SEMI-SUPERVISED TRANSDUCTIVE REGRESSION FOR SURVIVAL ANALYSIS IN MEDICAL PROGNOSTICS

By

FAISAL M. KHAN

A dissertation submitted to the

Graduate School - New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Casimir A. Kulikowski

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2016

ABSTRACT OF THE DISSERTATION

Semi-Supervised Transductive Regression for Survival Analysis in Medical Prognostics

By FAISAL M. KHAN

Dissertation Director:

Casimir A. Kulikowski

The central challenge in predictive modeling for survival analysis in medical prognostics is the management of censored observations in the data. While time-to-event predictions can be modeled as regression problems, traditional regression techniques are challenged by the censored characteristics of the data. In such problems the true target times of a majority of instances are unknown; what is known is a censored target representing some indeterminate time before the true target time. The information for most patients is incomplete and only known "up-to-a-point." Patients who have experienced the endpoint of interest (cancer recurrence, death, etc) during an often multi-year study are considered as *non-censored* or *events*. They may represent as little as 9% of the available sample. Most of the patients do not experience the endpoint or are lost to follow-up for various reasons (patient moved, died of other causes, etc.). These *censored* samples often represent most of the available sample. Modeling techniques which can correctly account for censored observations are crucial. Such censored samples can be considered

as semi-supervised targets, however most efforts in semi-supervised regression do not take into account the partial nature of unsupervised information; with samples treated as either fully labelled or unlabeled. This dissertation presents a novel transduction approach for semi-supervised survival analysis. The true target times are approximated from the censored times through transduction to improve predictive performance. The framework can be employed to transform traditional regression methods for survival analysis, or to enhance existing survival analysis algorithms for improved predictive performance. This proposed approach represents one of the first applications of semi-supervised regression to survival analysis and yields significant improvements in predictive performance for multiple applications in prostate and breast cancer prognostics.

ACKNOWLEDGEMENTS

I am eternally grateful for the encouragement, support, love and guidance of my beautiful family. My parents, Athar Mustafa Khan and Rubina Sultan. My wife Shafa Ahmed. My brother Bilal Mustafa Khan and my sisters Orubah Athar Khan and Maryam Athar Khan.

I would like to thank my advisor, Prof. Casimir A. Kulikowski for his support, guidance and understanding over the years. I truly could not have hoped for a better PhD advisor. I am also grateful to the members of my PhD Dissertation Committee, Prof. Kevin Chen, Prof. Konstantinos Michmizos and Prof. Georgios Mitsis.

I would also like to thank Prof. William M. Pottenger for the idea to apply to Rutgers and for providing me with encouragement and support over the years.

I am grateful to all my friends who have encouraged me to keep going and given me valuable advice during these years.

I would like to thank my mentors, bosses, and colleagues at Aureon Biosciences, Covance Inc, The Icahn School of Medicine at Mount Sinai, and Aetna Inc for supporting me in my graduate studies, and for accommodating the sometimes conflicting demands of working full-time as well as being a PhD student. In particular I am grateful to Mr. Robert Shovlin, Dr. Michael Donovan, and Dr. Ricardo Mesa-Tejada for encouraging me to apply for a PhD program and facilitating the process while at Aureon. Additionally, I would like to thank my bosses Dr. Krish Ghosh at Covance, Dr. Gerardo Fernandez at Mount Sinai and Mr. Gaurav Sharma at Aetna. I am thankful to Aureon Biosciences for the data used in this research. I would also like to thank Dr. Qiuhua Liu, Dr. Valentina

Bayer-Zubek and Dr. Hrishikesh Karvir, colleagues at Aureon, for discussing my research ideas and providing feedback and advice.

Finally, I would like to acknowledge that the work presented in this dissertation has already been published and presented in the following papers:

1. Faisal M. Khan and Casimir A. Kulikowski. *Impact of Prostate Biopsy Tumor Amount on Imaging Based Prognostics Employing Transductive Semi-Supervised Regression*. 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC16. August, 2016.

2. Faisal M. Khan and Casimir A. Kulikowski. *Predicting Advanced Prostate Cancer Endpoints from Early Indications via Transductive Semi-Supervised Regression*. IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). June 2016.

3. Faisal M. Khan and Casimir A. Kulikowski. *Transductive Semi-Supervised Survival Analysis in Medical Prognostics*. Workshop on Computational Biology, International Conference on Machine Learning, ICML 2016. June 2016.

4. Faisal M. Khan and Casimir A. Kulikowski. *The Role of Imaging Based Prostate Biopsy Morphology in a Data Fusion Paradigm for Transducing Prognostic Predictions*. SPIE Proceedings of Medical Imaging 2016. March 2016.

5. Faisal M. Khan and Casimir A. Kulikowski. *Survival Analysis via Transduction for Semi-Supervised Neural Networks in Medical Prognosis*. IEEE International Conference on Bioinformatics & Biomedicine (BIBM). November 2015.

6. Faisal M. Khan and Qiuhua Liu. *Medical Survival Analysis Through Transduction of Semi-Supervised Regression Targets*. International Journal of Knowledge Discovery in Bioinformatics. 2011, 2:3, 52-65.

7. Faisal M. Khan and Qiuhua Liu. *Transduction of Semi-Supervised Regression Targets in Survival Analysis for Medical Prognosis*. Biological Data Mining Workshop (BioDM), IEEE 11th International Conference on Data Mining ICDM 2011. pp1018-1025, December 2011.

DEDICATION

This dissertation is dedicated to my family, past, present and future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

There are two broad categories of applications for predictive time-to-event modeling in medical survival analysis. The first is prognostic, developing models for how a certain disease will progress. The purpose of such models includes understanding disease progression and prediction of how new patients will behave in the context of existing data. Examples include predicting when prostate cancer will recur in patients so therapy can be initiated early [18] or identifying which group of patients will benefit more from a certain therapy. The second purpose is explanatory, through factor analysis; to analyze disease processes and explore interaction effects between disease factors. An example is determining whether a potentially significant gene will continue to be relevant when combined with other predictors in a multivariate model [18] in order to possibly prioritize and identify candidate genes for targeted therapeutic drug development.

While time-to-event prediction is inherently a regression problem, survival analysis challenges computational modeling approaches due to the fact that healthcare data in such settings is characterized by censored and non-censored (event) observations. The term “censoring” in biostatistics describes the fact that the target survival time is not known for all samples in the survival analysis setting. For instance, patients might not experience death or cancer relapse during the course of a study, or be lost to follow-up.

The only time known is their last record of being healthy; hence the target time for regression is incomplete and only known "up-to-a-point." This concept is distinctly different from the notion of missing data in machine learning [57].

Censored observations contribute incomplete information as the event of interest (cancer recurrence, death from disease) may occur after patients are lost to follow-up. Simply omitting the censored observations [7, 58] or treating them as non-recurring samples in a classifier [61] both bias the resulting model and should be avoided. Additionally, in the field of healthcare diagnostics, due to the costs involved in identifying acceptable patients who will provide consent for inclusion in research, and then actively tracking them over a significant period of time, the sample size is often small, in the tens or hundreds. Since most of the samples may be censored [e.g., 91% in prostate cancer [17], 76% in breast cancer [48]) dropping such patients is a very unattractive option and accounting for them is of crucial importance for a model. Survival analysis represents a special example of the typical complexity in modeling noisy high-dimensional biomedical data to predict complex medical phenomena.

The core contribution of this dissertation is that a possible way of handling censored samples so common in time-to-event problems would be to consider them as semi-supervised targets. While there has been significant work in semi-supervised classification approaches [3, 9, 10, 25, 26, 33, 56], there has been limited work in semi-supervised regression [2, 13, 51, 63, 72]. Work thus far treats samples as either fully labeled or unlabeled and does not take into account the partial nature of unsupervised information, as is the case in time-to-event medical prognosis problems.

This dissertation presents a novel approach which treats the survival analysis problem as one of semi-supervised regression and transduces or learns through trial and error the appropriate target times. This framework for transducing the appropriate times can be applied to any regression algorithm, whether originally developed for survival analysis or not. In experiments with multiple algorithms on datasets for prostate and breast cancer, the proposed framework consistently yields significant improvement in predictive accuracy. This dissertation is largely an empirical evaluation, supporting the proposed advancements with experimental findings.

This dissertation is organized as follows. Chapter 2 presents background material and related work. It provides an overview of survival analysis, and different machine learning methods which have been employed for survival analysis. Finally, it presents a description of semi-supervised approaches in both classification and regression contexts. Chapter 3 further presents background material, with a special focus on SVRc (not a contribution of this dissertation), a survival analysis approach based on support vector regression. SVRc has previously been developed by us and is employed in many of the experiments presented in this dissertation, thus it necessitating a more detailed introduction.

Chapter 4 presents the proposed semi-supervised transduction approach. It discusses the idea behind the algorithm, and reviews the approach's pseudocode. Various subtle but important implementation details are reviewed, including optimizing the complexity of the algorithm. Finally, experiments with the Cox proportional hazards model, support vector and neural network approaches are presented. The results in

Chapter 4 have already been published in two workshops, a conferences and a journal [35, 36, 38, 40].

Chapter 5 explores the application of the semi-supervised framework in a unique prognostic modeling situation where observatoins from earlier in a disease's history are employed to model subsequent disease endpoints. Prostate cancer is a complex disease which advances in stages. While clinical failure (including metastasis) is a significant endpoint following a radical prostatectomy, it can often take years to manifest, usually too late to be optimistically treated. Instead the earlier endpoint of PSA Recurrence is frequently used as a surrogate in prognostic modeling. Our proposed approach leads to a significant increase in performance for predicting advanced prostate cancer from earlier endpoints. These results were presented at the 2016 29th IEEE International Symposium on Computer-Based Medical Systems (CBMS) [41].

Chapter 6 presents the application of the proposed approach in the analysis of prostate biopsy imaging features. One of the major uses of survival analysis methods is to explore the predictive power of features and especially their interactions in a multivariate setting. In biomedical prognostics, there has recently been the development of a new "data fusion" paradigm where related features compete. For researchers in biomedical imaging, of particular interest is how quantitative imaging characteristics compare with existing clinical variables which may be measuring the same biomedical properties. We explore our novel approach for comparing clinical characteristics in prostate cancer, like the Gleason grade, with quantitative imaging algorithms. These results were presented at the 2016 SPIE Medical Imaging Conference [39].

Chapter 7 evaluates how the novel semi-supervised framework improves the performance of biopsy based prostate cancer assays as the available amount of tumor for analysis decreases. For newly diagnosed prostate cancer patients with a positive biopsy, there are a variety of treatment options to consider. To aid physicians and patients in their decision making, a variety of predictive assays have emerged within the last decade, many of them imaging based. These assays build predictive models for survival analysis to provide personalized risk assessments for the patients. However, there have rarely been any published studies on how the amount of tumor in the positive prostate biopsy affects the predictive power of these imaging based assays. We assess how different amounts of tumor in the prostate biopsy affect the accuracy of imaging based prognostic models employing our semi-supervised framework. We show that the framework improves accuracy even with diminishing amounts of tumor, thereby enabling more accurate treatment decisions. These results were presented at the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [42].

Finally, Chapter 8 summarizes the findings and contributions of this dissertation and discusses potential future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

This chapter is divided into three sections. First in Section 2.1 we present an overview of survival analysis. Second in Section 2.2 we review existing literature in methods for survival analysis. Third in Section 2.3 we present related work in semi-supervised analysis including regression and transduction in the classification setting.

2.1 Overview of Survival Analysis

Healthcare data for prognostic modeling is usually obtained by tracking patients over the course of time in a well-designed study, perhaps lasting years. Often a predefined event such as the relapse of a disease or death due to disease is the focus of the study. The major difference between survival analysis and other time-to-event regression problems is that the event of interest is frequently not observed in many of the subjects. Rather, the information for most subjects is incomplete and only their last healthy time is recorded. Patients that did not experience the endpoint during the study or were lost to follow-up for any cause (ie the patient moved during a multi-year study) are considered as *censored*. All that is known about them is that they were disease-free up to a certain point, but what occurred subsequently is unknown. They may have actually experienced the endpoint of interest at a later point in time, but that is unknown.

Conversely, patients who have experienced the endpoint of interest (cancer recurrence, death, etc.) are considered as *non-censored* samples or *events*. In many medical prognosis problems, the vast majority of instances (76% or even 91%) are censored, so they cannot be dropped. The incomplete nature of the outcome targets in survival analysis prediction thus challenges traditional regression techniques and usually precludes their use. Instead, methods which can correctly account for censored observations are essential [14, 29, 35, 57, 67].

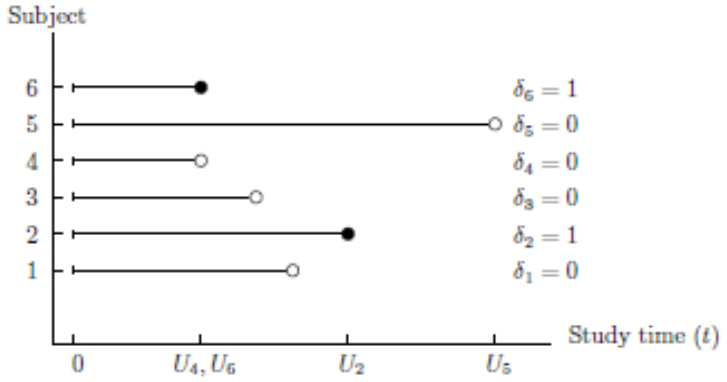


Figure 1: Illustration of survival time during a study. Event observations are indicated by solid dots, and censored observations by hollow dots. Reproduced from [57] by permission, ©IEEE 2014.

Figure 1 graphically represents a simplified example illustrating survival data for a study with six patients. Patients 2 and 6 show observed events, having experienced the endpoint of interest during the course of the study as indicated by solid dots. Patients 1, 3 and 4 were lost to follow-up during the study, and patient 5 reached the end of the study and was still healthy. Patients 1, 3, 4 and 5 are all considered as censored observations, if

they experienced the endpoint of interest, it was at some unobserved time following their last recorded observation.

If we let T_i denote the actual target time, C_i the censored time for a censored observation and U_i the observed time for all patients, then for events $U_i = T_i$ and for censored cases $U_i = C_i < T_i$. The survival outcomes for n patients is then represented by pairs of the random variables (U_i , δ_i) for $i = 1, \ldots, n$. The variable δ_i indicates whether the observed survival time U_i corresponds to an event ($\delta_i = 1$) or is censored ($\delta_i = 0$). Given a d-dimensional vector $x_i \in \mathbf{R}^d$, the data D for a medical prognosis problem can be represented as:

$$D = \{ U_{i,}, x_i , \delta_i \}^{n}_{i=1} .$$

$$U_i = \min (T_i , C_i)$$

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 0, & \text{for censored observation,} \\ 1, & \text{for exact observation.} \end{cases}$$

An important assumption is that T_i and C_i are independent conditional on x_i, meaning that the cause for censoring is independent of the survival time. In Figure 1, patients 4 and 6 have the same observed survival time ($U_4 = U_6$) however their censoring indicator variables are different ($\delta_4 = 0$, $\delta_6 = 1$).

Traditional statistical approaches to survival analysis attempt to estimate a survival function $S(t)$, the probability that the time-to-event is greater than a given time t, or Pr ($T_i > t$). The general problem is to learn S ($t \mid x$), the survival function conditional on the features of a patient in the data set. Once learned, this model is employed for prediction or explanatory factor analysis as described in the Introduction [35, 57].

The type of censoring described thus far (event-free and lost to follow-up) is known is "right-censoring," since information on the right-hand side of a timeline is unknown, as illustrated in Figure 1. The survival analysis problem is further confounded by the fact that non-censored patients actually experience the event-of-interest prior to their recorded time, that actually $U_i > T_i$. For instance, a cancer patient may visit a doctor every six months; so if recurrence is observed, it happened somewhere in the six months between the last "healthy" visit and the visit where the disease was detected. The term "left-censoring" describes this phenomenon where even the status of event patients is not completely known. Given the timelines involved, right-censoring is considered a significantly more important challenge and most survival analysis algorithms tend to ignore the left-censored nature of events. This dissertation as well concentrates on addressing the right-censoring problem, but it is important to be aware of left-censoring when working in survival analysis.

