Survival Analysis of Patients with Colorectal Cancer using Semi-supervised Learning

by

Vaishali Gadhiya

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science (MSc) in Computational Science

The Faculty of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis Titre de la thèse	Survival Analysis of Patients with Colorectal Cancer using semi-supervised Learning	
Name of Candidate Nom du candidat	Gadhiya, Vaishali	
Degree Diplôme	Master of Science	
Department/Program Département/Programme	Computational Sciences	Date of Defence Date de la soutenance February 24, 2021

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Peter Adamic
(Committee member/Membre du comité)

Dr. Rajan Patel
(External Examiner/Examinateur externe)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Dr. Lace Marie Brogden
Madame Lace Marie Brogden
Acting Dean, Faculty of Graduate Studies
Doyenne intérimaire, Faculté des études supérieures

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

Abstract

A present study introduces survival analysis for patients with Colorectal cancer. In cancer survival analysis gene selection is a significant task. The most significant invention in the clinical cancer research is to diagnose cancer more precisely dependent on the patient's gene expression profiles. For classification of High-risk and low-risk or survival time prediction for the patient's analytic operation, Cox proportional hazards model (COX) and accelerated failure time model (AFT) have been universally embraced. Limited number of samples and censored data are a major setback for training powerful and exact Cox classification model. Also, comparative phenotype tumors and prognoses are completely different diseases at the genotype and sub-molecular level. Subsequently, the utility of the AFT model for the survival time forecasting is restricted when such natural contrasts of the maladies have not been properly distinguished. To attempt to conquer these two fundamental issues, a novel semi-supervised learning technique has been implemented in this thesis, considering the Cox and AFT models to precisely foresee the treatment hazard and the endurance time of the patients. Furthermore, to choose the relevant genes that associate with the disease, the $L_{1/2}$ regularization approach has been used in the semi-supervised learning method. Semi-supervised learning model can powerfully improve the predictive performance of Cox and AFT models in endurance examination prove in the results of simulation experiments. These methods have been effectively applied on simulated data, Diffuse large B-cell lymphoma (DLBCL_2002) microarray gene expression and clinical datasets. These methodologies were tested on new real microarray gene expression and clinical datasets of Colorectal Cancer. The upsides of the proposed semi-supervised learning technique include: Increase in the number of training samples that are available from the censored data, high capacity for distinguishing the endurance hazard classes of patients in Cox, high anticipating precision for patient's endurance time in AFT model and robust efficiency of the proper gene selection. Semi-supervised learning model is one more applicable tool for endurance examination in clinical cancer research. The semi-supervised learning method was seen to be very strong in the detection of the correct simulated genes especially when the gene expressions are independent. The analysis was performed on the real data and the results showed the semi-cox is superior compared with single cox, single AFT and semi-AFT models.

Keywords: Cancer Survival Analysis, Endurance Examination, Semi-supervised learning, Gene Selection, Regularization, Cox-Proportional hazards model, Accelerated Failure Time Model

Acknowledgements

Firstly, I would like to thank my God, who got me this far, who blessed me with the right people to help me during the different stages of my study.

It gives me great pleasure to express my sincere gratitude to my thesis advisor Professor Dr. Kalpdrum Passi for the support of my master's study and his encouragement, valuable suggestions, discussion, guidance throughout my graduate studies. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a more desirable advisor and mentor for my master's study. Without his guidance and motivation this thesis would not have been possible.

I am also grateful to all my friends here in Sudbury and my friends in India for their encouragement and to help change my career path. I could not have achieved this without their help.

I wish to express my deepest appreciation and thank to my loving and caring parents and my dear, brothers for presenting me with constant support and continuous encouragement throughout my years of study, researching and financial help. This accomplishment would not have been possible without them. Thank you.

DEDICATIONS

This thesis is dedicated to my parents,

For their endless love, support & encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

WHO	World health Organization
MRI	Magnetic Resonance Imaging
CT	Computerized Tomography
COX	Cox Proportional Hazard
AFT	Accelerated Failure Time
LDA	Latent Dirichlet Allocation
EDA	Exploratory Data Analysis
SVM	Support Vector Machine
ROC	Receiver Operating Characteristics
GDC	Genomic Data Commons
MLP	Multi-layer Perception
DASA	Deep Active Survival Analysis
EHR	Electronic Health Records
PPI	Palliative Prognostic Index
IBS	Integrated Brier Score
CI	Concordance Index

Chapter 1

Introduction

As per the statistics from World health Organization (WHO) cancer is second leading cause of death world wild and is responsible for an estimated 9.6 million deaths in 2018 [1]. World-wide, about 1 out of 6 deaths is because of the cancer. Approximately 70% of deaths from cancer occur in low- and middle-income countries [1]. In Canada approximately 1 in 4 will die because of cancer and 1 from 2 will be suffering or develop cancer in their life, so cancer is the leading cause of death in Canada.

In Canada, an estimated new cancer cases and cancer deaths are 2,25,800 and 83,300 respectively expected in 2020 [2]. The most diagnosed cancers are expected to be lung cancer (29,800), breast cancer in females (27,400) and prostate cancer in males (23,300). The expected leading cause of cancer deaths is lung cancer, accounting for all cancer deaths 25.5% by the lung cancer, followed by colorectal, pancreatic and breast cancers are respectively 11.6%, 6.4% and 6.1% [2]. In 2019, an estimated 2,20,400 Canadians were diagnosed with cancer and 82,100 died from cancer. Lung, colorectal, breast and prostate cancers are expected to remain the most diagnosed cancers and accounted for 48% of all diagnoses in 2019 [3]. Table 1.1 shows the summary statistics of the cancer for 2020.

