin as well a role in platelet aggregation[2]. Fragments of fibrinopeptide A and fibrinogen alpha-chain have been noted for the identification of cancer specific features, including prostate cancer. [203]. The non-specificity of this fragment for prostate cancer versus other cancers and its inclusion in the NR cohort suggests that this protein is not a discriminator of the progression of CaP but rather a discriminator of cancer and normal tissue.

Discussion of top correlated QH-protein pair across all patients undergoing radical prostatectomy

Experts at the Hospital of University of Pennsylvania (HUP) have suggested that there may be biological significance of membrane primary amine oxidase based on its appearance as a high correlation QH-protein pair and the glandular patterns found in histological staining of membrane primary amine oxidase. This protein is also known by copper amine oxidase, semicarbazide-sensitive amine oxidase (SSAO), VP97, or vascular adhesion protein 1, and plays a role in monoamine oxidase activity[3].

---

[2]Information found via the UniProt (Universal Protein Resource) repository http://www.uniprot.org/uniprot/P02671

[3]Information found via the UniProt (Universal Protein Resource) repository http://www.uniprot.org/uniprot/Q16853

Monoamine oxidase A has been found to be one of the most highly differentially overexpressed genes between high Gleason grade 4 and 5 and Gleason 3, suggesting a role in the progression of prostate cancer [204, 205]. Furthermore, it has been found to be positively correlated with preoperative PSA levels and high grade Gleason 4 and 5 prostate cancers [206]. Its significance with regards to the min/max distance ratio, a descriptor of the variance in glandular morphology, suggests that this pattern could be used to identify presence of primary membrane amine oxidase via analysis of routine H& E stained whole slides.

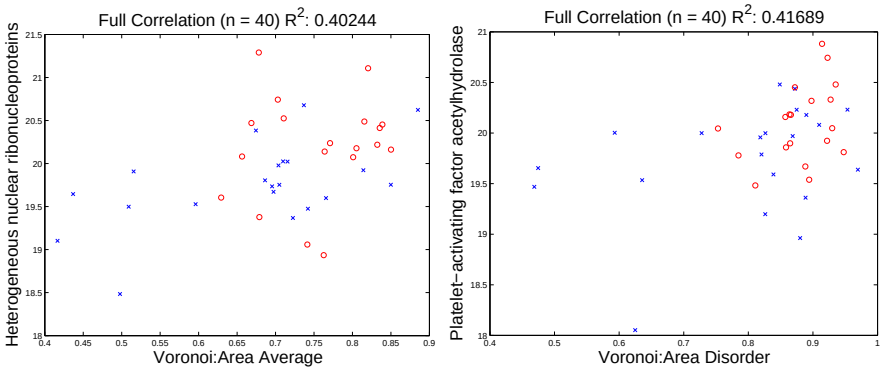
Summary of correlation study

It is interesting to note that the top correlated QH-protein pairs in these cohorts were related to glandular morphology, where the shape of the glands are the hallmark for prostate cancer and the basis of Gleason grading. The results of this correlation study suggest that in a preliminary cohort, we were able to capture previously studied proteins via quantitative histomorphometry via a high correlation coefficient $R^2 > 0.8$ for BCR and NR cohorts, and $R^2 > 0.7$ for the entire study as show in Table 8.3.

To summarize:

- 15 QH-protein pairs found with correlation $R^2 > 0.8$ across BCR cases only
- 65 QH-protein pairs found with correlation $R^2 > 0.8$ across NR cases only
- 1 QH-protein pair found with correlation $R^2 > 0.7$ across all 40 CaP cases

The top QH-protein pair found across all studies (Morph: Min/Max Distance Ratio and SSAO expression) showed not only correlation to each other, but was also correlated with time to recurrence following surgery as illustrated in Figure 8.2. Morph: Min/Max Distance Ratio showed a Spearman correlation of $R^2 = 0.4118$ and while SSAO showed $R^2 = 0.4208$ with time to recurrence.



(a) (b)

Figure 8.3: The most correlated of the statistically significantly ($p < 0.05$) predictive QH-protein expression pairs are shown via (a) Wilcoxon Rank Sum Test and (b) Student's t-test.

Investigation of predictive QH-protein pairs

To identify predictive pairs, QH and protein expression features found to show statistically significant ($p < 0.05$) differences between BCR and NR patients via the Wilcoxon Rank Sum test were correlated in order to mine potentially predictive QH-protein expression pairs. Preliminary work towards QH-based predictors of informative proteins do not appear to be within reach thus far. The best correlation values found within the significant cohort were found to have moderate correlation coefficient of approximately 0.4 as shown in Figure 8.3. Future work will aim to investigate these results on a cohort which shows greater differentiation (ie. Gleason 3 versus Gleason 4 and 5), as there were relatively few features that were statistically significantly predictive of BCR.

Appendix D: Comprehensive List of Quantitative Histomorphometry and Proteomic Markers

Table 8.1: Robust feature selection of quantitative histomorphometric features via leave-one-out cross-validated Student's t-test, Wilcoxon Rank Sum, and an intersection of the two significance tests $p < 0.05$.

Student's t-test
1: Voronoi:Area Disorder
2: Voronoi:Perimeter Disorder
3: Voronoi:Chord Disorder
4: Arch:Standard Deviation Nearest Neighbors in a 20 Pixel Radius
5: Arch:Standard Deviation Nearest Neighbors in a 30 Pixel Radius
6: CGT:mean tensor contrast energy
7: CGT:mean tensor contrast ave
Wilcoxon Rank Sum test
1: Arch:Standard Deviation Nearest Neighbors in a 20 Pixel Radius
2: Morph:Min / Max Invariant Moment 2
3: CGT:mean tensor contrast var
4: CGT:mean tensor intensity variance
5: CGT:range tensor entropy
6: CGT:range tensor energy
7: CGT:range tensor correlation
8: CGT:range tensor information measure1
9: CGT:range tensor information measure2
Intersection of Student's t-test and WRST
1: Arch:Standard Deviation Nearest Neighbors in a 20 Pixel Radius

Table 8.2: Robust feature selection of proteins via leave-one-out cross-validated Student's t-test, Wilcoxon Rank Sum, and an intersection of the two significance tests $p < 0.05$.