2.2 Related Work in Survival Analysis

2.2.1 Cox Proportional Hazards Model

The field of prognostic survival analysis has primarily been the focus of biostatisticians. The vast majority of practical research for new clinical trials, drug therapies, cancer prognosis, etc. in biological literature is performed with the Cox Proportional Hazards Model [14, 29, 36, 43, 67]. The Cox Model estimates the log hazard for a patient as a linear combination of the patient's features, plus a baseline hazard. The Cox Model makes the crucial but generally accurate assumption that the hazard function (the instantaneous rate of decline in survival at a point in time) is

proportional for all individuals at each time point; it is a constant ratio. This proportionality assumption is reflected in the general equation for the approach

$$h_i(t) = \exp\left(\sum_{j=1}^{p} b_j X_{ij}\right) h_0(t).$$

Where $h_i(t)$ is the hazard function for the i^{th} individual, b_j is the slope term for the j^{th} feature (which can be either categorical or continuous), X_{ij} is the value of feature j for individual i, $exp()$ refers to the exponential function, i.e., $exp(u) = e^u$ and $h_0(t)$ refers to the "baseline hazard function", the hazard function for an individual with simultaneous zero values for all features. Thus all hazard functions are assumed to be parallel to the baseline hazard function. Estimates of regression parameters (the b terms) are obtained via partial maximum likelihood estimation. The predicted hazard function for an individual allows predictions of an individual's survival. The Cox Model only employs censored patients' data in calculating the hazard function up to the time of censoring; afterwards they are excluded [22, 29, 30].

The Cox Model falls under the category of statistical semi-parametric approaches since the baseline hazard function is treated non-parametrically. To be more specific, the weights in the model are derived, however the baseline hazard function remains unspecified. It is a part of the Generalized Linear Model (GLM) family; however it can be observed that the parameters have a multiplicative effect on the hazard value which makes it different from other linear regression models [43, 49].

In general, the reliability of the Cox Model deteriorates if the number of features is greater than the number of events divided by ten [29]. Consequently, the Cox Model is

challenged by emerging trends in biology where large numbers of predictive factors such as genes are being analyzed in relatively small samples [17, 18, 52].

2.2.2 Machine Learning Approaches for Survival Analysis

While the field of survival analysis in medical applications has traditionally been the focus of statisticians, particularly biostatisticians, various machine learning approaches have also been explored. The use of decision trees adapted for censored data represent some of the earliest work in the field [28, 46, 54, 73]. Other techniques such as linear programming [48] have also been investigated.

An artificial neural network (ANN) is a complex modeling algorithm inspired by the biological neurons in a human brain. It consists of a series of network nodes at multiple layers which are "activated" through a mathematical function, often a sigmoid of the form:

$$\frac{1}{1+e^{-\theta^T x}}$$

Overall, given a training set of input vectors x_i, with a corresponding set of target vectors t_i, the algorithm minimizes the error function:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} ||y(x_n, w) - t_n||^2$$

Various forms of artificial neural networks have also been applied to survival analysis [5, 7, 61, 71, 74], with varying results. Some advantages arise from a neural network's ability to model nonlinearities. However, many have incorrectly treated the time-to-event problem as a classification problem rather than as regression, and often also struggled with the high dimensional data commonly found in biomedicine. NNci [71] is an implementation which treats the problem as one of regression, but modifies the use of the ANN's objective function to instead optimize the Concordance Index (CI). The CI is a performance measure of accuracy unique to survival analysis and is described further in Chapter 4. This implementation adapts NNs in a way that makes them directly applicable to the survival analysis problem in medical prognostics.

Widespread adoption of SVMs in various machine learning domains has also led to recent applications for survival analysis [22, 34, 35, 57, 58, 69]. However, approaches such as [22] treat the problem in a classification context rather than a regression problem. An adaptation of SVR has been proposed [69], however it only accounts for right-censored data, and while matching the performance of the Cox model, it yields no improvements over that standard. Another approach [58] only modified the error margin of the penalty function and not the penalty weight. In addition, left and right-censored cases were treated equivalently.

Another interesting avenue of research has been to combine the kernel concepts of methods such as SVMs with the Cox model to develop kernel Cox regression approaches [47].

2.2.3 SVRc

The Support Vector Regression (SVR) [59, 60] algorithm has proven to be a robust and useful tool in a variety of domains with an extensive body of literature describing its applications. However, since conventional SVR is unable to handle the censored data prevalent in survival analysis, it has not been more widely employed in the medical prognostics domain. Support Vector Regression for Censored Data (SVRc) [34, 45] is an approach for addressing this issue. This dissertation employs SVRc in many experiments, and thus the SVRc algorithm is further described in greater detail in Chapter 3.

2.3 Related Semi-Supervised Work

There has been a significant body of work in semi-supervised approaches for classification problems [3, 9, 10, 25, 26, 33, 56]. In many of these methods, the target class/label is learned or "transduced" by assigning different class labels to the unknown/unlabeled instances and selecting the one which has the best performance criteria in some optimization problem. Similar ideas are explored in semi-supervised regression.

2.3.1 Semi-Supervised Regression

The basic idea of transductive regression [13] is that given *m* labeled data and labels (x_1, y_1), …, (x_m, y_m) as well as *u* unlabeled data points x_{m+1}, …, x_{m+u}, transductive regression learning algorithms must accurately predict the labels y_{m+1}, …, y_{m+u}. To date, there have been various approaches developed for semi-supervised regression. In [13] there are two basic steps described for such algorithms. The first is local estimation where initial labels

of unlabeled datapoints are assigned based on their neighbors, through a weighted averaging scheme. In the second step, through global optimization, a hypothesis is selected that best fits the supervised labels and the estimated labels from the first step of the unlabeled samples.

The local linear semi-supervised regression algorithm [51] has two properties: 1) It fits a linear function at each point like in local linear regression; and 2) The estimation of the labels of any particular data point depends on the estimates for all the other samples in its neighborhood as in Gaussian Fields. Reference [2] proposed a family of learning algorithms that exploit the geometric distribution of features as a manifold regularization term. There are two proposed algorithms, a Laplacian-regularized SVM for classification and the Laplacian-regularized least-squares approach for regression. Reference [72] proposed a generalization of a well-known co-training algorithm for classification. The original algorithm trains two classifiers separately on two sufficient yet redundant attribute sets, each of which is sufficient for learning and is conditionally independent of the other given the class label. The approach employs the predictions of each classifier on labeled samples to augment the training set of the other. Instead of two attributes, [72] adopts two kNN regressors, each of which is refined iteratively with the help of unlabeled samples that are labeled by the latest updated version of the other regressor. On convergence, the final output is the average of the two models.

In all these approaches, a major challenge is the choice of the initial sample label transduced for the unsupervised instances. There are multiple choices for computing similarity measures between feature vectors to compute the initial labels of the unlabeled

data, such as the Euclidean distance, kNN, Markov random walks [63], or normalized Laplacian [2].

While many of these approaches work well for classical semi-supervised regression problems, where instances are fully labeled and unlabeled, their direct adoption for survival analysis is not ideal since they do not leverage the partial information of true outcome present in the partial labels (the censored times) of a majority of the instances. Additionally, classical semi-supervised regression problems do not reflect the typical circumstances of survival analysis where up to 91% of the instances may be unsupervised, but contain partial information. The relative scarcity of neighboring events with known target labels for censored instances may challenge them. For instance, [13] drops samples from analysis if there aren't enough neighbors to transduce a label for them.

Reference [1] applies semi-supervised methods for survival analysis. However, [1] does not leverage the concept of partial information in censored target times. Instead, unsupervised clustering is performed to recognize related genes, followed by supervised modeling. The method is semi-supervised in the sense that unsupervised gene discovery is paired with supervised prediction modeling. To the best of our knowledge, leveraging the partial knowledge of true outcome in the censored times for survival analysis is a largely neglected and potentially rich area of research. This dissertation proposes an overall framework for transducing times for censored instances in survival by leveraging the partial semi-supervised nature of the censored times.

CHAPTER 3

BACKGROUND AND RELATED WORK: FOCUS ON SVRc

Developed at AT&T Bell Labs by Vladimir Vapnik, Support Vector Machines (SVMs) [6, 59, 60, 70] have emerged as a powerful and compelling tool in the field of machine learning. Advantages that have driven widespread adoption include a grounding in statistical learning theory, extension of linear models to nonlinear problems [31], and applicability to high-dimensional data while overcoming the curse of dimensionality; numerous case studies have been published documenting the excellent performance of SVMs in various problem domains. Although the algorithm was initially developed in a classification setting, it was quickly adapted for time series prediction and regression problems.

The SVM algorithm is well grounded in statistical learning theory, but is abstractly a simple and intuitive linear algorithm; SVMs are linear models capable of linear and nonlinear modeling. Usually, linear models are incapable of representing a model with nonlinear relationships. SVMs employ linear models to represent both linear and nonlinear relationships by transforming the input feature space, into a new higher-dimensional feature space using a mapping. This transformation is facilitated through the use of mathematical functions called kernels. The SVM algorithm abstractly maintains a

linear relationship between outcomes and features; potential nonlinearities are encapsulated within the feature space via the kernel mapping. Consequently, complex pattern recognition, classification and regression approaches can abstractly be represented linearly.

The choice of the kernel function and the resultant feature space is important in theoretical and practical terms. It determines the functional form of the model; thus, different kernels may behave differently. For a mathematical function to be a valid kernel it must meet a set of conditions as outlined in [31]. Some of the most basic and common kernels are:

Linear: $\Phi(\vec{X},\vec{Y}) = (\vec{X} \bullet \vec{Y})$

Polynomial: $\Phi(\vec{X},\vec{Y}) = (\vec{X} \bullet \vec{Y})^d$

Radial Basis Function (RBF): $\Phi(\vec{X},\vec{Y}) = \exp(-\| \vec{X} - \vec{Y} \|^2 / (2\sigma^2))$

Sigmoid: $\Phi(\vec{X},\vec{Y}) = \tanh((\vec{X} \bullet \vec{Y}) + \Theta)$

This chapter is organized as follows. Section 3.1 presents the formulation of traditional Support Vector Regression. Section 3.2 then presents the adaptation of traditional SVR into SVRc, a modified version for survival analysis.

3.1 Support Vector Regression (SVR)

The Support Vector Regression (SVR) [59, 60] algorithm is an extension of SVMs to the regression setting. Once a SVR model has been learned, it can be applied to a new instance x through the following equation:

$$f(x) = W \bullet \Phi(x) + b$$

Thus, the SVR algorithm can abstractly be considered a linear algorithm similar to basic linear regression where the variables *m* and *b* are learned for the equation *y=mx+b*. The potential nonlinearity of a problem is encapsulated within the kernel function $\Phi(x)$, and the complexity of the problem is resolved within the higher dimensional feature space.

During SVR training, following the transformation of the data into the feature space, the algorithm learns the regression function *f(x)* that best fits the data in the feature space. SVR training involves minimizing the training error (empirical risk) controlled by a single regularization parameter *C* and a margin of error ε. This translates to obtaining the coefficients *W* and *b* through an optimization problem. Given a set of *n* input instance vectors $\vec{x}$ $(x_1, x_2, \ldots x_n)$ with corresponding target values $\vec{y}$ $(y_1, y_2, \ldots y_n)$, the algorithm minimizes the following objective function:

$$\min_{W,b} \quad \frac{1}{2} \| W \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

given the constraints:

$$y_i - (W \bullet \Phi(x_i) + b) \le \varepsilon + \xi_i$$
$$(W \bullet \Phi(x_i) + b) - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0, \qquad i = 1..n$$

where W is a vector with the weights of all the features in the higher dimensional kernel space, and $\Phi(x_i)$ represents the transformation of the instance x_i in the higher dimensional kernel space. The slack variables ξ_i, ξ_i^* make the constraints of the optimization problem feasible. The slack variables are characterized by the epsilon-insensitive loss function $|\xi|_\varepsilon$ where:

$$|\xi|_\varepsilon = \begin{cases} 0 & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases}$$

The variables C and ε shall now be explained in more detail in section 3.1.1.

3.1.1 Further Details of SVR Training

During SVR training, a common optimization function used is an epsilon insensitive loss function. In each iteration of the optimization, the algorithm attempts to find the best fit line for the data in the kernel space. However, since it is not possible to build a model that will perfectly fit all the training instances; an acceptable margin of error is set with the parameter ε, as illustrated in Figure 2. Instances for which the error *(f(x)-y)* of the model's prediction *f(x)* and actual target value *y* is within ε are considered

to be fitted by the model; instances outside the so-called "ε-insensitive tube" are poorly fit by the model.

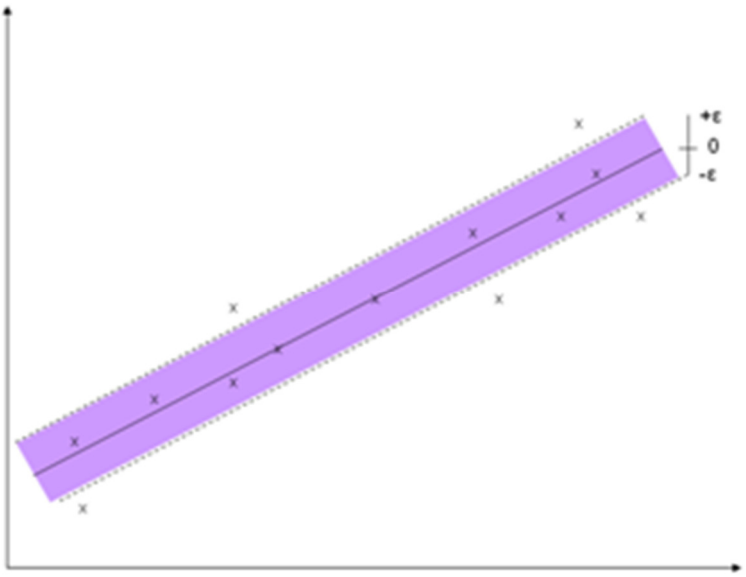


Figure 2: An illustration of the ε-insensitive tube, inspired by [59].

There is a model penalty associated with the instances that the line doesn't fit; this is controlled by the structural risk regularization parameter, *C*. During the training optimization, instances within the "ε - insensitive tube" have a penalty of zero, the model fits them correctly. The model receives a penalty for training records with errors *(f(x)-y)* greater than ε. The penalty is determined relative to the size of the error by a line with a slope of *C*. The larger the error, the larger the penalty as determined by *C*. Figure 3 illustrates this relationship [34].

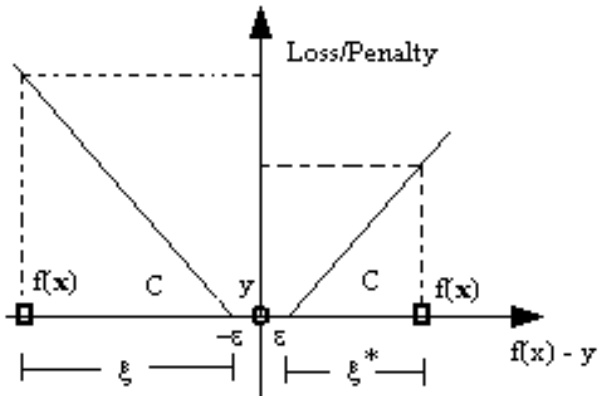


Figure 3: A graphical representation of the parameters C and ε in SVR. The x-axis represents the model error f(x) – y for an instance x [34].

The parameters ε and C achieve a balance between a good fit of the training data and the simplicity and generalization ability of the solution. The parameter ε sets a threshold of insignificant error in the function approximation. Simultaneously, it defines the complexity of the approximation. In SVR, the support vectors are the instances which have a difference between predicted and target values greater than ε. A smaller ε leads to more support vectors and an increased complexity of the approximation. If the approximation is too complex, it may lead to overfitting. The value of the parameter ε is closely related with the precision of the training data. If it is known that errors in measuring the target y are on the order of γ, then it does not make sense to have the value of ε less than γ. The parameter C controls relative importance of the two components of the functional: the relative risk, characterizing the quality of fit, and the complexity of the approximations.

Further details of the algorithm and its underlying mathematical theory are available in an excellent tutorial by Smola and Schölkopf [59].

3.2 Support Vector Regression for Censored Data: SVRc

The Support Vector Regression (SVR) [59, 60] algorithm has proven to be a robust and useful tool in a variety of domains with an extensive body of literature describing its applications. However, since conventional SVR is unable to handle the censored data prevalent in survival analysis, it has not been more widely employed in the medical prognostics domain. Support Vector Regression for Censored Data (SVRc) [34, 45] is an approach for addressing this issue. The key issue in applying conventional SVR to survival analysis is the inability to handle the differences between censored and event instances. The (left-censored) target regression values for events are fairly certain; the actual time may have occurred a short time prior to the recorded observation. The censored target values are extremely uncertain. The core SVRc concept is to account for the differences between these instances by asymmetrically modifying the ε - insensitive loss function optimized during training. The update introduces four new versions of both C (penalty slope) and ε (insensitive penalty threshold) parameters that account for censored and non-censored instances differently.

3.2.1 Events in SVRc

For events in the training cohort, the SVRc algorithm introduces four new parameters C_n^* and C_n which replace C, and ε_n^* and ε_n which replace ε. The approach

takes into account the left-censored nature of events; i.e. patients experiencing a disease event before it's detected during a visit to the doctor.