Table 1.1. Statistic of cancer disease in Canada for 2020 [4]

Summary statistic 2020			
	Males	Females	Both sexes
Population	18,732,178	19,009,979	37,742,157
Number of new cancer cases	145,006	129,358	274,364
Age-standardized incidence rate (World)	373.7	327.6	348.0
Risk of developing cancer before the age of 75 years (%)	35.3	30.4	32.8
Number of cancer deaths	45,721	40,963	86,684
Age-standardized mortality rate (World)	104.9	84.5	93.5
Risk of dying cancer before the age of 75 years (%)	10.4	8.8	9.6
5-year prevalent cases	527,888	495,373	1,023,261
Top 5 most frequent cancers excluding non-melanoma skin cancer (ranked by cases)	Prostate Colorectum Lung Bladder Kidney	Breast Lung Colorectum Corpus uteri Thyroid	Prostate Breast Lung Bladder Colorectum

Statistics in Table 1.1 were taken from Globocan 2020, which show that from the total population 37,742,157 around 2,74,364 Canadians suffered from cancer in 2020 from which 86,684 died. Analysis shows that male patients mainly suffer from prostate and colorectal cancer whereas female patients generally have breast and lung cancers. Risk factor of dying associated with cancer before the age of 75 is 9.6%. Based on Canadian cancer society report of estimation of 2020, the most diagnosed types of cancer in Canada are lung, breast, colorectal and prostate cancer (excluding non-melanoma skin cancer) [5]. In Canada almost 20% men have prostate cancer, as per the estimation of new cases for lung cancer and breast cancer, they are 14% and 25% respectively [6].

1.1 Stages of cancer

Once a person is diagnosed with cancer, a doctor or an oncologist diagnose the stage of the cancer. Cancer stage is determined based on the size of the cancer tumor, how far it has spread and whether it has affected other parts of the body. Generally, cancer is labeled in stages I to IV, where IV is the last and serious stage of cancer [7]. Determining the stage of cancer is very important because on that bases doctor plans the treatment which may include surgery chemotherapy or radiation therapy and also predicts the chances of recovery. Clinical tests for cancer include physical exam, blood test, Magnetic Resonance Imaging (MRI), Computerized Tomography (CT) scan, ultrasound, and many other imaging scans. Doctors also suggest having a biopsy, where a small piece of tissue is observed under a microscope.

- Cancer stage 0 means there is no disease, only abnormal cells with the possibility to become cancer. This is also called carcinoma in situ [7].
- Cancer stage 1 means the malignant growth is small and only in single region. Tumor is contained inside the organ it started in. This is also called biggening-stage cancer [7].
- Cancer stage 2 means the hazard risk is greater than stage 1, but it has not started to spread into the surrounding tissues [7].
- Cancer stage 3 means the cancer is larger and has grown into nearby tissues or lymph nodes [7].
- Cancer stage 4 means the malignancy has spread to different parts of the body. For example, different organs and bones. It is also called advanced or metastatic cancer [7].

Cancer stage does not change unless and until it comes back after treatment. If cancer comes back the doctors sometimes again diagnose with some tests (such as physical exams, imaging tests, endoscopy exams, biopsies, and maybe surgery) and follow some process to find out the amount or spread of cancer and stage again after initial stage that is called ‘cancer restaging’. Sometimes doctors conduct restaging to find out how cancer is responding to the treatment.

1.2 Cancer survival analysis

Survival analysis also known as time-to-event analysis, refers to a set of methods used to analyze the duration of time till specific event or end point occurred. Survival analysis is a subfield of statistics which is also used for modeling or structuring data is also analyzed for the information where results come in time. In context of statistics, censoring is a stage where some value of a measurement or observation is only partially available but in context of survival analysis, censoring occurs when there is some information about the subject or the patient, but we cannot assume exact event time for the patient. This is called as censored when information on time to event is not present due to failure in follow-up and absence of outcome event before the trial ends [8].

Censoring is one of the features of survival analysis data which not every cancer patient faces by the end of cancer survival time. Censoring is a form of missing value or partially known data. Censoring plays a role in the models same as in the non-parametric hazard and the condition of independent censoring. If the cancer patient failed to follow-up during the survival time or patient drops cancer observation then the patients time to live or time to survival cannot be observed or identified, that is called as censoring. When censoring occurs, due to the death of a patient then it is called non-random or informative censoring [9].

Currently, the main goal of health medicine area is to improve cancer prognosis. Many health professionals are working in the area of cancer survival analysis. Two conventional statistical methods are used by many researchers for cancer survival analysis, parametric and is semi-parametric where parametric method is used when distribution of survival time is known, and semi-parametric method is used when distribution is unknown. There are many ways or methods to estimating the absolute survival at a given time by varying the registration and follow-up periods of time. Table 1.2 shows the methods for estimating the absolute survival at a given time [10].

Table 1.2. Approaches to estimating the absolute survival

Approaches/methods	**Description**
Cohort analysis	The simplest way of computing survival probability is to compute the ratio or percentage of the number of subjects alive at the end of from the index date by the total number of subjects in the study at the beginning of the study.
Semi or partially complete analysis cancer	This approach is widely practiced in the estimation of survival by registries. In this not all patients diagnosed until the closing date of follow-up are included.
Complete analysis	In this approach to be used when there is no restriction on the potential follow-up time to equal. Rather, all subjects who are diagnosed as incident cancers until the closing date of the follow-up period qualify for inclusion in the analysis.
Period analysis	In this approach to deriving more up-to-date estimates of cancer patient survival by exclusively utilizing the survival information pertaining to the most recent incidence and follow-up periods.

The Cox proportional hazard regression model is the most widely used semi-parametric survival model in the medical science because it relies on fewer assumptions compared to parametric models [11]. The fundamental use of this model is the proportionality of the hazard function and to determine the risk factors. The hazard ratio of two people is independent of time presumed by proportional hazards (PH) models. It is inappropriate to use standard Cox PH model where PH assumption is not met as it may entail serious bias and loss of power when estimating or making inference about the effect of a given prognostic factor on mortality [11].

1.3 Semi-supervised learning

Machine learning algorithms are classified into three categories, first is supervised learning, second is unsupervised learning and third one is reinforcement learning. For training data for prediction in supervised learning we need labeled data whereas unsupervised algorithms do not need labeled data for training the machine for prediction as it can learn from unlabeled data. Unsupervised learning generally used to find new patterns in a dataset and to cluster the data into several categories based on several features [12]. Popular examples of algorithms are K-Means and Latent Dirichlet Allocation (LDA). In Reinforcement learning machine is trained to take sequence of decisions.