Student's t-test
1: Protein disulfide-isomerase A6
2: T-complex protein 1 subunit delta
3: Nidogen-1
4: ADP-ribosylation factor 3;ADP-ribosylation factor 1
5: Protein disulfide-isomerase
6: Glutathione S-transferase omega-1
7: Serine/arginine-rich splicing factor 3
8: Ras-related protein Rab-5C
9: ATP-dependent RNA helicase DDX3X;ATP-dependent RNA helicase DDX3Y
10: 40S ribosomal protein S17;40S ribosomal protein S17-like
11: Serine/arginine-rich splicing factor 7
12: 60S ribosomal protein L27
13: Proteasome subunit alpha type-4;Proteasome subunit alpha type
14: Collagen alpha-1(VIII) chain;Vastatin
Wilcoxon Rank Sum test
1: Protein disulfide-isomerase A6
2: T-complex protein 1 subunit delta
3: ADP-ribosylation factor 3;ADP-ribosylation factor 1
4: Protein disulfide-isomerase
5: Ras GTPase-activating-like protein IQGAP2
6: T-complex protein 1 subunit beta
7: Ras-related protein Rab-5C
8: ATP-dependent RNA helicase DDX3X;ATP-dependent RNA helicase DDX3Y
9: 40S ribosomal protein S17;40S ribosomal protein S17-like
10: Serine/arginine-rich splicing factor 7
11: Tubulin alpha-1A chain;Tubulin alpha-3C/D chain;Tubulin alpha-3E chain
12: Laminin subunit alpha-4
13: Collagen alpha-1(VIII) chain;Vastatin
14: Tubulin–tyrosine ligase-like protein 12
Intersection of Student's t-test and WRST
1: Protein disulfide-isomerase A6
2: T-complex protein 1 subunit delta
3: ADP-ribosylation factor 3;ADP-ribosylation factor 1
4: Protein disulfide-isomerase
5: Ras-related protein Rab-5C
6: ATP-dependent RNA helicase DDX3X;ATP-dependent RNA helicase DDX3Y
7: 40S ribosomal protein S17;40S ribosomal protein S17-like
8: Serine/arginine-rich splicing factor 7
9: Collagen alpha-1(VIII) chain;Vastatin

Table 8.3: Spearman's Rank correlation of quantitative histomorphometric features and protein expressions for the three previously defined cohorts

	QH feature	Protein	Correlation
	Cohort 1: BCR ($n = 21$)		
1:	Morph:Mean Perimeter Ratio	Calponin-1	0.834
2:	Morph:Min / Max Smoothness	Galectin-1	0.832
3:	GSG:Number of Edges	Cytoplasmic dynein 1 heavy chain 1	0.832
4:	Morph:Mean Distance Ratio	Calponin-1	0.829
5:	Voronoi:Area Minimum / Maximum	Acetyl-CoA acetyltransferase, mitochondrial	0.826
6:	Morph:Mean Area Ratio	Calponin-1	0.825
7:	CGT:range tensor contrast energy	Nucleoside diphosphate kinase	0.820
8:	Voronoi:Perimeter Standard Deviation	Beta-2-glycoprotein 1	0.817
9:	Morph:Median Invariant Moment 2	Calponin-1	0.812
10:	Morph:Mean Variance of Distance	Calponin-1	0.808
11:	Haralick:mean intensity intensity variance	Mitochondrial 2-oxoglutarate/malate carrier protein	0.808
12:	Morph:Min / Max Standard Deviation of Distance	Spectrin alpha chain, brain	0.806
13:	Morph:Min / Max Variance of Distance	Spectrin alpha chain, brain	0.806
14:	Morph:Standard Deviation Area Ratio	Copine-3	0.806
15:	Morph:Median Perimeter Ratio	Hemoglobin subunit beta	0.801
	Cohort 2: NR ($n = 19$)		
1:	Morph:Median Invariant Moment 2	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.893
2:	Morph:Mean Fourier Descriptor 7	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.889
3:	Morph:Mean Fourier Descriptor 7	Biglycan	0.886
4:	Morph:Median Invariant Moment 6	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.885
5:	Morph:Mean Long/Short Distance Ratio	Biglycan	0.884
6:	Morph:Median Fourier Descriptor 7	Prelamin-A/C;Lamin-A/C	0.877
7:	Morph:Mean Fourier Descriptor 7	Puromycin-sensitive aminopeptidase	0.858
8:	Morph:Median Long/Short Distance Ratio	Biglycan	0.856
9:	Morph:Median Invariant Moment 2	Biglycan	0.854
10:	Morph:Median Long/Short Distance Ratio	Serotransferrin	0.854
11:	Morph:Median Invariant Moment 1	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.849
12:	Morph:Median Invariant Moment 2	Puromycin-sensitive aminopeptidase	0.847
13:	CGT:range tensor contrast energy	Ras-related protein Rab-7a	0.846
14:	Morph:Median Standard Deviation of Distance	Endoplasmic reticulum resident protein 29	0.842
15:	Morph:Median Variance of Distance	Endoplasmic reticulum resident protein 29	0.842
16:	Morph:Mean Long/Short Distance Ratio	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.840
17:	Morph:Mean Long/Short Distance Ratio	Serotransferrin	0.837
18:	Morph:Standard Deviation Fourier Descriptor 1	Prelamin-A/C;Lamin-A/C	0.833
19:	Morph:Median Standard Deviation of Distance	Lysosome-associated membrane glycoprotein 2	0.833
20:	Morph:Median Variance of Distance	Lysosome-associated membrane glycoprotein 2	0.833
21:	Morph:Median Standard Deviation of Distance	Prostatic acid phosphatase;PAPf39	0.832
22:	Morph:Median Variance of Distance	Prostatic acid phosphatase;PAPf39	0.832
23:	Morph:Median Perimeter Ratio	Lysosome-associated membrane glycoprotein 2	0.831
24:	Morph:Median Fourier Descriptor 7	Puromycin-sensitive aminopeptidase	0.830
25:	Morph:Mean Fourier Descriptor 7	Prelamin-A/C;Lamin-A/C	0.830
26:	Morph:Standard Deviation Fourier Descriptor 2	Vitronectin;Vitronectin V65 subunit;Vitronectin V10 subunit;Somatomedin-B	0.830
27:	Morph:Min / Max Fourier Descriptor 10	Keratin, type II cytoskeletal 2 epidermal	0.828
28:	Morph:Standard Deviation Invariant Moment 5	Interleukin enhancer-binding factor 2	0.828
29:	Morph:Median Invariant Moment 1	Biglycan	0.826
30:	Morph:Mean Invariant Moment 1	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.825
31:	Morph:Median Fourier Descriptor 7	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.823
32:	Morph:Median Invariant Moment 2	Serotransferrin	0.823
33:	Morph:Standard Deviation Fourier Descriptor 2	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.823
34:	Morph:Median Perimeter Ratio	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.822
35:	Morph:Standard Deviation Fourier Descriptor 2	Filamin-B	0.821
36:	Morph:Median Perimeter Ratio	Trifunctional enzyme subunit beta, mitochondrial;3-ketoacyl-CoA thiolase	0.820
37:	Morph:Median Smoothness	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.819
38:	Morph:Mean Fourier Descriptor 7	Serotransferrin	0.818
39:	Morph:Standard Deviation Fourier Descriptor 6	T-complex protein 1 subunit epsilon	0.818
40:	Morph:Standard Deviation Fourier Descriptor 6	Macrophage migration inhibitory factor	0.818
41:	Morph:Median Invariant Moment 2	Prostate-specific antigen	0.816
42:	GSG:mean edge length	Lysosome-associated membrane glycoprotein 2	0.816
43:	Morph:Median Long/Short Distance Ratio	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.815
44:	Morph:Median Fourier Descriptor 4	Lysosome-associated membrane glycoprotein 2	0.813
45:	Morph:Standard Deviation Variance of Distance	Lamin-B2	0.812
46:	Morph:Standard Deviation Fourier Descriptor 5	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.812
47:	Morph:Median Standard Deviation of Distance	Retinol dehydrogenase 11	0.812
48:	Morph:Median Variance of Distance	Retinol dehydrogenase 11	0.812
49:	Arch:Standard Deviation Nearest Neighbors in a 10 Pixel Radius	T-complex protein 1 subunit epsilon	0.811
50:	Morph:Median Invariant Moment 6	Biglycan	0.810
51:	Morph:Mean Invariant Moment 4	T-complex protein 1 subunit epsilon	0.809
52:	Morph:Median Invariant Moment 4	Prelamin-A/C;Lamin-A/C	0.807
53:	Morph:Mean Distance Ratio	Lysosome-associated membrane glycoprotein 2	0.807
54:	Morph:Mean Invariant Moment 6	rRNA 2-O-methyltransferase fibrillarin	0.807
55:	CGT:range tensor contrast ave	rRNA 2-O-methyltransferase fibrillarin	0.807
56:	Morph:Median Perimeter Ratio	Prelamin-A/C;Lamin-A/C	0.806
57:	Morph:Standard Deviation Perimeter Ratio	Isocitrate dehydrogenase [NADP], mitochondrial;Isocitrate dehydrogenase [NADP]	0.805
58:	Morph:Median Invariant Moment 6	Vitronectin;Vitronectin V65 subunit;Vitronectin V10 subunit;Somatomedin-B	0.804
59:	Morph:Standard Deviation Fourier Descriptor 1	Fibrinogen alpha chain;Fibrinopeptide A;Fibrinogen alpha chain	0.804
60:	Morph:Median Long/Short Distance Ratio	Complement C3	0.802
61:	Morph:Median Invariant Moment 5	Puromycin-sensitive aminopeptidase	0.802
62:	Arch:Avg. Nearest Neighbors in a 20 Pixel Radius	Phosphatidylethanolamine-binding protein 1	0.802
63:	Morph:Median Invariant Moment 2	Lysosome-associated membrane glycoprotein 2	0.802
64:	Morph:Mean Perimeter Ratio	Trifunctional enzyme subunit beta, mitochondrial;3-ketoacyl-CoA thiolase	0.802
65:	Morph:Median Fourier Descriptor 7	Vimentin	0.800
	Cohort 3: All cases ($n = 40$)		
1:	Morph:Min / Max Distance Ratio	Membrane primary amine oxidase	0.722