The parameter ε_n^* defines the acceptable margin of error if the model's predicted value is greater than the actual target *(f(x)>y)*; if so, the penalty function is controlled by C_n^*. The parameter ε_n defines the acceptable margin of error if the model's predicted value is less than the actual target *(f(x)<y)*; if so the penalty function is controlled by C_n. The suggested relationships between these parameters are $\varepsilon_n > \varepsilon_n^*$ and $C_n < C_n^*$ to account for the left-censored nature of events. Consequently, if the model predicts an event as occurring before the actual target value, there is a relatively larger error margin and smaller penalty. Figure 4 illustrates this relationship. However, if one does not wish to account for the left censored nature of events, the parameters can simply be chosen to follow the relationship $\varepsilon_n = \varepsilon_n^*$ and $C_n = C_n^*$.

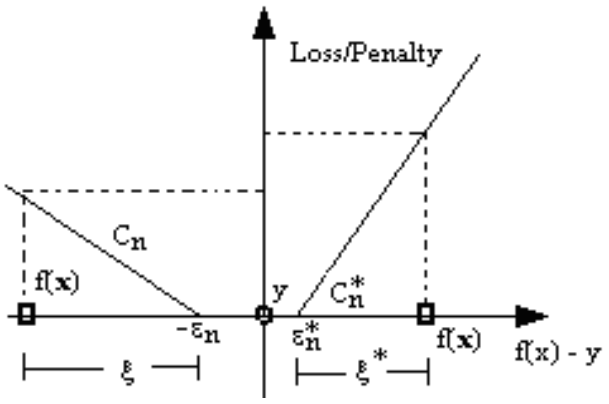


Figure 4: A graphical representation of the SVRc parameters for events. The x-axis represents the model error f(x) – y for an instance x [34].

3.2.2 Censored Instances in SVRc

The censored instances in the cohort are treated similarly as the events. The SVRc algorithm introduces four new parameters C_c^* and C_c which replace C, and ε_c^* and ε_c which replace ε. The algorithm accounts for the right-censored nature of the samples; patients experiencing an event of interest after their last recorded disease free time.

The parameter ε_c^* defines the acceptable margin of error if the model's predicted value is greater than the actual target *(f(x)>y)*; if so, the penalty function is controlled by C_c^*. The parameter ε_c defines the acceptable margin of error if the model's predicted value is less than the actual target *(f(x)<y)*; if so the penalty function is controlled by C_c. The suggested relationships between these parameters are $\varepsilon_c < \varepsilon_c^*$ and $C_c > C_c^*$ to account for the right-censored instances. If the model predicts a censored instance as occurring after the actual target value, there is a relatively larger error margin and smaller penalty. Figure 5 illustrates this relationship

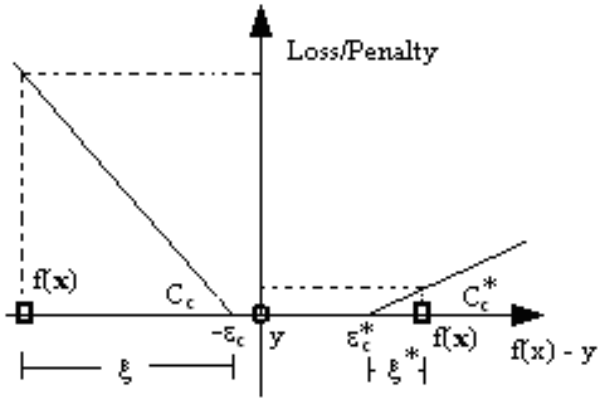


Figure 5: A graphical representation of the SVRc parameters for censored instances. The x-axis represents the model error f(x) - y for an instance x [34].

3.2.3 Overall SVRc Algorithm

For events in the training cohort, the SVRc algorithm introduces four new parameters to model the left-censored nature of events. For censored cases, there are an additional four new parameters to asymmetrically give censored predictions larger than the target time a wider margin of error with less penalty. In the context of the overall approach that encompasses both censored and event instances, the algorithm minimizes the following objective function:

$$\min_{W,b} \quad \frac{1}{2} \| W \|^2 + \sum_{i=1}^{n} (C_i \xi_i + C_i^* \xi_i^*)$$

given the constraints:

$$\begin{aligned} & y_i - (W \bullet \Phi(x_i) + b) \le \varepsilon_i + \xi_i \\ & (W \bullet \Phi(x_i) + b) - y_i \le \varepsilon_i^* + \xi_i^* \\ & \xi_i^{(*)} \ge 0, \qquad i = 1...n \end{aligned}$$

where:

$$\begin{aligned} & s = 1 \quad if \quad censored \\ & s = 0 \quad if \quad event \\ & C_i^{(*)} = s_i C_c^{(*)} + (1 - s_i) C_n^{(*)} \\ & \varepsilon_i^{(*)} = s_i \varepsilon_c^{(*)} + (1 - s_i) \varepsilon_n^{(*)} \end{aligned}$$

The suggested relationships between the eight parameters are $C_c^* < C_n < C_n^* = C_c$ and $\varepsilon_c^* > \varepsilon_n > \varepsilon_n^* = \varepsilon_c$. This is because the penalty for censored predictions less than the target time or event predictions greater than the target time should be equivalent and the largest, since these predictions are clearly incorrect. There should be a small adjustment for non-censored predictions before the target time due to the left-censored nature of events, and the greatest allowance should be made for censored predictions after the target time because they may in fact be correct.

This update of the SVR loss function for different error and structural risk parameters and the asymmetric relationships between those parameters is the core contribution of SVRc. The algorithm retains all the advantages of conventional SVR such as the mathematical mapping via kernels into higher dimensions, the concentration on the most important instances in a dataset, the decreased susceptibility to overfitting, and the ability to model with more features, plus it can now be applied to the field of survival analysis where censored data is prevalent. The SVRc algorithm is leveraged for multiple experiments in this dissertation and thus described in detail here; however the development of SVRc is not part of the research contributions of this dissertation.

CHAPTER 4

A FRAMEWORK FOR SEMI-SUPERVISED SURVIVAL ANALYSIS BY TRANSUDCING CENSORED REGRESSION TARGETS

This chapter is presented in three sections. Section 4.1 describes the proposed semi-supervised transduction framework for regression in survival analysis. Section 4.2 discusses accuracy metrics for evaluating the performance of survival models. Section 4.3 presents the results of applying the proposed framework to the Cox Model, neural network and support vector approaches for predicting medical prognoses in prostate and breast cancers.

4.1 Semi-Supervised Framework for Transducing Regression Targets in Survival Analysis

As discussed, the ability to leverage the incomplete information in the censored samples of time-to-event problems could provide significant advantages. If the "true" target as opposed to the censored target was known, the performance of predictive models would be increased.

We present an innovative approach that is, in essence, a wrapper around any regression function, whether developed for survival analysis or not. For each censored case (U_i , δ_i = 0), it iterates through possible target values between U_i and T_{max} (the maximum observed time U in the dataset). It then transduces or chooses a new target time $\hat{U}T_i$ which improves accuracy, maximizing some criterion for measuring predictive performance. The approach is extremely flexible, able to work with almost any regression function $F()$ and measure of accuracy $Criterion\ (y\ ,\ t)$. Given a dataset $D = \{\ U_{i,}\ , x_i\ , \delta_i\ \}\ ^{N}_{i=1}$, the algorithm can be described as:

$$\max_{Criterion\ (y\ ,U),\quad \hat{U}T} \qquad y = F(D = \{U_i\ , x_i\ , \delta_i\ \}_{i=1}^{n})$$

Given the constraints:

$$T_{max} = max\ (U_{i=1,\ldots,n})$$

$$U_i \leq \hat{U}T_i \leq T_{max} \qquad ; \qquad \text{if } \delta_i = 0$$

$$\hat{U}T_i = U_i \qquad ; \qquad \text{if } \delta_i = 1$$

A key issue is exploring the space of possible target values. Semi-supervised classification algorithms initially employed an exhaustive method assigning each class label to every unlabeled instance, in order to transduce the optimal label. Unfortunately, this led to a transduction complexity of C^n where C is the number of classes and n the number of unlabeled instances. Accordingly, researchers began to develop computationally more reasonable methods.

The proposed semi-supervised regression approach exploits a censored instance's own partial information of true outcome rather than its neighbor's labels to transduce optimal target times. The censored time represents the minimum possible value of the true target. The optimal target for each censored instance could thus be transduced by testing values in increments from the censored time to the maximum survival time in the training cohort. The initial idea was to replicate the exhaustive search of semi-supervised classification, but this is impractical. In one sample dataset, an average of 10 target values per each of the 341 censored cases would result in a transduction complexity of 10^{341}. To avoid this, the proposed technique is a singular transduction procedure which forgoes the exhaustive method. In this scenario, each instance is treated independently, and the best time for each censored case is found independent of the other censored cases. Consequently, a slight modification is required for the algorithm's optimization function described above, resulting in a singular rather than exhaustive transduction approach:

$$\left(\max_{Criterion\,(y,U)\,,\ \hat{U}T_i} y = F\ (D = \{U_i, x_i, \delta_i\}_{i=1}^{n}) \right)_{i=1}^{n}$$

Figure 6 below presents the proposed approach in pseudo code:

```
1  m=number censored
2  best_target=original_target
3  Model = Regression (original_target)
4  orig_criteria=Performance(original_target,Model)
5
6  for i=1:m
7      target=orig_target;
8      best_criteria=orig_criteria;
9      best_time=target(i)
10
11     for time=original_target(i):max_time
12         time=time+x
13         target(i)=time
14         Model=Regression (target)
15         critera=Performance(original_target,Model)
16         if(criteria>best_criteria)
17             best_criteria=criteria
18             best_time=time
19         end
20     end
21     best_target(i)=best_time;
22 end
23
24 Model=Regression(best_target)
25 Results=Performance(original_target,Model)
```

Figure 6: Pseudocode of proposed approach.

Initially, a traditional regression model is constructed (pseudocode line 3). Subsequently, each of the *m* censored instances in a training cohort is singularly transduced (pseudocode for-loop in lines 6-22). In each i^{th} singular transduction iteration (where i is between 1 and *m*), models are constructed as the target for the i^{th} censored instance increments from the i^{th} censored time to the maximum time in intervals of *x* units (pseudocode for-loop in lines 11-20). Meanwhile, the original censored targets are maintained for the other *m*-1 censored cases. The time that yields the best improvement in performance criteria is the transduced target label for that instance. In the subsequent iterations for other censored cases, the original censored time for this i^{th} instance is used

(not the new transduced target time). Hence, this procedure avoids favoring a particular censored instance by transducing its time first. On the other hand, the approach does preclude the discovery of transduction times that would result from an exhaustive search. Finally, once an optimal target time has been transduced for each censored instance, the final regression model is created using the new target times (pseudocode line 24).

One subtle but crucial point to note is that when evaluating the fit of the model on the training data (during the search to choose the best target time and at the end when the final model is built in pseudocode lines 14 and 24), the evaluation should be done with the original censored times rather than the new transduced times. Otherwise the resulting performance metrics may artificially be inflated as they will be calculated on the discovered targets that were derived precisely to improve performance. Subsequently when testing on an independent validation set where it is not possible to transduce the times but to simply apply the model, the resulting model would grossly overfit; as was observed. This is exactly why in the functions above we maximize *Criterion (y , U)* rather than *Criterion (y, $\hat{U}T$)*. Again this is important due to the unique nature of survival analysis in medical prognosis where most of the instances are censored.

4.2 Performance Metrics

In conventional regression, a useful metric of accuracy is the model's error in predicting the targets. However, in survival analysis this is not possible due to the prevalence of censored instances. For events, the prediction error can be easily assessed. For censored records, predictions are wrong only if they are less than the targets, otherwise the error, if any, is unknown. This requires alternate performance criteria.

The concordance index (CI) is the standard metric for assessing the predictive ability of a survival model [22, 29, 34, 49, 69, 71]. The CI measures the concordance between model results and the observed survival times. Survival analysis is inherently a ranking problem and the CI measures the accuracy of ranking a model's results against the patients' survival times. It is calculated in pair wise comparisons of all comparable patients in the cohort, and is the probability that a patient with a shorter survival time will have a smaller predicted result. The CI ranges from 0 to 1, with 0.5 indicating an absence of correlation, a random result. A value of 0 indicates perfect negative correlation, and 1 indicates perfect positive correlation. It is a linear transform of the Somers' d statistic, and is similar in interpretation to the area under the ROC curve (AUC) and the Mann-Whitney statistics [49, 69, 71].

This can be observed since for a two-class classification problem, the AUC has the form:

$$AUC = \frac{\sum_{(i,j)\in\theta} I(s_i, s_j)}{|\theta|}$$

Where s_i and s_j are classifier outputs for a positive sample *i* and a negative sample *j*. θ consists only of the pairs of positive samples *i* and negative samples *j*. Hence the AUC doesn't compare within the same class, but only considers pairwise comparisons between a pair of positive and negative samples. While acceptable for a standard classification task, the AUC cannot account for the critical survival time in a survival analysis problem. The CI makes pairwise comparisons of all patients within a dataset under two conditions:

1) Patients *i* and *j* who both experiences the event of interest and the event time t_i of patient *i* is less than patient *j*'s event time t_j
2) Only patient *i* is non-censored and $t_i < t_j$ (patient *j*'s follow-up censored time).

$$CI = \frac{\sum_{(i,j)\in\Omega} R(t_i, t_j)}{|\Omega|}$$

Here Ω represents the set of all possible comparisons between either two events, or all censored times after an event time, but not comparisons between two censored times [71].

One of the main uses of a survival model in medical applications is to stratify a patient population into high and low risk groups. Diverse risk profiles can lead to different and better targeted therapies and disease management for improved treatment. For a specific time point, patients can be stratified into high and low risk groups based on a model's predictions. The positive class identifies patients who are events prior to this time point, and the negative class identifies patients (censored or events) with targets after the time point. Censored patients with targets prior to the time point are excluded. Hence, in addition to evaluating a model's overall accuracy via the CI, the ability to correctly identify high and low risk groups is measured via the sensitivity and specificity of the low/high risk group classification. Since censored patients with targets earlier than the time point are excluded, it is often a good idea to evaluate the CI and the classification metrics at the same time.

Both the CI and the sensitivity-specificity pairing are metrics independently used in the medical literature [17, 18]. However, in order to assess both measures at the same time, we employed a performance criterion which combined both metrics, emphasizing

equally the CI and the product of the sensitivity and specificity. The product of the sensitivity and specificity is a good measure of both that has the same scale of accuracy as the CI. While in absolute theory, the CI may not have the same range as the product of sensitivity and specificity because CI values less than 0.5 imply negative correlation (similar to the AUC), this is not problematic from a practical perspective as all useful models must have Cis greater than 0.5. Consequently, in all the presented experiments the performance criterion for evaluation was:

$$Criterion = CI + (Sensitivit\,y * Specificit\,y)$$

4.3 Experimental Results

The proposed framework can be employed for any regression algorithm, whether or not it was originally developed for survival analysis. We focused our attention on evaluating the approach and combining it with the Cox Model, NNci and SVRc. Additionally, since NNci is an evolution of ANN and SVRc is a modification of traditional SVR, we also compared the improvement of the semi-supervised transduction framework for these core regression algorithms as well. The aim was to compare the performance of both the core and survival analysis versions of theses algorithms when the framework was layered on top.

We conducted experiments with four survival analysis datasets representing prognostic problems in prostate and breast cancer. In all experiments, the interval of time used to explore the target space (variable x in the pseudocode) was 10 months. All feature values were scaled between -1 and 1 based on the minimum and maximum values

in training. As noted, in all experiments the performance metrics were assessed according to the original times; no transduced targets were used in the accuracy assessments.

In developing medical prognostics, it is necessary to maintain separate training and validation sets (rather than combined cross-validation type approaches) due to FDA regulatory requirements for independent testing and validation. Similar requirements exist for health insurance companies as well when evaluating whether to cover and reimburse costs for a potentially expensive prognostic model and assay.

Study 1 [11] analyzed the endpoint of PSA Recurrence post-radical prostatectomy (RP) in patients treated for prostate cancer. High and low risk groups were assessed for PSA Recurrence at 5 years post-RP. The study consisted of 682 patients from the Memorial Sloan Kettering Cancer Center (MSKCC) split into 342 training and 340 validation patients. 40 features representing clinical, biomolecular and image morphometric domains were analyzed. Eighty-three percent of the training and eighty-seven percent of the validation patients were censored.

Study 2 [17] analyzed the endpoint of clinical failure (including metastatic disease progression) post-RP in 758 MSKCC prostate cancer patients. The patients were split into 373 training and 385 validation records; the same 40 features representing clinical, genetic and imaging information as study 1 were analyzed. High and low risk groups at 5 years were studied. Ninety-one percent of the training and ninety-two percent of the validation patients were censored.