Semi-supervised learning is also called inductive learning or transductive learning. Semi-supervised learning is a learning paradigm concerned with the study of how computers and natural systems such as humans learn in the presence of both labeled and unlabeled data. In machine learning, a small amount of labeled data combines with a large amount of unlabeled data during training, is one approach of semi-supervised learning. The algorithm is trained upon a combination of labeled and unlabeled data. Typically, we can use semi-supervised learning when the labeled data are scarce, expensive, or small size. In semi-supervised learning, unlabeled data is used to improve supervised learning tasks. The basic procedure involved is that first, the programmer will cluster similar data using an unsupervised learning algorithm and then use the existing labeled data to label the rest of the unlabeled data. The typical use cases of such type of algorithm have a common property among them – The acquisition of unlabeled data is relatively cheap while labeling the said data is very expensive. Semi-supervised learning method uses less amount of labeled training data to train the model as compared to the supervised learning model. Also supervised learning uses labeled data and it can combine many neural network models and training methods [13]. Figure 1.1 shows the labeling of the data by semi-supervised learning [14].

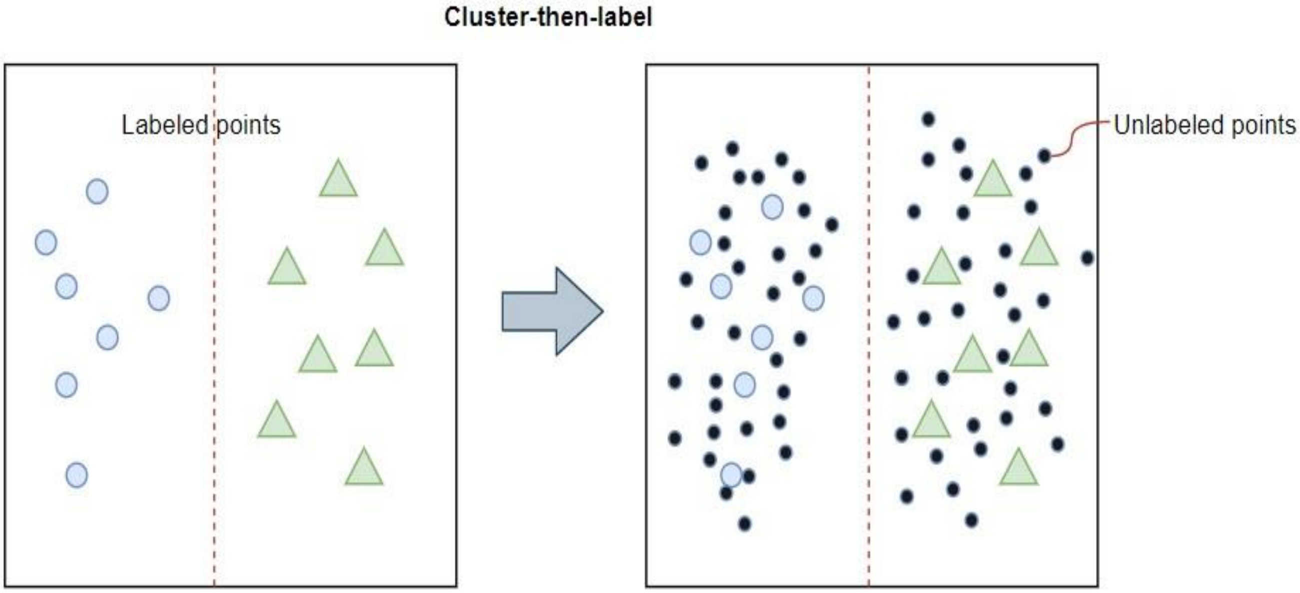

Figure 1.1. Labeling the data by semi-supervised method [14]

Semi-supervised learning classifiers are typically based on two main assumptions on data and joint data/ labels distributions:

1) Manifold assumption: it assumes that data classes span low dimensional manifold in the feature space.
2) Cluster assumption: it assumes that decision boundaries lie on low density regions

in data distribution. Furthermore, it is declared as 'labeling being smooth on data'. Consequently, similar data must have similar labels.

In case of one or both these assumptions apply, we can expect that the use of unlabeled samples may be beneficial for the overall performance with respect to a supervised learner. However, in general those assumptions may not be verified, and even if they were, the use of improper semi-supervised learner model may cause unlabeled data not leading to any improvement or even to performance worsening.

1.4 Motivation

Currently the main objective of some medical studies is to identify prognostic factors of patients' survival time based on clinical classification [15]. Our motivation for doing research in area of cancer survival analysis is that cancer is second leading cause of death in the world. A group of diseases that cause abnormal cells divide without control and overpass other tissues, called cancer. In addition, if they expand without any control, patient can die of cancer.

The reason behind measuring survival time of cancer patient is to estimate the time and days a patient or a group to patients to experience an event of interest. This time estimate is the duration between cancer identification and death event, so it is called as a 'time to event analysis'. For providing direct care to patients, and patients understandably wish to be given some estimate prognosis to clinician's survival statistics are become more popular. However, most survival data are just an evaluation of projected survival at the time of diagnosis. In reality, a patient's relative probability of survival for another 5 years changes—and generally improves—around the cancer experience. Recently published Canadian conditional survival data for a wide range of cancer diagnoses show that the prognosis did, in fact, except chronic lymphocytic leukemia all other cancers improve over time from cancer diagnosis [16].

Looking at the colorectal cancer data of Canada, the data show that except for esophageal cancer and ovarian cancer Canada is among the world leaders in survival for most of the seven cancers observed. While Canada’s overall average survival generally compares well, there is often more variation among the provinces than across the countries in this study [16].

The 2019-2029 Canadian Strategy for Cancer Control and its action plans acknowledge this variation in survival rates across Canada and strive to promote equity of access to cancer care and quality cancer care for all Canadians [16]. The above analysis was the main motivation to conduct research on cancer survival analysis with new semi-supervised learning model by combining the Cox and SP-AFT models using cancer data of high dimension and limited sample size.