References

[1] George Lee, Carlos Rodriguez, and Anant Madabhushi. Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 5(3):368–384, 2008.

[2] George Lee and Anant Madabhushi. Adaptive dimensionality reduction with semi-supervision (address) for classifying multi-attribute biomedical data. *Neurocomputing*, 2014. (Under Review).

[3] George Lee, Rachel Sparks, Sahirzeeshan Ali, Natalie N.C. Shih, Michael D. Feldman, Elaine Spangler, Timothy Rebbeck, John E. Tomaszewski, and Anant Madabhushi. Co-occurring gland tensors in localized subgraphs: Quantitative histomorphometry for postoperative prediction of biochemical recurrence in prostate cancer patients with intermediate-risk gleason scores.

[4] George Lee, Asha Singanamalli, Haibo Wang, Michael D. Feldman, Stephen R. Master, Natalie N.C. Shih, John E. Tomaszewski, and Anant Madabhushi. Supervised multi-view canonical correlation analysis (smvcca): Application towards an integrated histologic and proteomic biomarker for predicting recurrent prostate cancer.

[5] George Lee, Scott Doyle, James Monaco, Anant Madabhushi, Michael D Feldman, Stephen R Master, and John E Tomaszewski. A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: Preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 77–80. IEEE, 2009.

[6] George Lee and Anant Madabhushi. Semi-supervised graph embedding scheme with active learning (ssgeal): classifying high dimensional biomedical data. In *Pattern Recognition in Bioinformatics*, pages 207–218. Springer, 2010.

[7] P Tiwari, SE Viswanath, G Lee, and A Madabhushi. Multi-modal data fusion schemes for integrated classification of imaging and non-imaging biomedical data. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 165–168, 2011.

[8] George Lee, Rachel Sparks, Sahirzeeshan Ali, Anant Madabhushi, Michael D Feldman, SR Master, N Shih, and JE Tomaszewski. Co-occurring gland tensors in localized cluster graphs: Quantitative histomorphometry for predicting biochemical recurrence for intermediate grade prostate cancer. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 113–116. IEEE, 2013.

[9] George Lee, Sahirzeeshan Ali, Robert Veltri, Jonathan I Epstein, Christhunesa Christudass, and Anant Madabhushi. Cell orientation entropy (core): Predicting biochemical recurrence from prostate cancer tissue microarrays. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 396–403. Springer, 2013.

[10] Joshua Tenenbaum, Vin de Silva, et al. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2322, 2000.

[11] Jeffrey H Burkhardt, Mark S Litwin, Christopher M Rose, Roy J Correa, Jonathan H Sunshine, Christopher Hogan, and James A Hayman. Comparing the costs of radiation therapy and radical prostatectomy for the initial treatment of early-stage prostate cancer. *Journal of clinical oncology*, 20(12):2869–2875, 2002.

[12] F. Pinto et al. Clinical and pathological characteristics of patients presenting with biochemical progression after radical retropubic prostatectomy for pathologically organ-confined prostate cancer. *Urol Int*, 76(3):202–208, 2006.

[13] BJ Trock et al. Prostate cancer-specific survival following salvage radiotherapy vs observation in men with biochemical recurrence after radical prostatectomy. *JAMA*, 299(23):2760–2769, Jun 2008.

[14] AV D'Amico et al. Preoperative psa velocity and the risk of death from prostate cancer after radical prostatectomy. *N Engl J Med*, 351(2):125–135, Jul 2004.

[15] DF Gleason. Classification of prostatic carcinomas. *Cancer Chemother Rep*, 50(3):125–8., 1966.

[16] Jonathan I Epstein. An update of the gleason grading system. *Journal Of Urology (the)*, 183(2):433, 2010.