Study 3 [18] also analyzed the endpoint of clinical failure post-RP in a multi-institutional cohort of 1027 patients from the Mayo Clinic, Duke-Durham Veterans

Affairs Medical Center, University of Connecticut Health Science Center, and the University Hospital at Uppsala. It was split into 686 training and 341 validation records. A different set of 40 features representing clinical, genetic and imaging information was analyzed. High and low risk groups were at 8 years. The censoring rate in both training and validation datasets was eighty-seven percent.

Study 4 [48] was conducted on a publicly available cohort of 194 breast cancer patients. The patient data was split into 129 training and 65 validation records. The 32 features representing clinical and imaging characteristics were modeled and high and low risk groups at 3 years were calculated. Seventy-six percent of the training and seventy-seven percent of the validation patients were censored.

4.3.1 Results with the Cox Proportional Hazards Model

Table 1 below presents the results in both training and validation sets for all four studies of just the Cox Model by itself, and the proposed framework layered on top of the Cox Model. With the exception of Study 3, the proposed framework is improving the accuracy of the Cox Model in both training and validation results.

Table 1: Experimental Results of the Cox Model and the Cox Model with Transduction

Cox Model		**Cox Model with Transduction**	
Study 1			
Train CI:	0.87	Train CI:	0.87
Train Sensitivity:	0.75	Train Sensitivity:	0.76
Train Specificity:	0.90	Train Specificity:	0.90

Train Criteria:	**1.54**	**Train Criteria:**	**1.55**
Test CI:	0.73	Test CI:	0.73
Test Sensitivity:	0.44	Test Sensitivity:	0.45
Test Specificity:	0.84	Test Specificity:	0.84
Test Criteria:	**1.09**	**Test Criteria:**	**1.11**
Study 2			
Train CI:	0.94	Train CI:	0.93
Train Sensitivity:	0.86	Train Sensitivity:	0.90
Train Specificity:	0.91	Train Specificity:	0.87
Train Criteria:	**1.71**	**Train Criteria:**	**1.72**
Test CI:	0.80	Test CI:	0.81
Test Sensitivity:	0.47	Test Sensitivity:	0.63
Test Specificity:	0.87	Test Specificity:	0.85
Test Criteria:	**1.22**	**Test Criteria:**	**1.35**
Study 3			
Train CI:	0.78	Train CI:	0.78
Train Sensitivity:	0.77	Train Sensitivity:	0.70
Train Specificity:	0.81	Train Specificity:	0.84
Train Criteria:	**1.40**	**Train Criteria:**	**1.42**
Test CI:	0.67	Test CI:	0.67
Test Sensitivity:	0.52	Test Sensitivity:	0.45
Test Specificity:	0.79	Test Specificity:	0.79

Test Criteria:	**1.08**	**Test Criteria:**	**1.03**
Study 4			
Train CI:	0.81	Train CI:	0.82
Train Sensitivity:	0.81	Train Sensitivity:	0.81
Train Specificity:	0.78	Train Specificity:	0.80
Train Criteria:	**1.44**	**Train Criteria:**	**1.47**
Test CI:	0.60	Test CI:	0.61
Test Sensitivity:	0.55	Test Sensitivity:	0.55
Test Specificity:	0.65	Test Specificity:	0.70
Test Criteria:	**0.96**	**Test Criteria:**	**0.99**

4.3.2 Results with ANNs and NNci

We compared two adaptations of neural networks. The first was a basic ANN regression approach and the second was NNci, a specialized ANN developed specifically for Survival Analysis. We employed feed-forward ANNs in Matlab (with the Levenberg-Marquardt back-propagation method) with 3 hidden layers running for a maximum of 100 iterations per ANN optimization. Results of a basic ANN compared with our proposed approach layered on top of the basic ANN are presented in Table 3. In Table 3 we present the results of NNci and the proposed semi-supervised framework combined with NNci.

Table 2: Experimental Results of the Basic ANN and Basic ANN with Transduction

Basic ANN		**Basic ANN with Transduction**	
Study 1			
Train CI:	0.76	Train CI:	0.84
Train Sensitivity:	0.82	Train Sensitivity:	0.86
Train Specificity:	0.80	Train Specificity:	0.88
Train Criterion:	**1.42**	**Train Criterion:**	**1.60**
Test CI:	0.61	Test CI:	0.68
Test Sensitivity:	0.67	Test Sensitivity:	0.64
Test Specificity:	0.71	Test Specificity:	0.71
Test Criterion:	**1.09**	**Test Criterion:**	**1.13**
Study 2			
Train CI:	0.80	Train CI:	0.93
Train Sensitivity:	0.76	Train Sensitivity:	0.95
Train Specificity:	0.91	Train Specificity:	0.97
Train Criterion:	**1.49**	**Train Criterion:**	**1.86**
Test CI:	0.67	Test CI:	0.75
Test Sensitivity:	0.63	Test Sensitivity:	0.84
Test Specificity:	0.78	Test Specificity:	0.71
Test Criterion:	**1.16**	**Test Criterion:**	**1.35**
Study 3			
Train CI:	0.67	Train CI:	0.79

Train Sensitivity:	0.66	Train Sensitivity:	0.86
Train Specificity:	0.73	Train Specificity:	0.83
Train Criterion:	**1.15**	**Train Criterion:**	**1.51**
Test CI:	0.59	Test CI:	0.62
Test Sensitivity:	0.64	Test Sensitivity:	0.67
Test Specificity:	0.63	Test Specificity:	0.76
Test Criterion:	**0.99**	**Test Criterion:**	**1.13**
Study 4			
Train CI:	0.62	Train CI:	0.70
Train Sensitivity:	0.68	Train Sensitivity:	0.82
Train Specificity:	0.55	Train Specificity:	0.72
Train Criterion:	**1.00**	**Train Criterion:**	**1.30**
Test CI:	0.61	Test CI:	0.67
Test Sensitivity:	0.64	Test Sensitivity:	0.73
Test Specificity:	0.59	Test Specificity:	0.59
Test Criterion:	**0.99**	**Test Criterion:**	**1.10**

Table 3: Experimental Results of NNci and NNci with Transduction

NNci		**NNci with Transduction**	
Study 1			
Train CI:	0.61	Train CI:	0.76
Train Sensitivity:	0.43	Train Sensitivity:	0.80

Train Specificity:	0.81	Train Specificity:	0.82
Train Criterion:	**0.96**	**Train Criterion:**	**1.41**
Test CI:	0.62	Test CI:	0.73
Test Sensitivity:	0.44	Test Sensitivity:	0.74
Test Specificity:	0.83	Test Specificity:	0.71
Test Criterion:	**0.98**	**Test Criterion:**	**1.26**
Study 2			
Train CI:	0.81	Train CI:	0.88
Train Sensitivity:	0.86	Train Sensitivity:	0.95
Train Specificity:	0.80	Train Specificity:	0.92
Train Criterion:	**1.49**	**Train Criterion:**	**1.75**
Test CI:	0.74	Test CI:	0.80
Test Sensitivity:	0.79	Test Sensitivity:	0.84
Test Specificity:	0.73	Test Specificity:	0.75
Test Criterion:	**1.32**	**Test Criterion:**	**1.43**
Study 3			
Train CI:	0.73	Train CI:	0.74
Train Sensitivity:	0.78	Train Sensitivity:	0.69
Train Specificity:	0.65	Train Specificity:	0.78
Train Criterion:	**1.24**	**Train Criterion:**	**1.27**
Test CI:	0.66	Test CI:	0.65
Test Sensitivity:	0.70	Test Sensitivity:	0.76

Test Specificity:	0.65	Test Specificity:	0.67
Test Criterion:	**1.12**	**Test Criterion:**	**1.15**
Study 4			
Train CI:	0.60	Train CI:	0.71
Train Sensitivity:	0.95	Train Sensitivity:	0.77
Train Specificity:	0.04	Train Specificity:	0.68
Train Criterion:	**0.64**	**Train Criterion:**	**1.24**
Test CI:	0.58	Test CI:	0.66
Test Sensitivity:	0.45	Test Sensitivity:	0.64
Test Specificity:	0.51	Test Specificity:	0.65
Test Criterion:	**0.81**	**Test Criterion:**	**1.07**

These experimental results in four real world datasets for prostate and breast cancer from different institutions appear to confirm the validity of the proposed approach for ANNs. In all the experiments, whether we consider the basic ANN or NNci, both in training and validation, the transduction framework improves performance as measured by the defined Criterion. While independent components of the criterion do vary, the algorithm was designed to optimize the overall criterion, and it has performed well. Researchers can emphasize whichever measure of accuracy is more appropriate for their specific task, and the results seem to indicate that the proposed approach could improve results in not only training, but the all-important separate validation set. In the current experiments, the neural network architecture was fixed; simply the bias and weight terms

were optimized during transduction. Given the complexity of neural networks, future work would be to allow evolution of the architecture as well during transduction. This would significantly increase the complexity and execution time of the approach, but with research into optimization methodologies, could yield improved results.

4.3.3 Results with SVR and SVRc

We evaluated the performance of the proposed semi-supervised framework with SVRc, a current advanced approach for survival analysis. Additionally, since SVRc is a modification of traditional SVR, we also compared the improvement of our semi-supervised approach combined with SVR. The aim was to assess whether the semi-supervised approach layered on top of basic SVR would match the performance of SVRc. The SVRc parameters per [34] were $C_c^* = 1,\ C_n = 5,\ C_n^* = C_c = 6,\ \varepsilon_c^* = 12,\ \varepsilon_n = 5,\ \varepsilon_n^* = \varepsilon_c = 2$ and correspondingly the SVR parameters were set as $C = 6$ and $\varepsilon = 2$. Results of traditional SVR compared with our proposed approach layered on top of traditional SVR are presented in Table 4. In Table 5 we present the results of SVRc and the proposed semi-supervised framework combined with SVRc.

Table 4: Experimental Results of SVR and SVR with Transduction

SVR		**SVR with Transduction**	
Study 1			
Train CI:	0.74	Train CI:	0.79
Train Sensitivity:	0.68	Train Sensitivity:	0.77
Train Specificity:	0.79	Train Specificity:	0.83

Train Criteria:	**1.28**	**Train Criteria:**	**1.43**
Test CI:	0.71	Test CI:	0.72
Test Sensitivity:	0.59	Test Sensitivity:	0.62
Test Specificity:	0.68	Test Specificity:	0.73
Test Criteria:	**1.11**	**Test Criteria:**	**1.17**
Study 2			
Train CI:	0.74	Train CI:	0.80
Train Sensitivity:	0.76	Train Sensitivity:	0.76
Train Specificity:	0.79	Train Specificity:	0.87
Train Criteria:	**1.34**	**Train Criteria:**	**1.46**
Test CI:	0.71	Test CI:	0.76
Test Sensitivity:	0.68	Test Sensitivity:	0.74
Test Specificity:	0.70	Test Specificity:	0.80
Test Criteria:	**1.19**	**Test Criteria:**	**1.35**
Study 3			
Train CI:	0.70	Train CI:	0.73
Train Sensitivity:	0.69	Train Sensitivity:	0.66
Train Specificity:	0.71	Train Specificity:	0.80
Train Criteria:	**1.19**	**Train Criteria:**	**1.26**
Test CI:	0.64	Test CI:	0.67
Test Sensitivity:	0.61	Test Sensitivity:	0.61
Test Specificity:	0.70	Test Specificity:	0.74

Test Criteria:	**1.07**	**Test Criteria:**	**1.12**
Study 4			
Train CI:	0.64	Train CI:	0.67
Train Sensitivity:	0.63	Train Sensitivity:	0.63
Train Specificity:	0.67	Train Specificity:	0.77
Train Criteria:	**1.06**	**Train Criteria:**	**1.16**
Test CI:	0.62	Test CI:	0.67
Test Sensitivity:	0.55	Test Sensitivity:	0.55
Test Specificity:	0.77	Test Specificity:	0.84
Test Criteria:	**1.04**	**Test Criteria:**	**1.13**

Table 5: Experimental Results of SVRc and SVRc with Transduction

SVRc		**SVRc with Transduction**	
Study 1			
Train CI:	0.84	Train CI:	0.85
Train Sensitivity:	0.86	Train Sensitivity:	0.84
Train Specificity:	0.72	Train Specificity:	0.74
Train Criteria:	**1.46**	**Train Criteria:**	**1.47**
Test CI:	0.74	Test CI:	0.75
Test Sensitivity:	0.69	Test Sensitivity:	0.69
Test Specificity:	0.66	Test Specificity:	0.71
Test Criteria:	**1.20**	**Test Criteria:**	**1.24**

Study 2			
Train CI:	0.91	Train CI:	0.93
Train Sensitivity:	0.90	Train Sensitivity:	0.90
Train Specificity:	0.85	Train Specificity:	0.88
Train Criteria:	**1.68**	**Train Criteria:**	**1.72**
Test CI:	0.83	Test CI:	0.85
Test Sensitivity:	0.84	Test Sensitivity:	0.95
Test Specificity:	0.77	Test Specificity:	0.79
Test Criteria:	**1.48**	**Test Criteria:**	**1.60**
Study 3			
Train CI:	0.74	Train CI:	0.74
Train Sensitivity:	0.66	Train Sensitivity:	0.66
Train Specificity:	0.80	Train Specificity:	0.80
Train Criteria:	**1.27**	**Train Criteria:**	**1.27**
Test CI:	0.68	Test CI:	0.68
Test Sensitivity:	0.55	Test Sensitivity:	0.55
Test Specificity:	0.75	Test Specificity:	0.75
Test Criteria:	**1.09**	**Test Criteria:**	**1.09**
Study 4			
Train CI:	0.68	Train CI:	0.70
Train Sensitivity:	0.75	Train Sensitivity:	0.75

Train Specificity:	0.65	Train Specificity:	0.72
Train Criteria:	**1.17**	**Train Criteria:**	**1.24**
Test CI:	0.70	Test CI:	0.71
Test Sensitivity:	0.64	Test Sensitivity:	0.64
Test Specificity:	0.77	Test Specificity:	0.77
Test Criteria:	**1.19**	**Test Criteria:**	**1.20**

Concentrating on the more accurate assessment of performance in the test sets, in almost all the experiments the semi-supervised transduction framework we have proposed outperforms the underlying regression method; demonstrating the effectiveness of our proposed approach. One exception is in study 3 for SVRc where the approach doesn't improve performance, but doesn't hurt either; it maintains the same level of performance. This is not completely unexpected as the patients in study 3 were part of a study where a concerted effort was made to track patients, resulting in relatively longer follow up time. Hence, the censored time is already a good representation of the outcome and there may have been fewer "true" targets to learn.

One option to explore for study 3 would be to increase the maximum time allowed for transduction. Rather than transduce to the maximum time in the cohort, an even longer time could be chosen, thereby allowing more opportunities for censored cases to be transduced, and perhaps allowing for further improvements in overall results than is currently observed in study 3.

Of note is the fact that even with more information, traditional SVR does not outperform SVRc in study 3; only with the proposed semi-supervised algorithm does it match SVRc's performance.

It is interesting to note that the Cox model in aggregate, while benefitting from the transduction approach, appears to have the least incremental improvement when compared with the other algorithms. In addition, even the performance of the Cox model with transduction is usually worse than the performance of the other approaches with transduction. Combined with the manifest tendency of the Cox model to overfit more in training and have a larger decrease in validation performance, it may be suggested that the more advanced machine learning algorithms are desirable alternatives to the Cox model.

Another observation to note from a practical perspective in conducting these experiments is that the ANN experiments were time-consuming. There was significant trial and error in tuning the ANN parameters, including number of hidden nodes and layers. These are well known issues when working with ANNs, but also complicate the use of this family of algorithms from a practical perspective.

CHAPTER 5

PREDICTING ADVANCED PROSTATE CANCER ENDPOINTS FROM EARLY INDICATOINS VIA TRANSDUCTIVE SEMI-SUPERVISED REGRESSION

Prostate cancer is the most prevalent form of cancer and the second most common cause of cancer morbidity among men in the United States. The most common treatment is the surgical removal of the prostate through a radical prostatectomy (RP). Unfortunately, RP is no guarantee of a cure. Approximately 3-5% of men post-RP experience significant clinical failure (CF) including metastasis and/or death-of-cancer. While CF is a clinically meaningful endpoint, it can often take years to present; and when it does the disease maybe too advanced for effective treatment. Therefore, an earlier endpoint of prostate-specific-antigen-recurrence (PSAR) post-RP is frequently employed as a surrogate. This is however a noisier endpoint, which 15-25% of men experience post-RP. Not everyone with PSAR progresses to the more advanced stage of CF. Since PSAR occurs years earlier though, a physician and patient can start to make complex decisions about treatment options and impact on quality of life. Accurate prognosis is important as it is the principal factor in determining the treatment plan. In prognostic modeling, PSAR data is frequently employed to predict CF [11, 17].