1.5 Objectives

Following are some reasons and objectives for doing clinical cancer research in area of cancer survival analysis as well as using semi-supervised learning method with less labeled data of cancer patient.

1) Accurate prediction of survival time

Perfect predication of survival time is most important to determine in case of cancer patient. It is indices for mortality or recurrence of a disease, and to study the outcome of treatment. It is used to evaluate the impact of medicines or medical treatment on time until death. In this research biological analyses are performed or the selected genes using the semi-supervised method based on Cox and AFT models with L½ regularization to achieve accurate prediction and risk analysis.

2) Small size of labeled data

Semi-supervised learning methods were used in this research, which combines supervised and unsupervised learning methods using both limited amount of labeled data with a huge amount of unlabeled data during training. In this research small size labeled GENE data of cancer patient is used to train the model to generate large number of sample data based on learning algorithm.

3) To improve predictive performance

The most valuable part of this research is predictive modeling for patient survival time. For classification of patients in high-risk and low-risk and survival time prediction for the patients' medical cure, the Cox proportional hazards model (Cox) and accelerated failure time model (AFT) were used. The prediction performance can be improved by keeping the censored data in the same risk classes.

4) Risk Analysis

Cancer risk analysis can be split into two major types first is evolution of familial or genetic risk and assessment of environmental factors that may be causally related to cancer. A Hereditary Cancer Risk Assessment can help us and the doctors to take the significant decisions about the health of a patient. The risk analysis was done by classifying "low-risk" or "high-risk" using Cox model which depends on traditional supervised learning techniques.

1.6 Contributions

In this study, we have implemented the semi-supervised learning analysis proposed by Liang et al. (2016) [34] on simulated survival data and on two real datasets. The study includes the following steps:

- Cox proportional hazard model is explained and has been used for analyzing survival data.
- Accelerated failure time (AFT) model with lognormal distribution has been used to model survival data.
- Kaplan-Meier estimator has been explained and used for estimating the empirical distribution of the survival function.
- Mean imputation method has been used for imputing the censored survival times by using the survival function calculated by Kaplan-Meier estimator.
- Semi-supervised learning method has been explained and used to model survival data.
- Four methods of single cox, single AFT, semi-cox and semi-AFT have been used and compared by simulating survival data using different correlation values.
- Four methods of single cox, single AFT, semi-cox and semi-aft have been used for modeling DLBCL2002 data which also was modelled by Liang et al. (2016) [34] and the results were compared with each other.
- The four methods have been used on new real data (Colorectal cancer data). The results are presented and compared for the four methods in term of number of selected parameters, fraction of censoring before and after implementing semi-supervised learning method, Concordance index (CI) and integrated brier score (IBS).

Chapter 2

Literature Review

2.1 Related Work

Selection of model in Cancer Survival Analysis is one of the most important tasks. Many researchers are working in this area to identify a sub-type of cancer to increase patient survival rate. In Bioinformatics and Biomedical field semi-supervised learning method is used to analyze the gene expression data and map in training Dataset. This covers the work done previously in this area of Cancer Survival Analysis using Semi-Supervised learning method and cancer analysis using COX and AFT models.

Nattawut Thongpim et al. (2020); In this research they do analysis and predict survival of prostate cancer patient which is used by doctor to improve medical decision making in medical treatment process [17]. They used patient features age, family history, prostate specific antigen (PSA), Gleason group, Gleason score and pathology report for predicting. They have collected 78 prostate cancer patient data aged between 40-89 years old. Exploratory Data Analysis (EDA) is used to summarize the details of data availability and understanding of data. They used Prostate-Specific Antigen which is a type of protein produced by the cells of the prostate gland which indicates prostate disorder for predicting survival time using Cox regression. The conclusion of this research is that metastasis factors, PSA factors, Gleason score factors and age factors are the most effective factors in prediction.

Chai et al. (2017); In their study they combined a semi-supervised learning framework with Cox proportional hazard (COX) and accelerated failure time (AFT) for cancer research [18]. They also used self-paced learning method for effectively employing the information in the training data which helps AFT model to identify and include samples automatically into training and also helps to minimize interference of high noise. They used Cox model for classification and self-paced accelerated failure time (SP-AFT) model for prediction. They determine from their proposed system that COX-SP-AFT model can utilize more samples and estimate their survival time with more accuracy.

Wang et al. (2018); work in self-training of cancer sample data with assembling with semi-supervised learning for cancer classification [19]. They used semi-supervised learning as a self-training by utilizing unlabeled samples of biological data to improve the model performance. In their study they proposed self-training learning (ESTL) method for selecting unlabeled samples effectively with high-quality to reduce the noise. They used 102 samples among which 52 prostate cancer samples, trained highly accurate data using self-training learning and compared the results with traditional classification algorithms. The proposed classification approach improved 5% performance as compared to other traditional classifiers such as logistics regression and support vector machine (SVM) by using ROC (receiver operating characteristic) method.

Qiu et al. (2019); automated annotation of pathology reports and they investigated semi-supervised deep learning system to improve performance of information extraction system [20]. In this research they used 3,74,000 pathology reports from the Louisiana Tumor Registry and a novel conventional attention-based auto- encoder. From each report they extract information of six key features cancer primary site, sub-site, liberality, histology, behavior and grade and from total samples they used 80% sample reports for training information extraction system and rest of 20% is used for validation. After data preprocessing, they used a conventional attention network encoder with weight factor for feature mapping. They feed encoded document into task-specific classification layer simultaneously for join prediction in context of multitask supervised learning.

They conclude by indicating that the performance gains from the auto- encoders improved representation learning by utilizing unlabeled data.