[17] Jonathan I. Epstein. Update on the gleason grading system. *Ann Pathol*, 31(5 Suppl):S20–S26, Nov 2011.

[18] C. R. Pound, A. W. Partin, J. I. Epstein, and P. C. Walsh. Prostate-specific antigen after anatomic radical retropubic prostatectomy. patterns of recurrence and cancer control. *Urol Clin North Am*, 24(2):395–406, May 1997.

[19] Stephen J. Freedland, George S. Csathy, Frederick Dorey, and William J. Aronson. Percent prostate needle biopsy tissue with cancer is more predictive of biochemical failure or adverse pathology after radical prostatectomy than prostate specific antigen or gleason score. *J Urol*, 167(2 Pt 1):516–520, Feb 2002.

[20] JR Stark et al. Gleason score and lethal prostate cancer: does $3 + 4 = 4 + 3$? *J Clin Oncol*, 27(21):3459–3464, Jul 2009.

[21] M Han et al. Biochemical (prostate specific antigen) recurrence probability following radical prostatectomy for clinically localized prostate cancer. *J Urol*, 169(2):517–523, Feb 2003.

[22] WC Allsbrook, Jr, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein. Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol*, 32(1):81–88, Jan 2001.

[23] Michael W Kattan, Thomas M Wheeler, and Peter T Scardino. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of Clinical Oncology*, 17(5):1499–1499, 1999.

[24] Andrew J Stephenson, Peter T Scardino, James A Eastham, Fernando J Bianco, Zohar A Dotan, Christopher J DiBlasio, Alwyn Reuther, Eric A Klein, and Michael W Kattan. Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of Clinical Oncology*, 23(28):7005–7012, 2005.

[25] AI Hinev, D Anakievski, N Kolev, V Marianovski, and V Hadjiev. Validation of pre-and postoperative nomograms used to predict the pathological stage and prostate cancer recurrence after radical prostatectomy: a multi-institutional study. *Journal of BU ON.: official journal of the Balkan Union of Oncology*, 16(2):316, 2011.

[26] Matthew R Cooperberg, Stephen J Freedland, David J Pasta, Eric P Elkin, Joseph C Presti, Christopher L Amling, Martha K Terris, William J Aronson, Christopher J Kane, and Peter R Carroll. Multiinstitutional validation of the ucsf cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy. *Cancer*, 107(10):2384–2391, 2006.

[27] Matthew R. Cooperberg, Joan F. Hilton, and Peter R. Carroll. The capra-s score: A straightforward tool for improved prediction of outcomes after radical prostatectomy. *Cancer*, 117(22):5039–5046, Nov 2011.

[28] Carlos Cordon-Cardo, Angeliki Kotsianti, David A. Verbel, Mikhail Teverovskiy, Paola Capodieci, Stefan Hamann, Yusuf Jeffers, Mark Clayton, Faysal Elkhettabi, Faisal M. Khan, Marina Sapir, Valentina Bayer-Zubek, Yevgen Vengrenyuk, Stephen Fogarsi, Olivier Saidi, Victor E. Reuter, Howard I. Scher, Michael W. Kattan, Fernando J. Bianco, Thomas M. Wheeler, Gustavo E. Ayala, Peter T. Scardino, and Michael J. Donovan. Improved prediction of prostate cancer recurrence through systems pathology. *J Clin Invest*, 117(7):1876–1883, Jul 2007.

[29] Shahrokh F Shariat, Jochen Walz, Claus G Roehrborn, Alexandre R Zlotta, Paul Perrotte, Nazareno Suardi, Fred Saad, and Pierre I Karakiewicz. External validation of a biomarker-based preoperative nomogram predicts biochemical recurrence after radical prostatectomy. *Journal of Clinical Oncology*, 26(9):1526–1531, 2008.

[30] S. Doyle et al. A class balanced active learning scheme that accounts for minority class problems: Applications to histopathology. In *MICCAI*, 2009.

[31] Scott Doyle, Michael D Feldman, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. *BMC bioinformatics*, 13(1):282, 2012.

[32] Michael J. Donovan, Stefan Hamann, Mark Clayton, Faisal M. Khan, Marina Sapir, Valentina Bayer-Zubek, Gerardo Fernandez, Ricardo Mesa-Tejada, Mikhail Teverovskiy, Victor E. Reuter, Peter T. Scardino, and Carlos Cordon-Cardo. Systems pathology approach for the prediction of prostate cancer progression after radical prostatectomy. *J Clin Oncol*, 26(24):3923–3929, Aug 2008.

[33] E David Crawford, Kyle O Rove, Edouard J Trabulsi, Junqi Qian, Krystyna P Drewnowska, Jed C Kaminetsky, Thomas K Huisman, Mark L Bilowus, Sheldon J Freedman, W Lloyd Glover Jr, et al. Diagnostic performance of pca3 to detect prostate cancer in men with increased prostate specific antigen: a prospective study of 1,962 cases. *The Journal of urology*, 2012.

[34] Raymond A Clarke, Horst J Schirra, James W Catto, Martin F Lavin, and Robert A Gardiner. Markers for detection of prostate cancer. *Cancers*, 2(2):1125–1154, 2010.

[35] Emanuel F Petricoin, 3rd, David K. Ornstein, Cloud P. Paweletz, Ali Ardekani, Paul S. Hackett, Ben A. Hitt, Alfredo Velassco, Christian Trucco, Laura Wiegand, Kamillah Wood, Charles B. Simone, Peter J. Levine, W Marston Linehan, Michael R. Emmert-Buck, Seth M. Steinberg, Elise C. Kohn, and Lance A. Liotta. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst*, 94(20):1576–1578, Oct 2002.

[36] M.E. Wright et al. Mass spectrometry-based expression profiling of clinical prostate cancer. *Mol Cell Proteomics*, 4(4):545–554, Apr 2005.

[37] Adam M. Hawkridge and David C. Muddiman. Mass spectrometry-based biomarker discovery: toward a global proteome index of individuality. *Annu Rev Anal Chem (Palo Alto Calif)*, 2:265–277, 2009.

[38] Ali Khatami, Jonas Hugosson, Wanzhong Wang, and Jan-Erik Damber. Ki-67 in screen-detected, low-grade, low-stage prostate cancer, relation to prostate-specific antigen doubling time, gleason score and prostate-specific antigen relapse after radical prostatectomy. *Scand J Urol Nephrol*, 43(1):12–18, 2009.

[39] Sanaz Piran, Peter Liu, Ana Morales, and Ray E. Hershberger. Where genome meets phenome: rationale for integrating genetic and protein biomarkers in the diagnosis and management of dilated cardiomyopathy and heart failure. *J Am Coll Cardiol*, 60(4):283–289, Jul 2012.