Thus far, our proposed semi-supervised framework for transductive regression has only been applied to directly predict a medical prognostic endpoint. In the present chapter, we consider the interesting and practical problem where an earlier disease endpoint is used to predict a later one. We concentrate on the highly relevant prostate cancer space as, unlike other cancers, prostate cancer has a long multi-year horizon with multiple stages of the disease.

We applied the proposed transduction framework to build post-RP prognostic models using PSAR outcomes to predict the subsequent more advanced disease endpoint of CF. We analyzed three prostate cancer datasets. Dataset 1 [11] consisted of 262 patients with 8 clinical features, 37 of whom experienced PSAR (14% event rate). Dataset 2 [11] from a second institution consisted of 342 patients, 58 of whom experienced PSAR (17% event rate). Dataset 3 [11, 17] consisted of 340 new patients also from the second institution. Dataset 3 was unique because both the early PSAR endpoint and the later CF endpoint were available for all the patients. 43 patients experienced PSAR (13% event rate) and 12 experienced CF (3.5% event rate). Both Datasets 2 and 3 had 9 clinical features. The goal was to assess in Dataset 3 PSAR models built with Datasets 1 and/or 2.

We layered our semi-supervised transduction framework on top of both SVRc and the Cox Model, and compared the performance with and without the transductive semi-supervised regression. We performed two rounds of experiments. In Table 6, we present the first where PSAR models were built with Dataset 1 and validated for both PSAR and CF with Dataset 3. In Table 7 we present the second round where PSAR models were built with Dataset 2 and validated for both PSAR and CF with Dataset 3. As noted earlier

in Chapter 4, we maintained separate training and validation datasets which is the convention in developing medical prognostics. Additionally, as in all earlier experiments the performance metrics were assessed according to the original times; no transduced targets were used in the accuracy assessments.

Table 6: Results training on Dataset 1 and validating on Dataset 3

	SVRc	***SVRc with Transduction***	***Cox Model***	***Cox Model with Transduction***
	PSAR Training Performance			
CI	0.79	0.81	0.80	0.80
Sensitivity	0.77	0.87	0.80	0.70
Specificity	0.76	0.72	0.73	0.85
Criterion	**1.38**	**1.44**	**1.38**	**1.40**
	PSAR Validation Performance			
CI	0.74	0.76	0.77	0.80
Sensitivity	0.79	0.90	0.79	0.69
Specificity	0.62	0.58	0.59	0.75
Criterion	**1.23**	**1.28**	**1.24**	**1.32**
	CF Validation Performance			
CI	0.76	0.78	0.79	0.79
Sensitivity	0.83	1.00	1.00	1.00
Specificity	0.58	0.53	0.57	0.72
Criterion	**1.24**	**1.31**	**1.36**	**1.51**

Table 7: Results training on Dataset 2 and validating on Dataset 3

	Table Column Head			
	SVRc	***SVRc with Transduction***	***Cox Model***	***Cox Model with Transduction***
	PSAR Training Performance			
CI	0.78	0.79	0.80	0.81
Sensitivity	0.77	0.68	0.82	0.77
Specificity	0.73	0.83	0.71	0.75
Criterion	**1.34**	**1.35**	**1.38**	**1.39**
	PSAR Validation Performance			
CI	0.80	0.81	0.82	0.82
Sensitivity	0.74	0.69	0.79	0.79
Specificity	0.72	0.83	0.72	0.82
Criterion	**1.33**	**1.38**	**1..39**	**1.47**
	CF Validation Performance			
CI	0.88	0.88	0.88	0.88
Sensitivity	1.00	1.00	1.00	1.00
Specificity	0.68	0.78	0.68	0.75
Criterion	**1.56**	**1.66**	**1.56**	**1.63**

These prostate cancer experimental results appear to confirm the value of transductive semi-supervised regression for predicting late stage disease endpoints from earlier indications. For data from multiple institutions, existing survival analysis methods manifest an increase in predictive accuracy when the transduction framework is layered on top. In all the experiments, whether we consider SVRc or the Cox model, in training and both validations, the transduction framework improves performance as measured by the defined Criterion. While independent components of the criterion do vary, the

algorithm was designed to optimize the overall criterion, and it has performed outstandingly.

Not only is the accuracy for PSAR improved, but more importantly, CF is better predicted from the PSAR endpoint. In Table 8 there is a significant improvement in validation specificity. This is likely because all the CF patients experienced PSAR and the PSAR assessment of high risk captures them, but it probably also has a high number of false positives since PSAR is a noisier endpoint and not all patients with PSAR experience CF. The accuracy of predicting CF is higher since CF is a more concrete and relevant endpoint.

These results manifest the value of a novel transductive semi-supervised regression framework in the challenging problem of predicting advanced prostate cancer from earlier disease endpoints. This work presents the first innovative application of this recently developed technique for predicting subsequent endpoints from earlier ones and may be useful in other diseases as well, not just prostate cancer.

5.1 Deeper dive on features driving improvement

An interesting question to pose is whether there are differences in the features driving the improved prediction of validation performance for both SVRc and the Cox Model in the semi-supervised framework. We investigated the weights of all the clinical features in the models. It is difficult to compare the weights of a feature across models; the magnitude of the weight only makes sense within the context of a single model. Hence, we normalized the weights in each model by the highest weighted feature, thereby enabling a relative comparison of how important a particular feature is in a model.

Table 8: Weights of Features in Cox Models

	Cox Model		Cox Model with Transduction	
Feature	**Original Weight**	**Normalized Weight**	**Original Weight**	**Normalized Weight**
Clinical Stage	0.314	0.217	0.355	0.316
PSA	0.743	0.513	0.818	0.728
Dominant Biopsy Gleason Grade	-0.198	-0.137	0.134	0.119
Biopsy Gleason Sum	1.448	1.000	1.124	1.000
Dominant Prostatectomy Gleason Grade	0.872	0.602	1.002	0.891
Prostatectomy Gleason Sum	0.073	0.050	-0.139	-0.123
Seminal Vesicle Invasion	0.747	0.516	0.796	0.708
Positive Surgical Margin	0.306	0.212	0.261	0.232
Extra Capsular Extension	0.198	0.137	0.161	0.143

Table 9: Weights of Features in SVRc Models

	SVRc		SVRc with Transduction	
Feature	**Original Weight**	**Normalized Weight**	**Original Weight**	**Normalized Weight**
Clinical Stage	-2.939	-0.120	-4.496	-0.197
PSA	-10.329	-0.423	-8.397	-0.368
Dominant Biopsy Gleason Grade	-3.909	-0.160	-8.193	-0.359
Biopsy Gleason Sum	-24.394	-1.000	-22.826	-1.000
Dominant Prostatectomy Gleason Grade	-4.363	-0.179	-6.433	-0.282
Prostatectomy Gleason Sum	-11.852	-0.486	-8.784	-0.385
Seminal Vesicle Invasion	-23.926	-0.981	-25.963	-1.137
Positive Surgical Margin	-2.778	-0.114	-3.964	-0.174
Extra Capsular Extension	-4.259	-0.175	-3.499	-0.153

One interesting observation to note is that for both models with SVRc and the Cox Model, the dominant prostatectomy Gleason grade and the seminal vesicle invastoin status [11, 17] both have a much higher relative weight in the transduction framework than in the models without the transduction framework for both the Cox model and SVRc. The implication being that perhaps these features in particular are leading to an improved prediction. This is a noteworthy observation, since the roles of both features for predicting CF are very interesting to urologists and oncologists. In particular, the fact that the interaction of these two is intriguing as the dominant prostatectomy Gleason grade is a measure of how advanced the disease is and the seminal vesicle invasion status is a measure of how much the disease has proliferated/expanded into the surrounding

tissue and critical organs around the prostate. This study was not designed to fully explore these insights, but they are worth considering in future work.

CHAPTER 6

THE ROLE OF IMAGING BASED PROSTATE BIOPSY MORPHOLOGY IN A DATA FUSION PARADIGM FOR TRANSDUCING PROGOSTIC PREDICTIONS

A major focus area for precision medicine is in managing the treatment of newly diagnosed prostate cancer patients. For patients with a positive biopsy, clinicians aim to develop an individualized treatment plan based on a mechanistic understanding of the disease factors unique to each patient. Recently, there has been a movement towards a multi-modal view of the cancer through the fusion of quantitative information from multiple sources, imaging and otherwise.

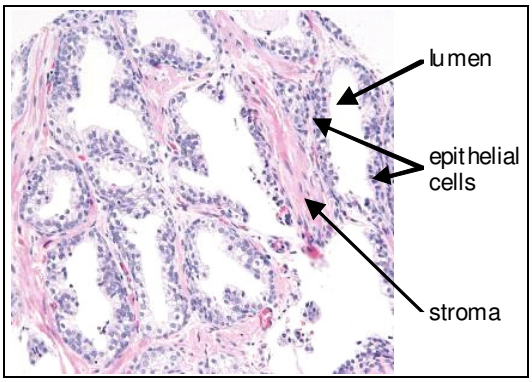
Simultaneously, there have been significant advances in machine learning methods for medical prognostics which integrate a multitude of predictive factors to develop an individualized risk assessment and prognosis for patients.

In this work, we apply our novel semi-supervised approach for support vector regression to predict the prognosis for newly diagnosed prostate cancer patients. We integrate clinical characteristics of a patient's disease with imaging derived metrics for biomarker expression as well as glandular and nuclear morphology. In particular, our goal was to explore the performance of nuclear and glandular architecture within the

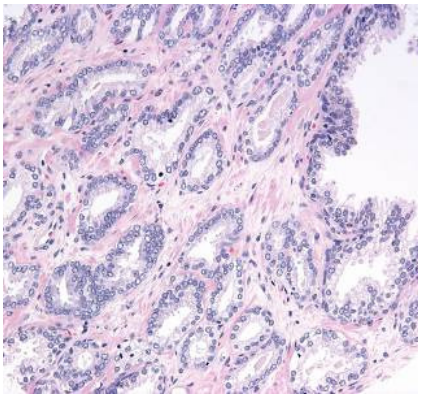
transduction algorithm and assess their predictive power when compared with the Gleason score manually assigned by a pathologist.

6.1 Background

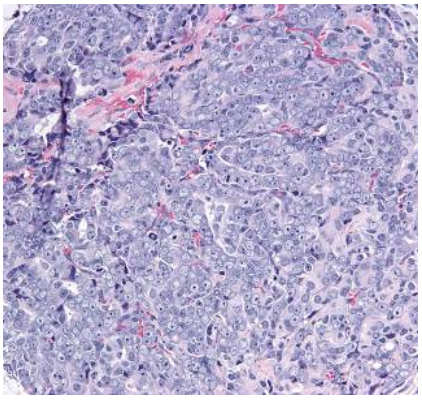
Prostate cancer is the most prevalent form of cancer and the second most common cause of cancer deaths among men in the United States. Accurate prognosis is important as it is the principal factor in determining the treatment plan. Prostate cancer is primarily assessed by the Gleason grading system which classifies the tissue architecture into five patterns of increasing severity [21, 27, 64, 23]. The Gleason grade characterizes tumor differentiation, i.e. the degree of tumor resemblance to normal tissue. In the lower risk Gleason grades of 1 through 3, the architecture consists primarily of isolated or touching gland rings surrounded by fibromuscular stromal tissue. Each gland is composed of a ring of epithelial cells surrounding a duct, the lumen. The connected glandular cytoplasm, or "epithelial unit", contains just one gland ring. As the cancer progresses to grade 4, epithelial units fuse together creating chains of gland rings, or "cribriform" sheets of rings. A second axis of variation in grade 4 and 5 disease is the increasing fragmentation of rings resulting in sheets of isolated cells and non-ring epithelial fragments (the terms "glandular" and "epithelial" are interchangeable). As the cancer progresses, epithelial cells replicate in an uncontrolled manner, disrupting the regular arrangement of gland units.



(a)



(b)



(c)

Figure 7: Samples of H&E stained prostate tissue with varying degrees of differentiation: (a) normal, (b) grade 2 well differentiated cancer associated with favorable outcomes and (c) grade 5 poorly differentiated cancer corresponding to aggressive disease [64].

There has been significant research in automatically approximating the Gleason grade and quantifying other aspects of prostate morphology [64, 23, 20, 62]. Simultaneously, advances have been made in automated quantification of molecular and protein biomarker expression [8, 53]. These quantitative image analyses from multiple modalities have become prevalent, yielding independent prognostic predictors of outcome. In recent years, there has been a trend towards integrating these independent predictors together into a "data-fusion" approach [18, 68, 24]. When combined together, these disparate information modalities provide a more comprehensive and powerful, personalized view of disease prognosis and staging. However, the fusion of these disparate information sources in a multivariate context is not trivial given the censored nature of outcome in survival analysis.

In this work, we explore the interaction of advanced imaging features for prostate morphology and biomarker quantification, with clinical variables, including the Gleason grade, in our novel semi-supervised framework for transduction regression targets in survival analysis. In particular, we aimed to explore the interaction of quantitative morphology with the pathological Gleason score. This represents one of the first explorations of multi-modal data fusion for semi-supervised prognostics.

6.2 Imaging Methods Employed

6.2.1 H&E Morphology

Morphological and architectural characteristics of the prostate tissue, such as epithelial nuclei and cytoplasm, provide critical information for the diagnosis, prognosis and therapeutic decision making of prostate cancer. The subjective and variable Gleason

grade assessed by expert pathologists in Hematoxylin and Eosin (H&E) stained specimens has been the standard for prostate cancer diagnosis and prognosis.

While there has been significant work in automatically approximating the Gleason grade and quantifying other aspects of prostate morphology, the majority of proposed approaches consider various tissue components such as lumens, nuclei and cytoplasm independently. Instead, regarding the entire glandular unit of epithelial nuclei, cytoplasm and stroma around a lumen would provide a more accurate and comprehensive morphological assessment of disease severity.

We leveraged a method proposed by Fogarasi et al. [23] for automated analysis of gland unit features from H&E images. The approach initially segments and classifies primary cellular components such as cytoplasm, nuclei, stromal fibroblasts, lumens, blood vessels and artifacts. This segmentation relies on cellular properties such as distance of tumor cells from lumens, as well as color, shape, texture and neighborhood properties. The relationships between these components are analyzed and leverage to construct distinct "gland units." Biological characteristics, such as logical and relative object positioning are employed to develop initial seeds which are optimized in an iterative classification process.

Gland units are objects created by uniform and symmetric grown around lumens that are seeds. Growth proceeds around these objects through spectrally uniform segmented epithelial cells. The accuracy of the border is determined by differentiating cytoplasm from the remaining tissue. Gland unit creation is thus a controlled object based region growing of epithelial cells. Region growing commences at the lumen boundaries, and continues through the tumor tissue until some biological boundary such

as a tear, stroma tissue, or another growing gland unit is reached. As each growth "ring" is added to the gland, the surrounding epithelial nuclei and cells are evaluated to be "within" or "outside" a gland unit.

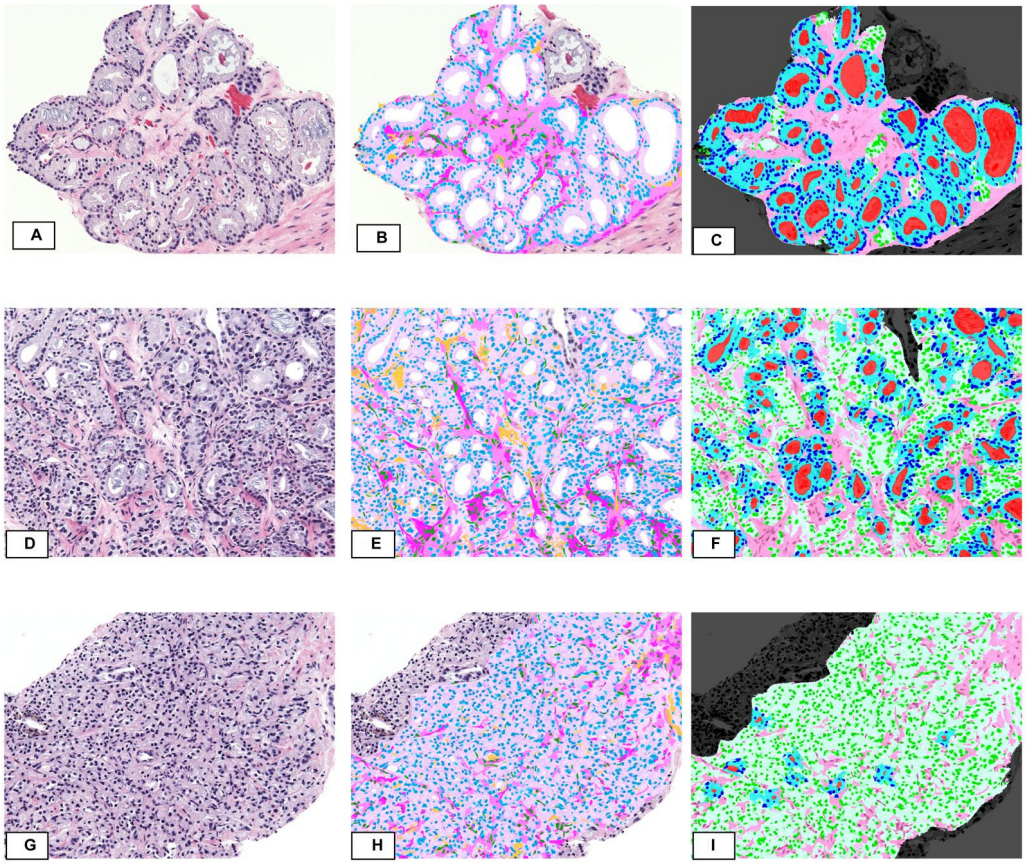


Figure 8: Images representing prostate cancer grades 3 (A-C), 4 (D-F) and 5 (G-I). Images representing the original H&E stain (A, D, G), primary object segmentation (B, E, H) and glandular object classification (C, F, I) are presented [23].