Kabir et al. (2019); have proposed classification model and survival analysis of prostate cancer patients by doing RNA sequencing of clinical data [21]. For detection of cancer and non- cancer they used machine learning and data mining approaches. Using cancer-sensitive genes they build regression tree with clinical attribute Gleason score as predictor and overall survival variable as the target variable. Decision tree is used for classification between cancer and non- cancer patients. Their assessment is to find gene-gene interactions and gene-environment interactions of prostate cancer. They used 550 patient instances among which 4978 were cancer patient samples and other are non-cancer patients from RNA-seq and clinical variable available in national cancer institute Genomic data commons (GDC). For research they used 36 parameters of patient instances which are associated with cancer from which Gleason score were used as a predictors and sample type were used in regression. For classification they used decision tree, random forest and multi-layered neural network whereas for investigation they used decision tree regression. They concluded the research by comparing classification and performance of all methods and found that multi-layer perceptron (MLP) performs betters than all other mentioned methods.

Huang et al. (2019); proposed highly accurate prediction model, a novel cox proportional hazards model for high dimensional genomics data in cancer prognosis [22]. In this study they proposed new strategy for using cox model for cancer survival duration prediction and gene selection. They combined self-paced learning (SPL) and smoothly clipped absolute deviation (SCAD) network-based regularization from which SPL allows learning from the low-level noise samples first to high-level samples by enhancing knowledge structure and SCAD network is used to make flatness between the coefficients of neighboring genes in biological network. They have taken 1000 patient instances from which 921 patients have invasive breast carcinoma After data collection the proposed model SPL with SCAD-Net penalty is applied to train the cox model and perform simulation analysis which expresses that SPL with the SCAD-Net penalty to the Cox model (SSNC) method is better than the other competitive methods in context of prediction and gene selection.

Nezhad et al (2019); proposed a solution for survival analysis framework for prostate cancer patients using deep learning and active learning, named as Deep Active Survival Analysis (DASA) [23]. In their study first they train survival model by labeling the examined data which are clinical features of patient instance. The feature representation was performed using deep learning to produce robust features from high- dimensional, sparse and complex Electronic Health Records (EHR). For that they set all instances in a pool of time-to-event instances and apply deep feature learning on both trained pools set. For experiments in this research, they used dataset from Surveillance, Epidemiology and end results (SEER) program. The proposed solution gives more accurate survival analysis for risk prediction, survival time estimation and treatment recommendation. The only limitation with this method is quality of result is bounded where labeled instances are limited, and data is high dimensional.

Katzman et al. (2018); In their study they proposed DeepSurv model which is a cox proportional Hazards deep multi-layer perceptron for prediction of risk occurrence for patients during survival and also give recommendation for personalized treatment [24]. Mainly Deepsurv model is used for prediction of patient risk and for recommendation which was tested on real medical studies and provide treatment recommendations. The research concluded that DeepSurv can compute both complex and nonlinear features without any prior selection and domain knowledge and compared to random survival forest method. DeepSurv method gives accurate results and better performance.

Hirozawa et al. (2018); In this research they proposed new survival prediction system based on extracted prognostic factors using machine learning and random forest regression [25]. For this study they used 2363 patient records and considered 55 prognostic factors of patients from which main factors are patient background, symptom, physical status, physical exam, medical activity, activity of daily living and blood exam. First, they applied random forest regression to rank and extract the prognostic factors. Then they applied regression trees to discretize the factors. At the end they replaced the two factors of Dyspnea at rest and Delirium with extracted factors. They compared proposed system with Palliative Prognostic Index (PPI) and concluded that proposed system of predicting survival is more accurate.

Yang et al. (2019); they introduced DeepCoxPH model which is a risk score estimation strategy based on deep learning and CoxPH [26]. The model is used to improve the risk stratification where abstracted weight from Deep learning (DL) and the hazard ratios from the CoxPH were transformed into risk score estimation for complete survival analysis. The model is divided into three parts, first deep learning (DL) was used for training network and toobtain weights, second coxPH was used to calculate hazard ratio (HR) and third by combining HR and DL weight risk degrees can be combined through matrix mortification or matrix addition to find high and low risk stratification. They have taken 1646 breast cancer patient records as a dataset. DeepCoxPH model used hazard rate to determine the effect of individual clinic pathological variables. This research determined that, combination of both risk weights in the CoxPH and DL models together can possible by the DeepCoxPH method and give more accurate risk stratification results.

Barsainya et al. (2018); the research carries the prediction survival duration after colorectal chemotherapy [27]. Looking for dataset side and parameters side they have used 14 factors from each patient records which are id, study, prior treatment, age, gender, obstruction, adherence factor, perforation factor, number of lymph nodes attacked, differentiation of tumor, censoring period, time period, extent of local spread, event type. They have tested training model on decision trees, KNN, regression, Bayesian classifier, neural networks for prediction task. They split total dataset into 80-20 ratio for training and testing purpose. They conclude that this model is specifically on post cancer results and statistics to assist patients in knowing if they need more treatment.

Cai et al. (2018); In this research they proposed the model which is used for prognosis prediction of early-stage lung cancer using support vector machine [28]. For this study they have used 174 data of early-stage lung cancer patients and 11 features from each patient instance for analysis. Borderline synthetic minority over-sampling technique algorithm is used for increasing sample based on SMOTE. They compare prognosis model labeling based on SVM and COX by dividing into 2 parts, first sample set contains data without preprocessing and other set contains data with preprocessing. They concluded that by using Borderline-SMOTE algorithm with SVM make insignificant allowance to the performance of this model and also this model provides effective and reliable prognosis for early-stage lung cancer.

Liang et al. (2016); In this research they implemented semi-supervised learning method using Cox and AFT model with L1/2 regularization for Lung, Diffuse Large B-cell Lymphoma (2002), DLBCL 2003 and AML cancer patients [34]. Model was developed with the limited patient's clinical parameters and high-dimensional microarray gene expression profiles including censored data. For this research they have used 240 number of patients and 7399 number of genes from National cancer Institute, center of cancer research which has 12 attributes like, patient's accession id, status of patient, subgroup, follow-up time, various signature types. For imputation of censored data, they have tested different approaches like Buckley James approach, Rank-based and mean imputation approach. In this research they have concluded that the semi-supervised learning approach with COX and AFT model is powerful tool to predict the survival time of the patients and gives accurate classification of patients. Also, L1/2 regularization approach helps to get proper biomarker selection.