[40] A. Krishan, A. Oppenheimer, W. You, R. Dubbin, D. Sharma, and B. L. Lokeshwar. Flow cytometric analysis of androgen receptor expression in human prostate tumors and benign tissues. *Clin Cancer Res*, 6(5):1922–1930, May 2000.

[41] Philippe L. Bedard, Stella Mook, Martine J. Piccart-Gebhart, Emiel T. Rutgers, Laura J. Van't Veer, and Fatima Cardoso. Mammaprint 70-gene profile quantifies the likelihood of recurrence for early breast cancer. *Expert Opin Med Diagn*, 3(2):193–205, Mar 2009.

[42] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung

Park, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

[43] Chungyeul Kim and Soonmyung Paik. Gene-expression-based prognostic assays for breast cancer. *Nat Rev Clin Oncol*, 7(6):340–347, Jun 2010.

[44] C Nicole Rosenzweig, Zhen Zhang, Xiaer Sun, Lori J. Sokoll, Katherine Osborne, Alan W. Partin, and Daniel W. Chan. Predicting prostate cancer biochemical recurrence using a panel of serum proteomic biomarkers. *J Urol*, 181(3):1407–1414, Mar 2009.

[45] Aissar Eduardo Nassif and Renato Tmbara Filho. Immunohistochemistry expression of tumor markers cd34 and p27 as a prognostic factor of clinically localized prostate adenocarcinoma after radical prostatectomy. *Rev Col Bras Cir*, 37(5):338–344, Oct 2010.

[46] T. D. Veenstra, T. P. Conrads, B. L. Hood, A. M. Avellino, R. G. Ellenbogen, and R. S. Morrison. Biomarkers: mining the biofluid proteome. *Mol Cell Proteomics*, 4(4):409–418, April 2005.

[47] Bao-Ling Adam, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, O John Semmes, Paul F. Schellhammer, Yutaka Yasui, Ziding Feng, and George L Wright, Jr. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*, 62(13):3609–3614, Jul 2002.

[48] Jamal A. Al-Ruwaili, Samantha E T. Larkin, Bashar A. Zeidan, Matthew G. Taylor, Chaker N. Adra, Claire L. Aukim-Hastie, and Paul A. Townsend. Discovery of serum protein biomarkers for prostate cancer progression by proteomic analysis. *Cancer Genomics Proteomics*, 7(2):93–103, 2010.

[49] Dana Faratian, Robert G. Clyde, John W. Crawford, and David J. Harrison. Systems pathology–taking molecular pathology into a new dimension. *Nat Rev Clin Oncol*, 6(8):455–464, Aug 2009.

[50] Manfred Dietel and Reinhold Schfer. Systems pathology–or how to solve the complex problem of predictive pathology. *Virchows Arch*, 453(4):309–312, Oct 2008.

[51] Anant Madabhushi, Scott Doyle, George Lee, Ajay Basavanhally, James Monaco, Steve Masters, John Tomaszewski, and Michael Feldman. Integrated diagnostics: a conceptual framework with examples. *Clin Chem Lab Med*, 48(7):989–998, Jul 2010.

[52] Anant Madabhushi, Shannon Agner, Ajay Basavanhally, Scott Doyle, and George Lee. Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Computerized medical imaging and graphics*, 35(7):506–514, 2011.

[53] Shailja V Parikh and James A De Lemos. Biomarkers in cardiovascular disease: integrating pathophysiology into clinical practice. *The American journal of the medical sciences*, 332(4):186–197, 2006.

[54] Carlos AG Van Mieghem, Nico Bruining, Johannes A Schaar, Eugene McFadden, Nico Mollet, Filippo Cademartiri, Frits Mastik, Jurgen MR Ligthart, Gaston A Rodriguez Granillo, Marco Valgimigli, et al. Rationale and methods of the integrated biomarker and imaging study (ibis): combining invasive and non-invasive imaging with biomarkers to detect subclinical atherosclerosis and assess coronary lesion biology. *The international journal of cardiovascular imaging*, 21(4):425–441, 2005.

[55] Yukun Chen, Subramani Mani, and Hua Xu. Applying active learning to assertion classification of concepts in clinical text. *J Biomed Inform*, 45(2):265–272, Apr 2012.

[56] Harubumi Kato, Toshihide Nishimura, Norihiko Ikeda, Tesshi Yamada, Tadashi Kondo, Nagahiro Saijo, Kazuto Nishio, Junichiro Fujimoto, Masaharu Nomura, Yoshiya Oda, et al. Developments for a growing japanese patient population: facilitating new technologies for future health care. *Journal of proteomics*, 74(6):759–764, 2011.

[57] Michael J. Donovan, Faisal M. Khan, Gerardo Fernandez, Ricardo Mesa-Tejada, Marina Sapir, Valentina Bayer Zubek, Douglas Powell, Stephen Fogarasi, Yevgen Vengrenyuk, Mikhail Teverovskiy, Mark R. Segal, R Jeffrey Karnes, Thomas A. Gaffey, Christer Busch, Michael Haggman, Peter Hlavcak, Stephen J. Freedland, Robin T. Vollmer, Peter Albertsen, Jose Costa, and Carlos Cordon-Cardo. Personalized prediction of tumor response and cancer progression on prostate needle biopsy. *J Urol*, 182(1):125–132, Jul 2009.

[58] R.E. Bellman. *Adaptive Control Processes.* Princeton University Press, 1961.

[59] I Guyon and A Elisseeff. An introduction to variable and feature selection. *J. Mach Learn Res*, 3:1157–1182., 2003.

[60] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomeld, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(531):531–537, 1999.

[61] Yonghong Peng. A novel ensemble machine learning for robust microarray data classification. *Computers in Biology and Medicine*, 36(6):553–573, 2006.

[62] Chao Shi and Lihui Chen. Feature dimension reduction for microarray data analysis using locally linear embedding. In *APBC*, pages 211–217, 2005.

[63] S. D. Der, A. Zhou, B. R. Williams, and R. H. Silverman. Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 95(26):15623–15628, Dec 1998.

[64] Rosalia Maglietta, Annarita D'Addabbo, Ada Piepoli, Francesco Perri, Sabino Liuni, Graziano Pesole, and Nicola Ancona. Selection of relevant genes in cancer diagnosis based on their prediction accuracy. *Artif Intell Med*, 40(1):29–44, May 2007.

[65] Te Ming Huang and Vojislav Kecman. Gene extraction for cancer diagnosis by support vector machines–an improvement. *Artif Intell Med*, 35(1-2):185–194, 2005.

[66] Gulisa Turashvili, Jan Bouchal, Karl Baumforth, Wenbin Wei, Marta Dziechciarkova, Jiri Ehrmann, Jiri Klein, Eduard Fridman, Jozef Skarda, Josef Srovnal, Marian Hajduch, Paul Murray, and Zdenek Kolar. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer*, 7(55), 2007.