Without the addition of stop conditions, uncontrolled growth of gland units would occur. Consequently small lumens are ignored as gland seeds and the controlled region-

growing algorithm continues in a manner which constrains the collision against other morphological objects. Subsequently, all meaningful cellular components such as epithelial and stromal nuclei are evaluated in relation to these gland units to create morphological features. In Figure 8, we present representative images from this analysis.

6.2.2 IF Morphology and Biomarkers

In multispectral immunofluorescence (IF) microscopy [64, 53, 65], multiple proteins in the tissue specimen are simultaneously labeled with different fluorescent dyes. Each dye has a distinct emission spectrum and its associated antibody binds to its target protein within a tissue compartment (ie nuclei or cytoplasm). The stained slide is illuminated under a fluorescence microscope with a light source for a specific wavelength. This excitation light is absorbed by the fluorescent dye causing it to emit light of a longer wavelength. The intensity of the emitted light is a measure of the target protein's concentration. In multiplexed IF images, the tissue is labeled with several antibodies at the same time. Each antibody is labeled with a unique fluorescent dye with distinct spectral characteristics. The tissue is then imaged with a multispectral camera, then spectrally un-mixed, to yield multiple images with one image per individual dye/antibody. Two common dyes that reveal the tissue structure are DAPI (a nuclear stain) and CK18 (stains epithelial cytoplasm). Nuclear objects are segmented and then separated using a co-localization scheme into epithelial nuclei positive for both DAPI and CK18 and stromal nuclei positive for DAPI but not CK18. Subsequently prognostic biomarkers such as AR (androgen receptor) are evaluated within each co-localized

compartment. Figure 9 illustrates a sample prostate gland unmixed into DAPI, CK18 and AR specific images.

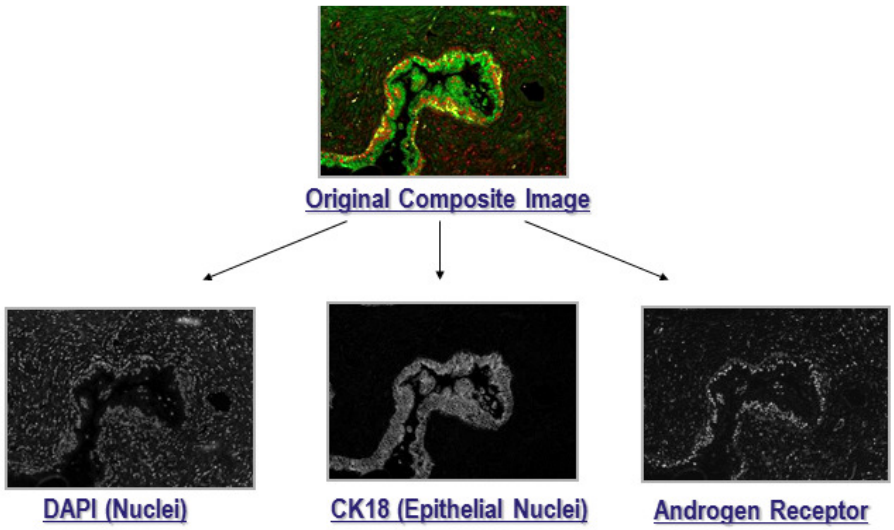
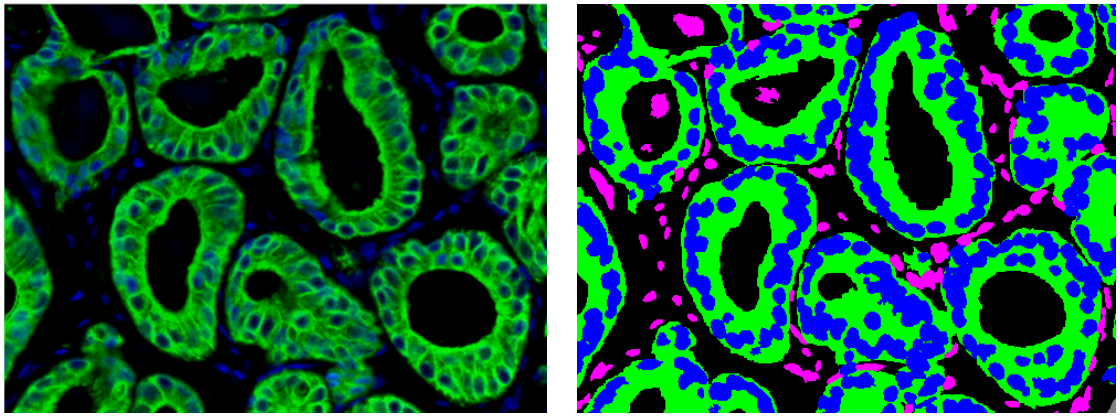


Figure 9: Sample composite image of a prostate gland spectrally unmixed into individual images representing DAPI, CK18 and AR biomarkers [64].

In this work we build upon previous work in IF biomarker quantification [53, 65]. Specifically, we analyzed expression of AR and Ki67 prostate biomarkers as proposed by Sapir et al [53]. Quantification of a biomarker is achieved in two stages. First, a biomarker relevant comparted is detected. Then, the signal is separated from the background within the compartment via intensity thresholding. Following the definition of epithelial and stromal nuclei, as well as epithelial cytoplasm, background autofluorescence and non-specific binding effects are filtered out. An interactive model based thresholding technique is used to classify whether each nuclei is positive for a particular biomarker. The expression of each biomarker can then be quantized and normalized (epithelial signal normalized by stromal expression). Features representing

the relative rise of the biomarker in the epithelial disease specific compartments were recognized to be prognostic as they measure the dynamic range of biomarker expression in an image.



(a) (b)

Figure 10: A multiplex IF pseudo-color image consisting of the DAPI counterstain (blue) and the CK18 biomarker (green); and (b) segmented epithelial nuclei (blue), stroma nuclei (purple) and epithelial cytoplasm (green) [64].

Additionally, these tissue objects can be analyzed for morphological properties such as distance based minimum-spanning-tree (MST) measures, as well as the fractal dimension of the glandular boundaries. MST, fractal and wavelet features proposed by Tabesh et al [64] were employed in this analysis.

The MST connecting the centroids of all epithelial nuclei in the tissue is the basis for extracting feature characterizing tissue architecture. The MST of a graph is defined as the tree connecting all vertices (i.e., epithelial nuclei centroids) such that the sum of the lengths of the lines (edges) connecting the vertices is minimized. Many algorithms exist for constructing the MST of a graph. We used the well-known Prim's algorithm

[12, 64]. Let $G = \{V, E\}$ denote a graph with vertices V and edges E, and let $G_{\text{MST}} = \{V_{\text{MST}}, E_{\text{MST}}\}$ denote the MST of G. The algorithm starts by adding an arbitrary vertex v in V to V_{MST}, that is, $V_{\text{MST}} = \{v\}$. Then, the algorithm finds the nearest vertex in the rest of the graph to the current G_{MST}. That is, the shortest edge e connecting the vertices u and v is found such that $u \in V_{\text{MST}}$ and $v \notin V_{\text{MST}}$. Then, G_{MST} is updated by adding v to V_{MST} and adding e to E_{MST}. The process of adding vertices is continued until all of them are included in V_{MST}. Figure 11 illustrates the MST of the epithelial nuclei in Figure 10.

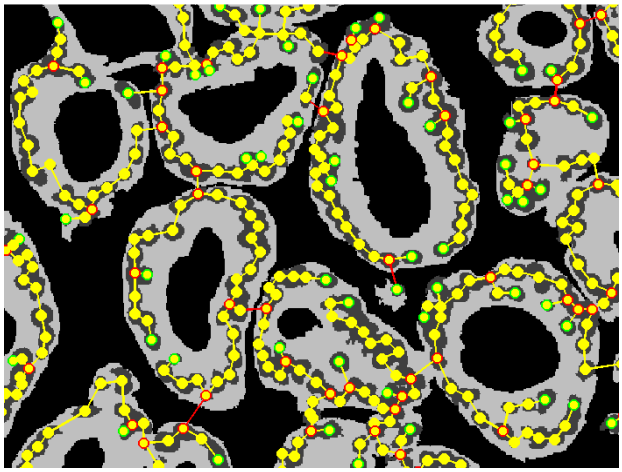


Figure 11: MST connecting the epithelial nuclei in Figure YY. Segmented epithelial nuclei are marked in grey, and stromal nuclei and other compartments are masked out. Epithelial nuclei centroids and intra-gland MST edges are marked in yellow and inter-gland edges are marked in red [64].

The fractal dimension of the boundaries between the glands and the surrounding stroma provides a quantitative measure of the irregularity of the shape of the boundary.

In general, the fractal dimension is a measure of the space-filling capacity of an object. The fractal dimension of a straight line is one, whereas the fractal dimension of a more irregular planar curve is between 1 and 2. Gland boundaries with lumen and stroma are defined as pixels that have at least one non-gland and one gland pixel among their 4-connected neighbors. As lumens and stroma appear similar in our multiplex IF images, we used morphological operations to distinguish them. We defined lumens as pixels belonging to "holes" in the gland regions, i.e., pixels that cannot be reached by flood-filling the non-gland region starting from pixels on the edge of the image. Two features were considered, namely, the fractal dimension of gland-stroma boundaries, and the fractal dimension of gland boundaries with both stroma and lumens. We estimated these features using the box-counting algorithm described in [64]. A detailed description of fractal theory is available in [66].

6.3 Results

We analyzed prostate biopsies from a mult-institutional cohort of 1027 patients. Each patient had clinical data available including age, the clinical stage, Gleason grade and PSA (prostate specific antigen) level. Each patient had up to 3 H&E and 6 IF images captured which were then quantized and analyzed to develop predictive features. We then evaluated all the features in a multi-variate fusion approach leveraging our semi-supervised regression framework with SVRc. We trained the semi-supervised transductive models on 686 patients and validated on 341 patients. Three different types of models were created. Model 1 was solely based on the PSA, clinical Gleason and IF biomarker (AR and Ki67) expression features. Model 2 was developed without the

clinical Gleason and with the morphometric IF and H&E features added. Model 3 was built with all the feature modalities represented. Results are presented in Table 10.

Table 10: Results of all three models in training and test data sets

	Model 1: Clinical and IF Biomarker Features	**Model 2: PSA, IF biomarker and IF and H&E Morphology features**	**Model 3: All Features**
Train CI	0.75	0.76	0.75
Train Sensitivity	0.67	0.70	0.67
Train Specificity	0.77	0.78	0.76
Train Criterion	**1.27**	**1.31**	**1.26**
Test CI	0.69	0.67	0.68
Test Sensitivity	0.48	0.64	0.55
Test Specificity	0.81	0.74	0.80
Test Criterion	**1.08**	**1.14**	**1.11**

As can be observed, the quantitative morphological features not only improved the predictive performance, but removing the pathologist assigned clinical Gleason increases the accuracy of the prediction in both training and test sets. This is likely due to the removal of the subjective, noisy and non-robust manual assessment of the Gleason grade.

6.4 Chapter 6 Summary

This work presents an application of our unique semi-supervised approach for medical prognosis in the context of fusing multi-modal features from positive prostate biopsies. It represents an evaluation of the Gleason score with metrics for morphology derived from quantitative image analysis in this context. The results on a multi-institutional cohort of 1027 prostate biopsy patients indicate that morphometric IF and H&E features when fused with other characteristics in a multi-model framework, improve predictive performance, especially with the absence of a pathologically assigned Gleason score. This is the first exploration of an interaction of advanced imaging features for prostate morphology and biomarker quantification, with clinical variables, including the assessment of quantitative prostate biopsy architecture versus the Gleason grade in the context of a data fusion paradigm which leverages a semi-supervised approach for risk prognosis. We plan further analysis of multi-modal data fusion for semi-supervised prognostics.

CHAPTER 7

ASSESSING THE IMPACT OF PROSTATE BIOPSY TUMOR AMOUNT ON IMAGING BASED PROGNOSTICS EMPLOYING TRANSDUCTIVE SEMI-SUPERVISED REGRESSION

Prostate cancer is the most common form of cancer diagnosed in American men and the second deadliest of all cancers affecting men [17]. Newly diagnosed patients with a positive prostate biopsy and their physicians face a variety of potential treatment options including surgery, radiation therapy, active surveillance, and more. Which option is best for the individual patient is not always clear, and there have been a number of assays developed to analyze a patient's tumor specimen and provide a personalized assessment of cancer severity and risk [4, 15, 16, 18, 44]. Some of these assays employ image analysis algorithms to extract morphometric and biomolecular characteristics from the tumor specimen as features in predictive models for risk assessment. A practical challenge however is that there is often not enough tumor present in the biopsy specimen for analysis. Even if sufficient tumor is present, the amount of cancerous material may affect the accuracy of the predictive models. To the best of our knowledge, there have

been limited published studies on how different amounts of tumor in a prostate biopsy would affect the performance of imaging features in a predictive model [37].

The predictive models for these prognostic assays are often constructed analyzing the features of prognostic risk and predicting the time to cancer progression (including metastasis) based on these disease characteristics. Scientists leverage statistical and machine learning techniques for survival analysis in these endeavors [17, 34, 57]. In this chapter we explore how the prognostic performance of our semi-supervised framework is affected as automated image analysis algorithms extract morphometric and biomolecular features from varying amounts of tumor.

7.1 Background on Prostate Biopsy Image Analysis

For prostate cancer patients with a positive biopsy, clinicians aim to develop an individualized treatment plan based on a mechanistic understanding of the disease factors unique to each patient. Two main information sources are the architecture of the tumor morphology and biomolecular mechanisms of the disease as assessed by biomarkers [90, 17, 23, 53, 64]. There has been significant research in image analysis of prostate morphology as well as automated quantification of molecular and protein biomarker expression [23, 53, 64]. These quantitative image analyses from multiple modalities have become prevalent, yielding not only independent prognostic predictors of outcome, but also features which can be combined into multivariate models [17, 37]. In this work, we explore morphometric features from H&E (hematoxylin and eosin) and IF (immunofluorescent) images, as well as IF biomarker features.

7.1.1 H&E Morphology

Morphological and architectural characteristics of the prostate tissue, such as epithelial nuclei and cytoplasm, provide critical information for the diagnosis, prognosis and therapeutic decision making of prostate cancer. We leveraged a method proposed by Fogarasi et al [23] for automated analysis of gland unit features from H&E images.

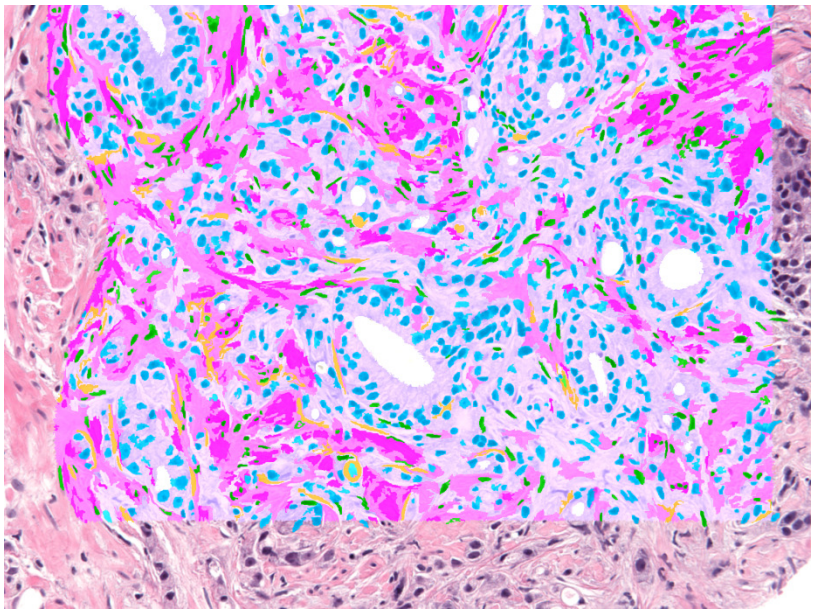
7.1.2 IF Morphology and Biomarkers

In multispectral IF microscopy [53, 64] multiple proteins in the tissue are simultaneously labeled with different fluorescent dyes. Each dye has a distinct emission spectrum and its associated antibody binds to its target protein within a tissue compartment (ie nuclei or cytoplasm). The stained slide is illuminated under a fluorescence microscope with a light source for a specific wavelength. This excitation light is absorbed by the fluorescent dye causing it to emit light of a longer wavelength. The intensity of the emitted light is a measure of the target protein's concentration. Two common dyes that reveal the tissue structure are DAPI (a nuclear stain) and CK18 (stains epithelial cytoplasm). Nuclear objects are segmented and then separated using a co-localization scheme into epithelial nuclei positive for both DAPI and CK18 and stromal nuclei positive for DAPI but not CK18. Subsequently prognostic biomarkers such as AR (androgen receptor) or Ki67 are evaluated within each co-localized compartment.