Imani et al. (2019); In this paper they proposed survival analysis of cancer recurrence using random forest model for breast cancer patients [29]. Model was constructed using sampling and bootstrapping into big data and taking surveillance, epidemiology, and end result data of year 1973 to 2015 as input. For this research they have used 1,631,572 patient entries from SEER database and from each patient instance 14 attributes were taken for analysis which are Id, age, year, month, sex, race, marital status, grade stage, surgery, primary site, histology, sequence number, laterality, behavior and diagnostic. The patterns of event occurrence to reveal is key properties of breast cancer so the main outcome they focus is time-to-event. The result is based on age, surgery history, stage and histological grade four important attributes which impact the most in recurrence of breast cancer. The proposed method estimates and predicts the survival function by sampling and bootstrapping. The result of this methods shows that recurrences of breast cancer lies between 2% to 6%. Milad Zafar Nezhad et al (2019) [23] is also used same dataset of SEER Medicare data in their research active learning-based survival analysis using a novel sampling strategy where as Farhad [29] uses random survival forest to study the impact of each patient attribute on recurrence of cancer. They both apply deep learning for feature reduction and extraction, when data is high-dimensional, complex, and sparse. Table 2.1 shows a summary of the literature review.

Table 2.1. Summery of Literature Review Analysis

Ref No	Title	Dataset	Methodology
18	A new semi-supervised learning model combined with cox and sp-aft models in cancer survival analysis	78	Cox regression
17	On Predicting Survival Opportunities for Prostate Cancer by COX Regression in PSU Patients Data	2000	Cox proportional hazard accelerated failure time
19	Semi-supervised learning with ensemble self-training for cancer classification	102	Self-training learning (ESTL) method
20	Semi-Supervised Information Extraction for Cancer Pathology Reports	3,74,000	Semi-supervised deep learning
21	Classification Models and Survival Analysis for Prostate Cancer Using RNA Sequencing and Clinical Data	550	Multi-layer perceptron, decision tree
22	A novel Cox proportional hazards model for high- dimensional genomic data in cancer prognosis	1000	Cox model
23	A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer	Not mention	Deep Active Survival Analysis (DASA)
24	DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network	Not mention	Cox proportional Hazards deep multi- layer perceptron
25	New survival prediction system for terminal patients based on machine learning	2363	Machine learning, random forest regression
26	Identifying Risk Stratification Associated with a Cancer for Overall Survival by Deep Learning-Based CoxPH	1646	Risk score estimation strategy based on deep learning and CoxPH
27	Analysis and Prediction of Survival after Colorectal Chemotherapy using Machine Learning Models	Not mention	Decision Tree Classifier

28	The earlystage lung cancer prognosis prediction model based on support vector machine	174	Support vector machine
29	Random Forest Modeling for Survival Analysis of Cancer Recurrences	1,631,572	Random forest model
34	Semi-supervised learning method using COX and AFT model with L1/2 regularization	4	COX and AFT model with and without semi-supervised learing

In this thesis, semi supervised learning has been used with COX and AFT models for predicting risk and survival time for the cancer patients. Logistic regression or logistic model has been used in epidemiological studies to predict a discrete outcome (say, yes/no) from a set of independent variables whereas cox proportional hazard model is a semi-parametric regression method, which used for estimating the hazard function to analyzing the risk factors for multivariate prognostic factors. The cox model is used when the result is in numerical value such as the “time to live” and logistic regression is used for a binary event such as “dead” or “alive”. Another difference is that in Logistic Regression Odds is being modeled which is in unit-less whereas in Cox Regression Hazard rate is being modeled which is in time^{-1} units. Cox model is semi parametric, so it was used as semi supervised learning with labeled and unlabeled data both and it evaluates hazard ratio. Semi-supervised learning takes both small amount of labeled data as well as a larger set of unlabeled data where Reinforcement learning is a training algorithm with a reward system, providing feedback when an artificial intelligence agent performs the best action in a particular situation and is also used as a tool for training AI models. Reinforcement learning uses the estimated errors as rewards and is about taking suitable action to maximize reward in a particular situation. Semi-supervised learning is a combination of the supervised and unsupervised learning families. This group of models consists of algorithms that use the assessed errors as rewards or sentences. When separating significant features from the data is difficult and labeling examples is a time-intensive task for experts, semi-supervised learning method is useful. So, for predicting risk and survival time for the cancer patients, semi-supervised learning method was used with COX and AFT model in this research.

Chapter 3

Materials and Methodology

3.1 Dataset properties

In this study two real datasets were used for the survival analysis. The first dataset was collected by Rosenwald et al, (2002) [30] and the Colorectal cancer dataset (Smith et al, 2010) has patients' gene expression profiles and artificial evaluation are available on NCBI (Gene expression Omnibus) [31]. In this research, the Diffuse large B-cell lymphoma (DLBCL 2002) dataset was used for the implementation of four methodologies to verify the previous research which is done by Liang et al. 2016 [34]. If the latest dataset is used the results would be different based on the survival times because the survival time will be different with the different patients. Another dataset of colorectal cancer was used which was the latest available dataset on NCBI and the results may vary for new datasets because the survival time will be different with the different patients.

3.1.1 Diffuse large B-cell lymphoma (DLBCL2002) Data

The dataset is examination samples of 240 patients' gene expression with analyzes of DNA microarrays. The dataset was taken after chemotherapy from specimens of tumor (diffuse large B-cell lymphoma) biopsy. In this study it will be shown as DLBCL (2002) [30]. It includes 7399 gene expression variables. There were 240 patients in this data where 102 of them (42.5%) were alive at the time of data collection and 138 of them were dead (57.5%). It means the survival time of 102 (42.5%) of the patients is right-censored and for the rest of the patients the survival time is known.

All the patients were diagnosed to have diffuse large B-cell lymphoma. The patients which had tumor biopsy specimens available were selected. They received chemotherapy and the monitoring was continued for up to 7 years. 57% of the patients were dead at the end of monitoring. 56% of the patients were male and 44% of them were female. The median age for the patients was 63 years. More information about the data collection can be found in Rosenwald et al, (2002) [30].