[67] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andeas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D.Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[68] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *J Comput Biol*, 7(3-4):559–583, 2000.

[69] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–267, Jan 2000.

[70] Aik Choon Tan and David Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2(3 Suppl):S75–83, 2003.

[71] Le Song, Justin Bedo, Karsten M. Borgwardt, Arthur Gretton, and Alex Smola. Gene selection via the bahsic family of algorithms. *Bioinformatics*, 23:490–498, 2007.

[72] Li Li, Wei Jiang, Xia Li, Kathy L. Moser, Zheng Guo, Lei Du, Qiuju Wang, Eric J. Topol, Qing Wang, and Shaoqi Rao. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 85:16–23, 1995.

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

[74] Huiqing Liu, Jinyan Li, and Limsoon Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform*, 13:51–60, 2002.

[75] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering of tumor and

normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci*, 96(12):6745–6750, 1999.

[76] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, , and William R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[77] Mira Park, Jae Won Lee, Jung Bok Lee, and Sueuck Heun Song. Several biplot methods applied to gene expression data. *Journal of Statistical Planning and Inferenece*, 138:500–515, 2007.

[78] Emanuel F Petricoin, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, and Lance A Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359(9306):572–577, 2002.

[79] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–2912, Mar 1999.

[80] Sanghwa Yang, Jihye Shin, Kyu Hyun Park, Hei-Cheul Jeung, Sun Young Rha, Sung Hoon Noh, Woo Ick Yang, and Hyun Cheol Chung. Molecular basis of the differences between normal and tumor tissues of gastric cancer. *Biochim Biophys Acta*, 2007.

[81] Laura J. van 't Veer1, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Ren Bernards Peter S. Linsley and, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:430–536, 2002.

[82] Scott L. Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M. Sturla, Michael Angelo, Margaret E. McLaughlin, John Y. H. Kim, Liliana C. Goumnerova, Peter M. Black, Ching Lau, Jeffrey C. Allen, David Zagzag, James M. Olson, Tom Curran, Cynthia Wetmore, Jaclyn A. Biegel, Tomaso Poggio, Shayan Mukherjee, Ryan Rifkin, Andrea Califano, Gustavo Stolovitzky, David N. Louis, Jill P. Mesirov, Eric S. Lander, and Todd R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[83] William A. Freije, F. Edmundo Castro-Vargas, Zixing Fang, Steve Horvath, Timothy Cloughesy, Linda M. Liau, Paul S. Mischel, and Stanley F. Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer Research*, 64(18):6503–6510, 2004.

[84] Margaret A. Shipp, Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C.T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich,

Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval1, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S. Lander, Jon C. Aster, and Todd R. Golub. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8:68–74, 2002.

[85] David G. Beer, Sharon L.R. Kardia, Chiang-Ching Huang, Thomas J. Giordano, Albert M. Levin, David E. Misek, Lin Lin, Guoan Chen, Tarek G. Gharib, Dafydd G. Thomas, Michelle L. Lizyness, Rork Kuick, Satoru Hayasaka, Jeremy M.G. Taylor, Mark D. Iannettoni, Mark B. Orringer, and Samir Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8:816–823, 2002.

[86] Dennis A. Wigle, Igor Jurisica, Niki Radulovich, Melania Pintilie, Janet Rossant, Ni Liu, Chao Lu, James Woodgett, Isolde Seiden, Michael Johnston, Shaf Keshavjee, Gail Darling, Timothy Winton, Bobby-Joe Breitkreutz, Paul Jorgenson, Mike Tyers, Frances A. Shepherd, and Ming Sound Tsao. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, 62:3005–3008, 2002.

[87] Jieping Ye, Tao Li, Tao Xiong, and Ravi Janardan. Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 1(4):181–190, 2004.

[88] A. Golugula et al. Supervised regularized canonical correlation analysis: integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery. *BMC bioinformatics*, 12(1):483, 2011.

[89] G Hughes. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, 14(1):55–63, 1968.

[90] Vladimir Pestov. Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 2012.

[91] Yvan Saeys, Iaki Inza, and Pedro Larraaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, Oct 2007.

[92] Peyman Jafari and Francisco Azuaje. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med Inform Decis Mak*, 6:27, 2006.

[93] Jeffrey G Thomas, James M Olson, Stephen J Tapscott, and Lue Ping Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11(7):1227–1236, 2001.

[94] Gyan Bhanot, Gabriela Alexe, Babu Venkataraghavan, and Arnold J. Levine. A robust meta-classification strategy for cancer detection from ms data. *Proteomics*, 6(2):592–604, Jan 2006.

[95] Jiangsheng Yu and Xue-Wen Chen. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. *Bioinformatics*, 21 Suppl 1:i487–i494, Jun 2005.

[96] Rianne Hupse and Nico Karssemeijer. The effect of feature selection methods on computer-aided detection of masses in mammograms. *Phys Med Biol*, 55(10):2893–2904, May 2010.

[97] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (2nd ed.* Wiley, 2000.

[98] Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.

[99] Jian J. Dai, Linh Lieu, and David Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):1–15, 2006.

[100] Kevin Dawson, Raymond L. Rodriguez, and Wasyl Malyj. Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using isomap, a nonlinear algorithm. *BMC Bioinformatics*, 6:195, 2005.

[101] Caroline Truntzer, Catherine Mercier, Jacques Estve, Christian Gautier, and Pascal Roy. Importance of data structure in comparing two dimension reduction methods for classification of microarray gene expression data. *BMC Bioinformatics*, 8(90), 2007.

[102] Anna Andersson, Tor Olofsson, David Lindgren, Bjorn Nilsson, Cecilia Ritz, Patrik Eden, Carin Lassen, Johan Rade, Magnus Fontes, Helena Morse, Jesper Heldrup, Mikael Behrendtz, Felix Mitelman Mattias Hoglund, Bertil Johansson, and Thoas Fioretos. Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations. *PNAS*, 102(52):19069–19074, 2005.

[103] Yi Zhu, Rong Wu, Navneet Sangha, Chul Yoo, Kathleen R. Cho, Kerby A. Shedden, Hidetaka Katabuchi, and David M. Lubman. Classifications of ovarian cancer tissues by proteomic patterns. *Proteomics*, 6:5846–5856, 2006.

[104] Marco A. Mendez, Christian Hodar, Chris Vulpe, and Mauricio Gonzalez. Discriminant analysis to evaluate clustering of gene expression data. *Federation of European Biochemical Societies*, 522:24–28, 2002.

[105] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[106] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–899, 2006.

[107] T.F. Cox M.A.A. Cox. *Multidimensional Scaling.* Chapman and Hall., 2001.