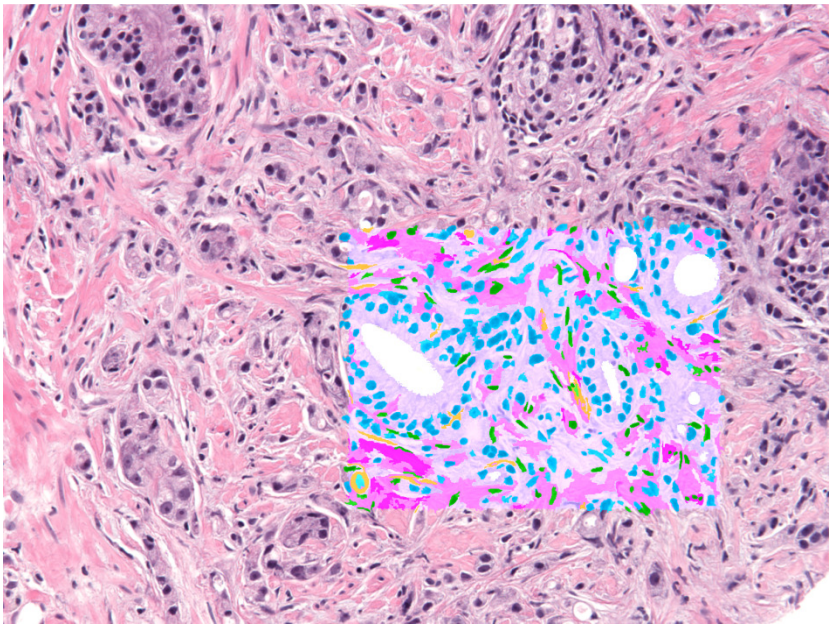
We analyzed expression of AR and Ki67 prostate biomarkers as proposed in [53]. Additionally, tissue objects like epithelial nuclei (DAPI and CK18 positive nuclei) can be analyzed for morphological properties such as distance based minimum-spanning-tree (MST) measures proposed in [64].

7.2 Study Design

The purpose of this study was to assess the impact of decreasing tumor on prostate biopsy prognostic models built with the transductive regression framework. We employed a dataset of 226 patients with positive prostate cancer biopsies [37]. This dataset was from a previous tumor analysis of a SVRc based prostate biopsy model [18] and had concluded that the imaging features were robust down to 20% of the field-of-view Each patient had one H&E image, and two IF images: one for AR and a second for Ki67. All images were acquired at a 20x magnification field-of-view, and had tumor in at least 80% of the image. Images were then masked by expert pathologists using pre-defined masks representing 80%, 60%, 40%, 20% and 10% of the field-of-view. Pathologists identified areas of tumor with these masks which were then analyzed to extract H&G gland unit morphology features, IF MST features and AR and Ki67 biomarker expression features. In summary, three images for 226 patients with five masks of decreasing tumor levels led to a total of 3390 images analyzed in this study. The figures below illustrate sample segmented images for a representative patient at the 80% and 20% mask levels. Variations in the amount of tumor analyzed are evident in these representative samples. It is interesting to note how the different tumor amounts available for analysis changes the segmentation and classification of tissue objects as illustrated by the different colors for the same part of the tissue in the different masks.

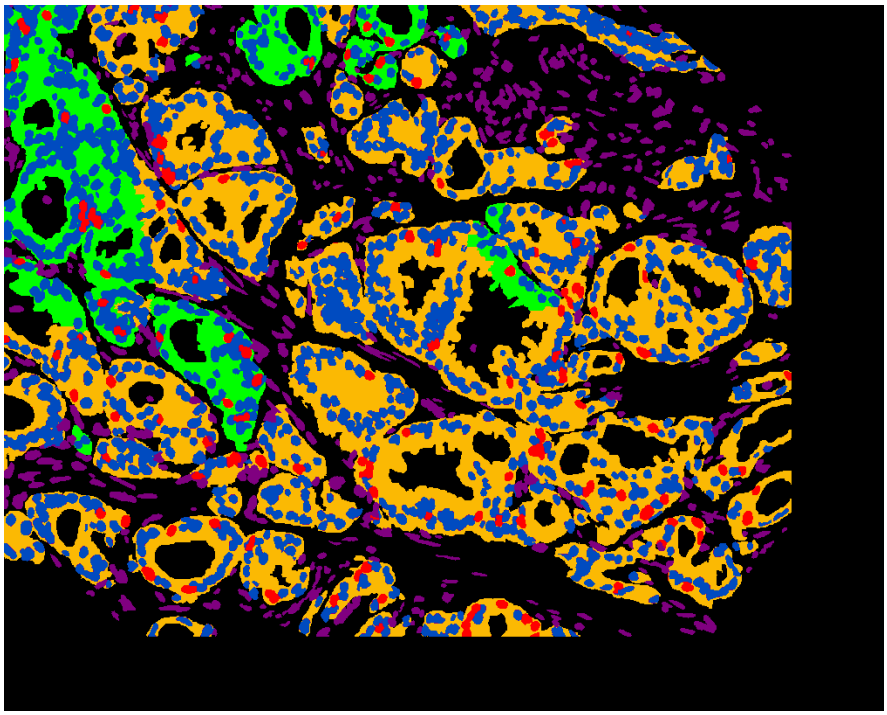


(a)

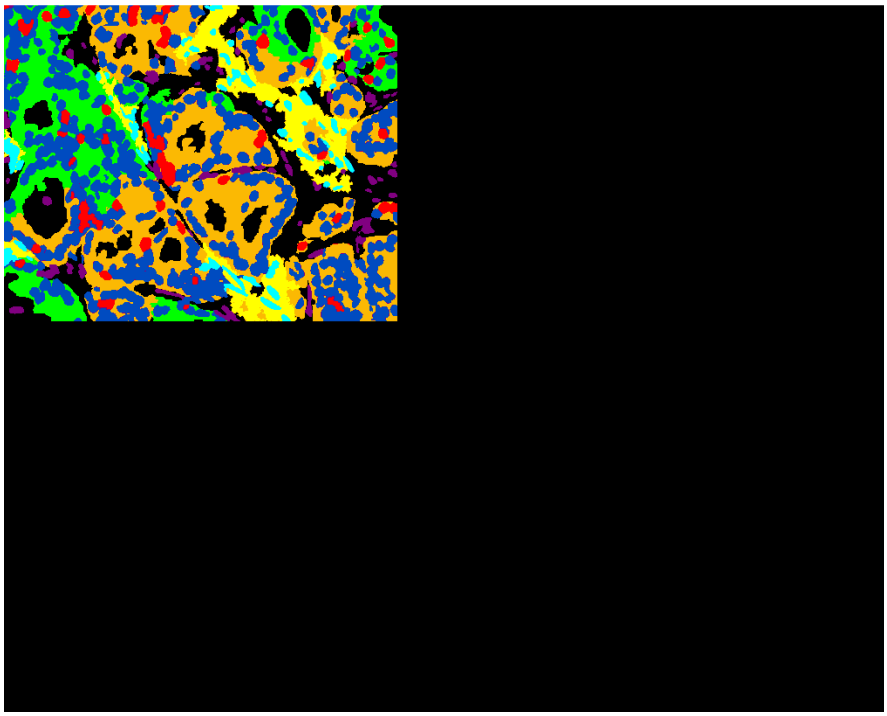


(b)

Figure 12: Segmented H&E image at 80% (a) and 20% (b) tumor mask levels. Regions outside the mask are illustrated as the original H&E image and the analyzed area has segmented components [37].



(a)



(b)

Figure 13: Masked and segmented IF image at 80% (a) and 20% (b) tumor mask levels [37].

Following the extraction of morphological and biomolecular imaging features at the four different tumor mask levels, we then constructed models for prostate cancer progression. First the 226 patient cohort was split into training and validation sets, each with 113 patients. As a reminder, in developing medical prognostics, it is necessary to maintain separate training and validation sets (rather than combined cross-validation type approaches) due to FDA regulatory requirements for independent testing and validation.

We then constructed models to predict significant disease progression (including metastasis) and validated them. A new model was created and validated for each tumor level. We built models with SVRc alone and SVRc combined with our semi-supervised transductive regression framework. We employed the same optimization criterion for the framework which has previously proven successful.

7.3 Experimental Results and Discussion

The complete results of the study are presented in Table 11. These experiments appear to confirm that the semi-supervised transductive regression framework for survival analysis performs better with reduced amounts of tumor in the prostate biopsy. At all tumor levels, the accuracy in the validation set of the performance criterion is higher when the framework is layered on top of SVRc rather than with SVRc alone. In fact, for all tumor levels except 10%, the accuracy is better in training as well. While independent components of the criterion do vary, the algorithm is designed to optimize the overall criterion, in which it has succeeded.

Table 11: Results of training and testing SVRc models at decreasing tumor levels with and without the semi-supervised transduction framework

Tumor Mask Level	80%	60%	40%	20%	10%
SVRc Training Performance					
CI	0.70	0.72	0.73	0.76	0.67
Sensitivity	0.64	0.82	0.73	0.82	0.73
Specificity	0.78	0.68	0.82	0.70	0.67
Criterion	**1.20**	**1.28**	**1.33**	**1.33**	**1.16**
SVRc Validation Performance					
CI	0.69	0.68	0.66	0.59	0.68
Sensitivity	0.73	0.73	0.55	0.27	0.64
Specificity	0.77	0.67	0.73	0.75	0.59
Criterion	**1.25**	**1.17**	**1.06**	**0.79**	**1.06**
SVRc with Transduction Training Performance					
CI	0.74	0.76	0.75	0.79	0.67
Sensitivity	0.82	0.82	0.82	0.73	0.73
Specificity	0.75	0.80	0.75	0.85	0.67
Criterion	**1.36**	**1.42**	**1.37**	**1.41**	**1.16**
SVRc with Transduction Validation Performance					
CI	0.70	0.69	0.68	0.60	0.68
Sensitivity	0.73	0.64	0.82	0.45	0.73
Specificity	0.77	0.83	0.68	0.62	0.64
Criterion	**1.26**	**1.22**	**1.24**	**0.88**	**1.15**

It is interesting to note that the validation performances are not very different at the 80% tumor level when there is a significant amount of tumor available to analyze for the imaging features. But as the available tumor amount decreases, the value of the semi-supervised framework becomes apparent. Another observation is that there is a decrease

for both SVRc alone and with the transduction framework at the 20% level. The transduction framework still does better, but this may be due to some artifact in the imaging features which is emerging at the 20% tumor level.

These are the results of a robust study designed to assess the impact of decreasing tumor amounts on prostate biopsy prognostic assays. The results have proven the value of the proposed semi-supervised transductive regression framework for building prostate biopsy prognostic models with imaging features extracted from progressively smaller proportions of tumor in the biopsy specimen. The study analyzed different imaging domains, with prognostic features for each domain analyzed for five different masks of decreasing size. This works represents one of the few published studies of how different proportions of tumor can affect prostate cancer assays, and oncology predictive assays in general. We would urge scientists developing these assays and clinicians using them that such robustness and sensitivity studies should carried out regularly. Our results suggest that a semi-supervised transductive regression framework for survival analysis may be beneficial in ensuring robust results as the amount of tumor available for analysis decreases.

CHAPTER 8

SUMMARY AND CONCLUSION

While model-based medical survival analysis has been employed by biostatisticians since the 1970s, modern machine learning approaches such as NNci and SVRc can improve the predictive power of such analyses. Despite the fact that these prediction methods have multiple ways of accounting for censored cases, none of them up to now had employed semi-supervised approaches to leverage the partial information about survival endpoint times. In this dissertation, we provided evidence that a transduction framework when combined with machine learning methods can be a powerful tool for improving the accuracy of survival analysis in a range of medical prognostic problems. Interestingly, the core ideas behind the proposed approach are scalable with a large variety of regression algorithms and can be applied to a wide scope of survival analyses.

It seems straightforward to expand our transduction framework to work with other regression algorithms. There are a variety of such algorithms to explore including the ones described in Chapter 2 as well as recent promising innovations such as Deep Survival which proposes a hybrid mix of deep learning and the Cox Model [32], or Conditional Random Fields that take into account the sequential structure of the input

data and/or the labels. Additionally, we anticipate that further tuning the SVRc parameters and NNci neural network architectures could also yield improved results.

This dissertation has presented a new methodology validated empirically through a rather constrained set of clinical results. Motivated by these promising evidence-based analytics, future areas of development could include the elucidation of more theoretical foundations for this line of research. Specifically, two main opportunities arise. The first one is to develop a more theoretical understanding of the joint CI and sensitivity/specificity criterion which could help clarify how we can assess performance under semi-supervised leaning formulations of dependency-state-networks (ie conditional random fields or other deeper semantically-motivated causal or associated categorical temporally-constrained networks) or alternatively time series model-based predictions.

An interesting extension of this performance criterion would be to consider each component (CI, sensitivity and specificity) as an axis in a three dimensional space, and the overall criterion to optimize would be a (perhaps Euclidean) distance of a point representing these three metrics from the origin. As performance of each metric improves, the distance would increase. Ideally, such an approach would be robust to improvements in one metric over the others. This would also have the added benefit of allowing visualization of performance, and perhaps even the construction of manifolds in the three dimensional space as different models are constructed. Another alternative would be to introduce weighting parameters to differentially weigh the CI and the product of sensitivity and specificity. This would allow more control over the importance of the individual metrics within the overall criterion.

The second set of opportunities to pursue is more controllable, through abstracted and simplified simulations. These would allow us to explore the behavior of the transduction framework in various settings where there is more control of the data generation under various model assumptions than what we had in the experiments analyzed in the presented dissertation. In this way, we could investigate how different rates of censoring, either through simulated datasets or artificially introduced, might affect performance under different types and degrees of structure in a model's assumed dependencies. Based on the results presented in previous chapters, we speculate that our approach will prove its value in problems with high degrees of censoring, but this needs further investigation. Relatedly, how should one subsample the feature subsets, and how this affects the results, remain to be seen. When assumed dependency structures in an underlying model are strong, one can also assume that predictions will be better accounted for and the transductive "guessing" of outcome end-points less necessary for these kinds of problems. An interesting question is whether one could identify "translational" situations where transduction might still help, though not as much as in weakly structured problems with data censoring. Expanding on the areas of research described in the previous paragraph, as such models are constructed simulating various situations, a manifold of the how the performance criterion behaves in a three dimensional space would facilitate a geometric interpretation of the results.

Our experimental applications were concentrated on oncological problems, with data concerning prostate and breast cancer outcomes. The proposed approach in this dissertation could be extended to any type of disease state where survival is a major problem, and for a variety of medical prognostic applications. Furthermore, there is no

reason to believe that the proposed transduction framework need be limited to the medical domain. On the contrary, it could be employed for survival or failure-time analysis problems in industrial manufacturing, customer churn prediction, reliability and the induced or controlled analysis of equipment failures, among others.