The heading of the collected gene expression for DLBCL 2002 dataset shown in Table 3.1. The current data for gene expression include 7399 genes for 294 patients. Each of the unique IDs in Table 3.1 are referred to as gene expression. There are 7399 uniqueIDs. The names of patients are shown as MCL94_46 and there are 294 patient names. while the survival data is recorded for 240 patients, matching between these two datasets is required before starting the analysis

Table 3.1. Head of Gene-expression of DLBCL2002 - 7399 gene expression x 240 cases

UNIQID	MLC94_46	MLC96_45	MLC91_27	MLC96_84	MLC96_43	MLC91_28
27481	2212	2893	3517	2890	1224	5736
17013	3436	2009	3216	2888	3833	4647
24751	1839	1938	2817	2050	2600	3865
27498	4664	2518	4446	3795	3521	6550
27486	3029	1842	4458	2364	3242	4601
30984	1	333	43	364	37	65
summary statistics:						
UNIQID	Min	1st Qu.	median	mean	3rd Qu.	Max
27481	720	2013	2740	2996	3630	8273
17013	541	1707	2425	2696	3428	8040
24751	210	1462	1946	2153	2749	7150

27498	483	1813	2690	3017	3815	12009
27486	352	1617	2226	2608	3209	13034
30984	1	151	310.5	354.4	474.8	1551

The samples survival time, status of the patient or censoring (binary variable represents whether the patient is alive = 0 (censored) or dead = 1), International Prognostic Index group or IPI-group and gene expression subgroups are the major columns of this data set. The head of dataset for survival is presented in Table 3.2.

Table 3.2. Survival times and events for DLBCL2002

DLBCL_sample	Analysis_Set	Time(years)	status	Subgroup	IPI_Group
2	Training	4	Alive	GCB	Low
4	Training	4.9	Alive	GCB	Medium
6	Training	5.6	Alive	GCB	Low
7	Training	12.1	Alive	GCB	Medium
8	Training	0.6	Dead	ABC	Medium
11	Training	0.3	Dead	GCB	High

In Table 3.2, survival time for those with status Alive are considered as censored data. The analysis set column includes 160 training and 80 validation cases. Out of 240 patients, 138 patients are dead (status = 1) and 102 of them are alive (status = 0). In the subgroup columns there are 73 cases as ABC, 115 cases as GCB and 52 of them are as Type III. The IPI-group includes 32 cases as high, 82 of them as low and 108 of them as medium, the rest of the cases (18) IPI-group are missing. The frequency table for these parameters is presented in the Table 3.3.

Table 3.3. Frequency and summary table for the survival data

Analysis_Set		Time (years)		status		Subgroup		IPI_Group	
Training Validation	160 80	min 1st Qu Median 3rd Qu Max	0.0 0.9 2.8 7.1 21.8	Alive Dead	102 138	ABC GCB Type III	73 115 52	High Low Medium Na	32 82 108 18

Two-way cross tabulation for the Subgroup versus IPI_Group is shown in Table 3.4. The chi-squared (χ^2) tests of independence do not show a significant dependence between IPI_Group and Subgroup. ($\chi^2 = 3.772, df = 4, p - value = 0.4377$)

Table 3.4. Cross tabulation of IPI_Group and Subgroup

SubGroup	IPI_Group		
	High	Low	Medium
ABC	13	20	35
GCB	12	43	53
Type III	7	19	20

Cross tabulation for IPI_Group versus survival time divided in binary categories are presented in Table 3.5. There is a significant dependence between survival time categories and the IPI_group for survival time > mean ($\chi^2 = 17.531, df = 2, p - value = 0.000156$), survival time > median ($\chi^2 = 33.11, df = 2, p - value = 6.461e - 08$) and for log(survival time) > 0 ($\chi^2 = 15.092, df = 2, p - value = 0.0005283$)

Table 3.5. Cross tabulation of IPI_Group versus survival time categories

Survival time	IPI_Group		
	High	Low	Medium
time < mean(time)	26	36	73
time >= mean(time)	6	46	35
χ^2= 17.531, df = 2, p-value = 0.000156			
time < median(time)	26	22	63
time >= median(time)	6	60	45
χ^2= 33.11, df = 2, p-value = 6.461e-08			
log(time) < 0	13	9	34
log(time) >= 0	19	73	74
χ^2= 15.092, df = 2, p-value = 0.0005283			

It can be seen from Table 3.5. that for low and medium IPI-Group the frequency count for log(time) >= 0 is much more than those of log(time) < 0 and high category of IPI-Group. The count of Low IPI_Group for time >= median(time) is 60 while for time < median(time) the maximum time (63) can be seen for the medium category of IPI_group.

The Chi-squared test of independence for the survival time categories versus the subgroup column of the data shows significant dependence. For the survival time < mean (survival time) the test has p-value slightly less than 5% and it is more than 1%. so, it is not significant at 1% significance level. But for categories of survival time < median (survival time) and log(time) < 0 there is a significant dependence. The GCB category has significantly more counts for log(time) >= 0 compared with GCB of log(time) < 0 and other categories. Figure 3.6 shows the cross-tabulation of the subgroup verses survival time categories.

Table 3.6. Cross tabulation of subgroup versus survival time categories

Survival time	Subgroup		
	ABC	GCB	Type III
time < mean(time)	51	60	36
time >= mean(time)	22	55	16
χ^2= 7.668, df = 2, p-value = 0.02162			
time < median(time)	47	42	30
time >= median(time)	26	73	22
χ^2= 15.613, df = 2, p-value = 0.0004071			
log(time) < 0	27	19	16
log(time) >= 0	46	96	36
χ^2= 10.605, df = 2, p-value = 0.00498			

The dimensions of first data which include gene expression was 294x7399 and the second data which include survival time has dimension of 240x12. There are in total 240 patients which could be matched through these two datasets. Column for matching in the first data is the heading which include names like: "MLC94-46_LYM009_de novo untreated", "MLC96-45_LYM186_de novo untreated" , "MLC91-27_LYM427_de novo untreated" and … . for all these names after character LYM the sample ID could be seen for example in the mentioned names the sample ids are 9, 186 and 427, respectively. In the first column of the survival data (DBCL_sample) these values could be found. For each patient in gene expression data and survival data these sample ids were matched, and data was merged together.