[108] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 22(8):888–905, 2000.

[109] Sam Roweis and Lawrence Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[110] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[111] Anant Madabhushi, Jianbo Shi, Mark Rosen, John E. Tomaszeweski, and Michael D. Feldman. Graph embedding to improve supervised classification and novel class detection: Application to prostate cancer. In *MICCAI*, pages 729–737, 2005.

[112] Pallavi Tiwari, Anant Madabhushi, and Mark Rosen. A hierarchical unsupervised spectral clustering scheme for detection of prostate cancer from magnetic resonance spectroscopy (mrs). In *MICCAI*, volume 2, pages 278–286, 2007: 278-286.

[113] S. Weng, C. Zhang, Z. Lin, and X. Zhang. Mining the structural knowledge of high-dimensional medical data using isomap. *Med. Biol. Eng. Comput.*, 43:410–412, 2005.

[114] Jens Nilsson, Thoas Fioretos, Mattias Hglund, and Magnus Fontes. Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics*, 20(6):874–880, 2004.

[115] G.J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley-IEEE, 2004.

[116] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

[117] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, pages 857–864, 2003.

[118] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[119] M. Brand. Charting a manifold. *Advances in Neural Information Processing Systems*, pages 985–992, 2003.

[120] Tuomo Kakkonen, Niko Myller, Erkki Sutinen, and Jari Timonen. Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, 11(3):275–288, 2008.

[121] Bernhard Schlkopf, Sebastian Mika, Alex Smola, Gunnar Rtsch, and Klaus-Robert Mller. Kernel pca pattern reconstruction via approximate pre-images. 1998.

[122] Z. Zhang and H. Zha. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. *SIAM Journal on Scientific Computing*, 26:313, 2004.

[123] P. Demartines and J. Herault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on neural networks*, 8(1):148–154, 1997.

[124] T. Rohlfing et al. Information fusion in biomedical image analysis: Combination of data vs. combination of interpretations. *IPMI*, 3565:150–161, 2005.

[125] Pallavi Tiwari, John Kurhanewicz, Mark Rosen, and Anant Madabhushi. Semi supervised multi kernel (sesmik) graph embedding: identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy. *Med Image Comput Comput Assist Interv*, 13(Pt 3):666–673, 2010.

[126] Brian McFee, Carolina Galleguillos, and Gert Lanckriet. Contextual object localization with multiple kernel nearest neighbor. *IEEE Trans Image Process*, 20(2):570–585, Feb 2011.

[127] Ana L N Fred and Anil K Jain. Combining multiple clusterings using evidence accumulation. *IEEE Trans Pattern Anal Mach Intell*, 27(6):835–850, Jun 2005.

[128] Satish Viswanath and Anant Madabhushi. Consensus embedding: theory, algorithms and application to segmentation and classification of biomedical data. *BMC Bioinformatics*, 13:26, 2012.

[129] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of online learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[130] G.R.G. Lanckriet et al. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomp.*, pages 300–311, 2004.

[131] Darrin P Lewis, Tony Jebara, and William Stafford Noble. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics*, 22(22):2753–2760, 2006.

[132] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[133] Maurice S. Bartlett. Further aspects of the theory of multiple regression. In *Proceedings of the Cambridge Philosophical Society*, volume 34, pages 33–40. Cambridge Univ Press, 1938.

[134] Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976.

[135] Sue E Leurgans, Rana A Moyeed, and Bernard W Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 725–740, 1993.

[136] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[137] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4, 2010.

[138] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput*, 16(12):2639–2664, Dec 2004.

[139] Chenping Hou, Feiping Nie, and Yi Wu. Semi-supervised dimensionality reduction via harmonic functions. In *Modeling Decision for Artificial Intelligence*, pages 91–102. Springer, 2011.

[140] Buyue Qian and Ian Davidson. Semi-supervised dimension reduction for multi-label classification. In *AAAI*, 2010.

[141] Masashi Sugiyama, Tsuyoshi Idé, Shinichi Nakajima, and Jun Sese. Semi-supervised local fisher discriminant analysis for dimensionality reduction. *Advances in Knowledge Discovery and Data Mining*, pages 333–344, 2008.

[142] Xin Yang, Haoying Fu, Hongyuan Zha, and Jesse Barlow. Semi-supervised nonlinear dimensionality reduction. *International Conference on Machine Learning*, pages 1065–1072, 2006.

[143] Haitao Zhao. Combining labeled and unlabeled data with graph embedding. *Neurocomputing*, 69(16-18):2385–2389, 2006.

[144] Daoqiang Zhang et al. Semi-supervised dimensionality reduction. In *SIAM International Conference on Data Mining*, 2007.

[145] Jakob J Verbeek and Nikos Vlassis. Gaussian fields for semi-supervised regression and correspondence learning. *Pattern Recognition*, 39(10):1864–1875, 2006.

[146] Scott Doyle, James Monaco, Michael Feldman, John Tomaszewski, and Anant Madabhushi. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC Bioinformatics*, 12:424, 2011.

[147] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

[148] Ying Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inf. Comput. Sci.*, 44 (6):1936–1941, 2004.

[149] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[150] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artif. Intell. Res.*, 4:129–145, 1996.

[151] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[152] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, pages 999–1006, 2000.

[153] Lijun Zhang, Chun Chen, Jiajun Bu, Deng Cai, Xiaofei He, and Thomas S Huang. Active learning based on locally linear reconstruction. *IEEE Trans Pattern Anal Mach Intell*, Jan 2011.

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

[155] Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online choice of active learning algorithms. *The Journal of Machine Learning Research*, 5:255–291, 2004.

[156] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.

[157] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[158] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[159] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, 1987.

[160] Hema Raghavan, Omid Madani, and Rosie Jones. Active learning with feedback on features and instances. *The Journal of Machine Learning Research*, 7:1655–1686, 2006.

[161] RM Haralick, K Shanmugam, and I Dinstein. Textural features for image classification. *IEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

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

[163] Pallavi Tiwari, S Viswanath, J Kurhanewicz, Akshay Sridhar, and Anant Madabhushi. Multimodal wavelet embedding representation for data combination (maweric): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection. *NMR in Biomedicine*, 25(4):607–619, 2012.

[164] Liang Sun, Shuiwang Ji, and Jieping Ye. A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1024–1031. ACM, 2008.

[165] Yaqian Guo, Trevor Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

[166] Ignacio González, Sébastien Déjean, Pascal GP Martin, Alain Baccini, et al. Cca: An r package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14, 2008.

[167] Roy D Yates and David J Goodman. *Probability and stochastic processes*. John Wiley & Sons USA, 1999.

[168] Tingkai Sun and Songcan Chen. Class label versus sample label-based cca. *Applied Mathematics and computation*, 185(1):272–283, 2007.