It is important to note that this dissertation does not address the crucial issue of feature selection. Being motivated by and focused on medical problems, these present with great difficulties and expense in assessing underlying states, or measurable features such as gene expressions which usually require time and non-trivial costs for assays to be run, hence maintaining performance with a minimal feature set is of paramount importance. This dissertation's research does not suggest how the many different approaches to feature selection could affect the transduction results in systematic ways. Feature selection procedures could be executed prior to the transduction of targets in order to choose the best feature set to work with, and again would also vary with any prior assumptions of underlying (hidden) state dependencies that could help structure and constrain the expected temporal model results in more detail. On one hand, the advantages of this approach are a reduction in overall computational complexity and that the transduction will be conducted on supposedly meaningful features, further reducing the impact of noise in the data. On the other hand, the disadvantages are that the translation of features across different models is not always straightforward and prior feature selection could yield less informative features, as they would be derived from a suboptimal, non-transduced model. While many of the features would be the same, it is possible that some features which may not have been selected may be important in the context of the final transduced model, and other features which may initially be important

may lose their relevance following the generation of improved targets. Alternatively, if feature selection is performed after transduction, the features selected will be better suited for the transduced model. However, the disadvantages are that the transduction procedure would contain potentially noisy and unnecessary dimensions that could adversely affect performance and would be less efficient due to increased complexity. Existing wrapper type approaches for feature selection in semi-supervised classification could be employed [50]. An interesting idea would be to adapt a Laplacian score for feature selection which has already been explored in the semi-supervised regression setting [19], though the generality of these results is hard to assess.

Adapting the incremental period for the target time in each iteration could also be beneficial empirically in our approach. In the current set of experiments an interval of 10 months was heuristically derived given the length of the maximum time in the training cohorts, the average censored time and observed execution time of the program.

An important point is that our work at present only addresses right-censoring in non-event patients. As discussed in the introduction, survival analysis is further complicated by the left-censored nature of events. Therefore, another area of future research could be to extend our work to transduce the event times as well, decreasing them slightly in order to improve performance.

Additional ideas under consideration are to leverage the metrics such as the internal SVR error, instead of the external model evaluation criteria, for selecting the best model. Additionally, each instance is now transduced independently. It is worth testing whether having a dependent order of transduction would improve the results: first transducing the censored instance with the highest SVR error, and then keeping its

derived target value when moving on to the case with the second highest error, etc. This is an approach related to active learning concepts [55] where one aims to achieve as high a predictive accuracy with as few labeled instances as possible.

Since it is computationally intensive to conduct an exhaustive search of the target space, perhaps five to ten instances with either the highest error or the maximal weight (SVM alpha values) could be selected and an exhaustive search could be conducted within this target space. While this would attempt to mitigate the singular nature of our proposed approach, it would also increase its computational complexity. Hence, further algorithmic optimization to reduce the computational complexity is also a crucial area of future work. The proposed approach scales with the number of censored cases being transduced, the length of the maximum time in the training cohort, and the time interval of increase in each iteration.

Showing initially promising sets of results with notably improved overall prediction performance over existing methods, the proposed transduction framework approach is, to our knowledge, the first application of semi-supervised learning to survival analysis. As noted from the concluding comments and suggestions for futures discussed above, this dissertation's approach and its related lines of research suggest the value of further investigation and research.

In summary, we have presented a novel semi-supervised approach for transducing regression targets in survival analysis problems, with a focus on medical prognosis. Our method can be combined with almost any regression algorithm, whether designed for survival analysis or not. The innovative procedure manifests a marked improvement in the performance of current algorithms. In experiments representing prostate and breast

cancer, our proposed method has outperformed the current leading algorithms for survival analysis. Additionally, the method has proven its utility in various medical prognostic applications where survival analysis algorithms are employed, such as building models for late stage disease endpoints from earlier indications, evaluating the interaction of quantitative image analysis metrics with clinical characteristics in a data fusion paradigm, and assessing the impact of decreasing tumor in a prostate biopsy assay. This dissertation represents one of the first applications of semi-supervised learning for survival analysis and has introduced the notion of leveraging the partial knowledge of true outcome in censored times.

REFERENCES

[1] Bair, E., and Tibshirani, R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLOS Biology. 2 (2004), 511-522.

[2] Belkin, M., Niyogi, P. and Sindhwani, V. Manifold Regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research. 9 (2006), 2399-2434.

[3] Bennet, K. and Demiriz, A. Semi-Supervised support vector machines. Advances in Neural Information Processing Systems, 11 (1999), 368-374.

[4] Blume-Jensen, P., Berman, D., Rimm, D., et al. Development and Clinical Validation of an in situ Biopsy Based Multi-Marker Assay for Risk Stratification in Prostate Cancer. Clinical Cancer Research, 2015.

[5] Brown, S.F., Branford, A.J., and Moran, W. On the use of artificial neural networks for the analysis of survival data. IEEE Trans. on Neural Networks, (1997), 1071-1077.

[6] Burges, C. A Tutorial on Support Vector Machines for Pattern Recognition. Knowledge Discovery and Data Mining. Vol. 2 (1998), 121-167.

[7] Burke, H. B., Goodman, P. H., Rosen, D. B. et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, (1997), 857-862.

[8] Camp, R., Chung, G., and Rimm, D., Automated subcellular localization and quantification of protein expression in tissue microarrays, Nature Medicine, vol. 8, 1323-1327 (2002).

[9] Chapelle, O., Sindhwani, V. and Keerthi, S. Optimization techniques for semi-supervised support vector machines. Journal of Machine Learning Research. 9 (2008), 203-233.

[10] Chen, Y., Wang, G. and Dong, S. Learning with progressive transductive support vector machines. Pattern Recognition Letters. 14 (2003), 1845-1855.

[11] Cordon-Cardo, C., Kotsianti, A., Donovan, M.J., Capodieci, P., Verbel, D.A., et al. Improved prediction of prostate cancer recurrence through systems pathology, J. Clin. Invest., Vol. 117 (2007), 1876-1883.

[12] Cormen, T. H., Leiserson, C.E., Rivest, R.L. and Stein, C. Introduction to Algorithms, 2nd ed. MIT Press, Cambridge, MA, 2001.

[13] Cortes, C. and Mohri, M. 2007. On Transductive Regression. Advances in Neural Information Processing Systems (NIPS), 2007

[14] Cox, D. R. Regression Models and Life-Tables (with Discussion). Journal of the Royal Statistical Society. (1972), pp. 187-220.

[15] Cullen, J., Rosner, I.L., Brand, T.C., et al. A Biopsy-based 17-gene Genomic Prostate Score Predicts Recurrence After Radical Prostatectomy and Adverse Surgical Pathology in a Racially Diverse Population of Men with Clinically Low- and Intermediate-risk Prostate Cancer. European Urology. 68 (2015), pp. 123-131.

[16] Cuzick, J., Stone, S., Fisher, G. et al. Validation of an RNA cell cycle progression score for predicting death from prostate cancer in a conservatively managed needle biopsy cohort. British Journal of Cancer. 11 (2015), pp. 382-389.

[17] Donovan, M., Hamann, S., Clayton, M. et al. A Systems Pathology Approach for the prediction of prostate cancer progression after radical prostatectomy, J. Clin. Oncol., 28 (2008), pp. 3923-3929.

[18] Donovan, M., Khan, F.M., Fernandez, G. et al. Personalized Prediction of Tumor Response and Cancer Progression from the Prostate Needle Biopsy. J. of Urol. 182 (2009), pp. 125-132.

[19] Doquire, G. and Verleysen, M.. A Graph Laplacian Based Approach to Semi-Supervised Feature Selection for Regression Problems. Neurocomputing, 121 (2013) pp. 5-13.

[20] Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., and Tomaszeweski, J., Automated grading of prostate cancer using architectural and textural image features. Proc. IEEE Int. Symp. Biomed. Imaging, (2007), pp. 1284-1287.

[21] Epstein, J., [Prostate Biopsy Interpretation], Lippincott-Raven, Philadelphia (1995).

[22] Evers, L. and Messow, C.M. Sparse kernel methods for high-dimensional survival data. Bioinformatics, 24 (2008), 1632-1638.

[23] Fogarasi, S., Khan, F.M., Pang, H., et al. Glandular Object Based Tumor Morphometry in H&E Biopsy Samples for Prostate Cancer Prognosis. Proc. SPIE 7963. (2011).

[24] Fourati, H. [Multisensor Data Fusion: From Algorithms and Architectural Design to Applications], CRC Press, Boca Raton & London & New York, (2016).

[25] Fung, G. and Mangasarian, O. Semi-supervised support vector machines for unlabeled data classification. Optimization Methods and Software. 15 (2001).

[26] Gammerman, A., Vovk, V. and Vapnik, V. Learning by Transduction. Proceedings on Uncertainty in Artificial Intelligence. 1998, 148-156.

[27] Gleason, D., ["The veteran's administration cooperative urologic group: Histologic grading and clinical staging of prostatic carcinoma," Urologic Pathology: The Prostate], Lea and Febiger, Philadelphia, 171-198 (1977).

[28] Gordon, L. and Olshen, R. Tree-structured survival analysis. Cancer Treatment Reports. 69 (1985), 1065-1068.

[29] Harrell, F. E. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York. 2001.

[30] Harrell, F.E., Klee, K., Califf, R., Pryor, D. and Rosait, R. Regression modeling strategies for improved prognostic prediction." Stat. Med. 95 (1984), 634-635.

[31] Herbich, Ralf. Learning Kernel Classifiers: Theory and Algorithms. Cambridge: The MIT Press, (2002).

[32] Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T. and Kluger, Y. Deep Survival: A Deep Cox Proportional Hazards Network. Workshop on Computational Biology, International Conference on Machine Learning, ICML 2016. June 2016.

[33] Kemp, C., Griffiths, T., Stromsten, S., and Tenenbaum, J. Semi-supervised learning with trees. Advances in Neural Information Processing Systems. 16 (2004).

[34] Khan, F. and Zubek, V. Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. Proceedings of the Eighth IEEE International Conference on Datamining, ICDM08. 2008, pp. 863-868.

[35] Khan, F. and Liu, Q. Transduction of Semi-Supervised Regression Targets in Survival Analysis for Medical Prognosis. Biological Data Mining Workshop (BioDM), IEEE 11th International Conference on Data Mining ICDM 2011. December 2011, pp 1018-1025.

[36] Khan, F. and Liu, Q. Medical Survival Analysis Through Transduction of Semi-Supervised Regression Targets. International Journal of Knowledge Discovery in Bioinformatics. 2011, 2:3, pp. 52-65.

[37] Khan, F.M., Fogarasi, S.I., Powell, D., Fernandez, G., Mesa-Tejada, R. and Donovan, M. An Analysis of the Impact of Tumor Amount on the Predictive Power of a Prostate Biopsy Predictive Assay. SPIE Proceedings of Medical Imaging 2011, vol 7966. 2011.

[38] Khan, F.M. and Kulikowski, C. Survival Analysis via Transduction for Semi-Supervised Neural Networks in Medical Prognostics. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015.

[39] Khan, F.M. and Kulikowski, C. The Role of Imaging Based Prostate Biopsy Morphology in a Data Fusion Paradigm for Transducing Prognostic Predictions. SPIE Proceedings of Medical Imaging 2016. March 2016.

[40] Khan, F.M. and Kulikowski, C. Transductive Semi-Supervised Survival Analysis in Medical Prognostics. Workshop on Computational Biology, International Conference on Machine Learning, ICML 2016. June 2016.

[41] Khan, F.M. and Kulikowski, C. Predicting Advanced Prostate Cancer Endpoints from Early Indications via Transductive Semi-Supervised Regression. IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). June 2016**.**

[42] Khan, F.M. and Kulikowski, C. Impact of Prostate Biopsy Tumor Amount on Imaging Based Prognostics Employing Transductive Semi-Supervised Regression. 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC16. August, 2016**.**

[43] Khosla, A., Cao, Y., Lin, C., Chiu, H., Hu, J., and Lee, H. An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 183-192, 2010.

[44] Klein, E., Haddad, Z., Yousefi, K. et al. Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. Urology, 90 (2016), pp. 148-152..

[45] Land, W.H., Qiao, X., Margolis, D., and Gottlieb, R. A new tool for survival analysis: evolutionary programming/evolutionary strategies (EP/ES) support vector regression hybrid using both censored/non-censored (event) data. Procedia Computer Science (6), pp 267-272, 2011.

[46]] LeBlanc, M., and Crowley, J. Relative risk trees for censored survival data. Biometrics 48, 411-425.

[47] Li, H., and Luan, Y. Kernel Cox regression analysis for linking gene expression profiles to censored survival data. Proceedings of the Pacific Symposium on Biocomputiong 2003, 65-75.

[48] Mangasarian, O., Street, W.N., and Wolberg, W.H. Breast Cancer Diagnosis and Prognosis via Linear Programming. Mathematical Programming Technical Report 94-10, University of Wisconsin.

[49] Raykar, V., Steck, H., Krishnapuram, B., Dehing-Oberije, C., and Lambin, P. On Ranking in Survival Analysis: Bounds on the Concordance Index. Advances in Neural Information Processing Systems. 20 (2008), 1209-1216.

[50] Jiangtao, R., Zhengyuan, Q., Wei, F., Cheng, H., Yu, P.. Forward Semi-supervised Feature Selection. Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD'08, pp970-976, 2008.

[51] Rwebangira, M. and Lafferty, J. Local linear semi-supervised regression. Technical report, School of Computer Sciences, Carnegie Mellon University, CMU-CS-08-999.

[52] Saidi O., Cordon-Cardo C., Costa J. Technology Insight: Will systems pathology replace the pathologist? Nature Clinical Practice. 4 (2006), 39-45.

[53] Sapir, M., Khan, F., Vengrenyuk, Y., Fernandez, G., Mesa-Tejada, R., Hamman, S., Teverovskiy, M., and Donovan, M. Improved Automated Localization and Quantification of Protein Multiplexes via Multispectral Fluorescence Imaging in Heterogeneous Biopsy Samples. Proc. IEEE Int. Symp. Biomed. Imaging, 157-160 (2010).

[54] Segal, M. Regression trees for censored data. Biometrics. 44 (1988), 35-48.

[55] Settles, B. Active Learning Literature Survey, Computer Sciences Technical Report 1648. University of Wisconsin–Madison, http://pages.cs.wisc.edu/~bsettles/pub/settles.activelearning.pdf, retrieved 2010-09-14.

[56] Seeger, M. A taxonomy of semi-supervised methods. Chappelle, O., Scholkopf, B., and Zien, A. Eds. Semi-Supervised Learning. MIT Press, 2006.

[57] Shiao, H., and Cherkassky, V. Learning Using Privileged Information (LUPI) for Modeling Survival Data. 2014 International Joint Conference on Neural Networks (IJCNN), July 2014, 1042-1049.

[58] Shivaswamy, P., Chu, W., and Jansche, M. A Support Vector Approach to Censored Targets. Seventh IEEE International Conference on Data Mining. 2007, 655-660.

[59] Smola, A., and Scholkopf, B. A Tutorial on Support Vector Regression. ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2.

[60] Smola, A., and Scholkopf, B. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press.2002.

[61] Snow, P., Smith, D.S., and Catalona, W.J. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. J. Urology. 152 (1997), 1923-1926.

[62] Sparks, R., Madabhushi, A., Statistical shape model for manifold regularization: Gleason grading of prostate histology. Computer Vision and Image Understanding 117 (2013), , pp. 1138-1146.

[63] Szummer, M. and Jaakkola, T. Partially labeled classification with Markov random walks. Advances in Neural Information Processing Systems. 2000, 945-952.

[64] Tabesh, A., Vengrenyuk, Y., Teverovskiy, M., Khan, F., Sapir, M., Powell, D., Mesa-Tejada, R., Donovan, M.. and Fernandez, G. Robust Tumor Morphometry in Multispectral Fluorescence Microscopy. Proc. SPIE 7260, (2009).

[65] Teverovskiy, M., Vengrenyuk, Y., Tabesh, A., et al. Automated Localization and Quantification of Protein Multiplexes via Multispectral Fluorescence Imaging. Proc. IEEE Int. Symp. Biomed. Imaging, 300-303 (2008).

[66] Theiler, J. "Estimating fractal dimension," J. Opt Soc. Am. A, vol. 7, pp. 1055-1073, 1990.

[67] Therneau, T. and Grambsch, P. Modeling Survival Data: Extending the Cox Model. NY: Springer-Verlag, 2000.

[68] Tiwari, P., Viswanath, S., Lee, G., and Madabhushi, A. Multi-modal data fusion schemes for integrated classification of imaging and non-imaging biomedical data. Biomedical Imaging: From Nano to Micro, 2011 IEEE International Symposium on (2011).

[69] Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Support vector machines for survival analysis. In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (Plymouth, England, July 2007). CIMED2007. pp. 1-8.

[70] Vapnik, V. Statistical Learning Theory. Wiley and sons. 1998.

[71] Yan, L., Verbel, D., and Saidi, O. Predicting prostate cancer recurrence via maximizing the concordance index. KDD04 Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 479-485.

[72] Zhou, Z and Li, M. Semi-supervised regression with co-training. International Joint Conference on Artificial Intelligence, 2005.

[73] Zubek, V.B., Verbel, D., and Saidi, O. Censored Time Trees for predicting time to PSA Recurrence. Proceedings of the Fourth International Conference on Machine Learning and Applications, 221-226. 2005.

[74] Zupan, B., Demsar, J., Kattan, M.W., et al. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial Intelligence in Medicine, 20 (2000), pp. 59-75.