[169] Jerome L Myers, Arnold D Well, and Robert Frederick Lorch. *Research design and statistical analysis*. Routledge, 2010.

[170] JP Monaco et al. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. *Med Image Anal*, 14(4):617–629, Aug 2010.

[171] Robert W Veltri et al. Nuclear roundness variance predicts prostate cancer progression, metastasis, and death: A prospective evaluation with up to 25 years of follow-up after radical prostatectomy. *The Prostate*, 70(12):1333–1339, 2010.

[172] Kourosh Jafari-Khouzani and Hamid Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Trans on Biomedical Engineering*, 50(6):697–704, 2003.

[173] Ali Tabesh et al. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Trans on Medical Imaging*, 26(10):1366–1378, 2007.

[174] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *ISBI*, pages 1284–1287, 2007.

[175] WA Christens-Barry and AW Partin. Quantitative grading of tissue and nuclei in prostate cancer for prognosis prediction. *Johns Hopkins Apl Technical Digest.*, 18:226–233, 1997.

[176] Cagatay Bilgin et al. Cell-graph mining for breast tissue modeling and classification. In *IEEE Eng in Med and Biol Soc, EMBS*, pages 5311–5314, 2007.

[177] Sahirzeeshan Ali, Robert Veltri, Jonathan .I. Epstein, C. Christudass, and A Madabhushi. Cell cluster graph for prediction of biochemical recurrence in prostate cancer patients from tissue microarrays. In *Proc. SPIE 8676, Medical Imaging*, Digital Pathology, 2013.

[178] Sahirzeeshan Ali and Anant Madabhushi. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. *IEEE Transactions on Medical Imaging*, 31(7):1448–1460, 2012.

[179] R. Sparks and A. Madabhushi. Gleason grading of prostate histology utilizing manifold regularization via statistical shape model of manifolds. In *Proc. of SPIE Vol*, volume 8315, pages 83151J–1, 2012.

[180] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[181] Morten W Fagerland. t-tests, non-parametric tests, and large studiesa paradox of statistical practice? *BMC Medical Research Methodology*, 12(1):78, 2012.

[182] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[183] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50:163–170, 1966.

[184] Katherine J. Heaton and Stephen R. Master. Peptide extraction from formalin-fixed paraffin-embedded tissue. *Curr Protoc Protein Sci*, Chapter 23:Unit23.5, Aug 2011.

[185] Jacek R. Wisniewski, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. Universal sample preparation method for proteome analysis. *Nat Methods*, 6(5):359–362, May 2009.

[186] Juri Rappsilber, Matthias Mann, and Yasushi Ishihama. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using stagetips. *Nat Protoc*, 2(8):1896–1906, 2007.

[187] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

[188] Sally J Deeb, Rochelle CJ D'Souza, Jürgen Cox, Marc Schmidt-Supprian, and Matthias Mann. Super-silac allows classification of diffuse large b-cell lymphoma subtypes by their protein expression profiles. *Molecular & Cellular Proteomics*, 11(5):77–89, 2012.

[189] S. Naik et al. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *International Symposium on Biomedical Imaging*, pages 284–287. IEEE, 2008.

[190] Rachel Sparks and Anant Madabhushi. Novel morphometric based classification via diffeomorphic based shape representation using manifold learning. *Med Image Comput Comput Assist Interv*, 13(Pt 3):658–665, 2010.

[191] J.P. Monaco, J.E. Tomaszewski, M.D. Feldman, et al. Detection of prostate cancer from whole-mount histology images using markov random fields. In *MIAAB*, 2008.

[192] Juha Reunanen. Overfitting in making comparisons between variable selection methods. *The Journal of Machine Learning Research*, 3:1371–1382, 2003.

[193] F. Markowetz and R. Spang. Molecular diagnosis. classification, model selection and performance evaluation. *Methods Inf Med*, 44(3):438–443, 2005.

[194] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[195] Yair Goldberg and Michael R Kosorok. Support vector regression for right censored data. *arXiv preprint arXiv:1202.5130*, 2012.

[196] Faisal M Khan and Valentina Bayer Zubek. Support vector regression for censored data (svrc): a novel tool for survival analysis. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 863–868. IEEE, 2008.

[197] Kristina Schwamborn, Ren C. Krieg, Marcus Reska, Gerhard Jakse, Ruth Knuechel, and Axel Wellmann. Identifying prostate carcinoma by maldi-imaging. *Int J Mol Med*, 20(2):155–159, Aug 2007.

[198] J.R. Quinlan. Bagging, boosting, and c4.5. *AAAI/IAAI*, 1:725–730, 1996.

[199] J.R. Quinlan and R.L. Rivest. Inferring decision trees using the minimum description length principle. *Inform. Comput.*, 80(3):227–248, 1989.

[200] Jun Kong, Lee AD Cooper, Fusheng Wang, Jingjing Gao, George Teodoro, Lisa Scarpace, Tom Mikkelsen, Matthew J Schniederjan, Carlos S Moreno, Joel H Saltz, et al. Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *PLOS ONE*, 8(11):e81049, 2013.

[201] Jennifer A Tuxhorn, Gustavo E Ayala, Megan J Smith, Vincent C Smith, Truong D Dang, and David R Rowley. Reactive stroma in human prostate cancer induction of myofibroblast phenotype and extracellular matrix remodeling. *Clinical Cancer Research*, 8(9):2912–2923, 2002.

[202] Sridhar Ramaswamy, Ken N Ross, Eric S Lander, and Todd R Golub. A molecular signature of metastasis in primary solid tumors. *Nature genetics*, 33(1):49–54, 2002.

[203] MA Karpova, SA Moshkovskii, IY Toropygin, and AI Archakov. Cancer-specific maldi-tof profiles of blood serum and plasma: biological meaning and perspectives. *Journal of proteomics*, 73(3):537–551, 2010.

[204] Vincent Flamand, Hongjuan Zhao, and Donna M Peehl. Targeting monoamine oxidase a in advanced prostate cancer. *Journal of cancer research and clinical oncology*, 136(11):1761–1771, 2010.

[205] Lawrence True, Ilsa Coleman, Sarah Hawley, Ching-Ying Huang, David Gifford, Roger Coleman, Tomasz M Beer, Edward Gelmann, Milton Datta, Elahe Mostaghel, et al. A molecular correlate to the gleason grading system for prostate adenocarcinoma. *Proceedings of the National Academy of Sciences*, 103(29):10991–10996, 2006.

[206] Donna M Peehl, Marc Coram, Htet Khine, Stephen Reese, Rosalie Nolley, and Hongjuan Zhao. The significance of monoamine oxidase-a expression in high grade prostate cancer. *The Journal of urology*, 180(5):2206–2211, 2008.