3.1.2 Colorectal Data

The colorectal data (Smith et al, 2010) [31] includes 177 patients from Moffitt Cancer Center. This data includes gene expression profiles derived from invasive mouse colon cancer cells. The data includes 54675 gene expressions for these 177 patients. The heading for the gene expression data is presented in Table 3.7.

Table 3.7. Heading of the gene expression data for colorectal cancer, 54675 x 177

ID_REF	GSM437093	GSM437094	GSM437095	GSM437096	GSM437097	GSM437098
1007	11.7	11.6	11.2	11.8	11.1	11.2
1053	8.85	8.63	8.23	8.73	8.52	8.81
117	7.45	7.34	8.12	7.29	7.82	7.51
121	9.91	9.96	9.89	9.59	9.55	9.68
1255	4.89	4.90	4.77	4.84	4.80	5.10
1294	8.45	7.98	7.77	8.39	8.16	7.95
summary statistics:						
ID_REF	Min	1st Qu.	median	mean	3rd Qu.	Max
1007	10.03	11.16	11.37	11.37	11.58	12.48
1053	7.598	8.536	8.765	8.764	9.036	9.811
117	7.095	7.447	7.591	7.663	7.765	9.451
121	9.495	9.731	9.838	9.844	9.950	10.364
1255	4.648	4.846	4.902	4.930	4.983	5.411
1294	7.444	7.911	8.111	8.139	8.364	9.093

The gene expression has a very low interquartile range for this data. The standard deviation of the genes is also low and ranges from 0.13 to 0.39 for these 6 columns. Figure 3.8 shows the heading of the survival data for colorectal cancer.

Table 3.8. Heading of the survival data for colorectal cancer, 177 x 13

Sample_Colorectal	Age	Gender	Ethnicity	Grade	Status	Survival time(month)
GSM437093	73	male	caucasian	Medium	Alive	143.0
GSM437094	63	male	caucasian	High	Alive	123.0
GSM437095	51	male	other	Medium	Dead	29.0
GSM437096	56	female	caucasian	High	Alive	119.0
GSM437097	50	male	caucasian	Medium	Alive	59.5
GSM437098	63	female	other	Medium	Dead	68.8

The survival time is set as months. The frequency count and the summary of the survival data is presented in Table 3.9.

Table 3.9. Frequency counts for the parameters in survival data

Age		Gender		Ethnicity		Grade		Status		Survival time (month)	
Min	26	male	96	black	9	High	16	Alive	104	Min	0.92
1st Qu.	57	female	81	caucasian	151	Low	27	Dead	73	1st Qu.	22.78
Median	66			hispanic	1	Medium	134			Median	42.27
Mean	65.48			other	16					Mean	48.12
3rd Qu.	75									3rd Qu.	67.82
Max	92									Max	142.55

The survival time median is 42.2 months-, 54% of the patients are male and 46% are female. Regarding the censoring, there are 104 (58.8%) patients as Alive which are said to be right censored. Only 41.2% of the patients are not censored. As it could be seen in Table 3.10 The cross tabulation for the parameters versus the survival time > median (survival time) is not significant in this data which holds the null hypothesis of independence. The Pearson correlation between Age and survival time is not significantly different from zero. which shows the correlation between Age and survival time is zero. (t = -1.75, df = 175, p-value = 0.08)

Table 3.10. Cross tabulation of Grade, Gender versus survival time categories

<table>
<tr><td rowspan="2">Survival time</td><td colspan="4">Grade</td></tr>
<tr><td>High</td><td colspan="2">Low</td><td>Medium</td></tr>
<tr><td>time < median(time)</td><td>7</td><td colspan="2">18</td><td>63</td></tr>
<tr><td>time >= median(time)</td><td>9</td><td colspan="2">9</td><td>71</td></tr>
<tr><td colspan="5">χ^2= 3.722, df = 2, p-value = 0.1555</td></tr>
<tr><td rowspan="2">Survival time</td><td colspan="4">Gender</td></tr>
<tr><td colspan="2">female</td><td colspan="2">male</td></tr>
<tr><td>time < median(time)</td><td colspan="2">40</td><td colspan="2">48</td></tr>
<tr><td>time >= median(time)</td><td colspan="2">41</td><td colspan="2">48</td></tr>
<tr><td colspan="5">χ^2= 0.006696, df = 1, p-value = 0.9348</td></tr>
<tr><td rowspan="2">Survival time</td><td colspan="4">Ajcc_stage</td></tr>
<tr><td>I</td><td>II</td><td>III</td><td>IV</td></tr>
<tr><td>time < median(time)</td><td>6</td><td>29</td><td>20</td><td>33</td></tr>
<tr><td>time >= median(time)</td><td>18</td><td>29</td><td>36</td><td>6</td></tr>
<tr><td colspan="5">χ^2= 29.259, df = 3, p-value = 1.976e-06</td></tr>
</table>

In Table 3.10, the only chi-squared test that has p-value < 0.01 is the Stage verse survival-, so classified stage and survival time are not independent from each other

3.2 Methodology

For survival analysis Cox proportional hazard model (Cox, 1972) [32], Accelerated Failure Time (AFT) (Wei,1992) [33] and semi-supervised learning methods including semi-Cox and semi-AFT proposed by Liang et al. 2016 [34] have been used in this study. These methods were used and compared with the simulated data and real data DBCL2002 [30] and Colorectal data [31].

3.2.1 Cox proportional hazard model

Cox proportional hazard model is a semiparametric regression model. The idea is that the log of hazard ratio is linearly dependent with the covariates. The hazard ratio is constant over time. using p covariates, the i^th^ individual hazard ratio is given by the formula (1).

$$log\,\{\frac{h_i(t)}{h_0(t\,)}\} \;=\; \beta_1\,X_{1i}\; +\beta_2\,X_{2i}+\ldots+\beta_p\,X_{pi} \tag{1}$$

where $\beta_{\mathbf{1}}\,,\,\ldots\,,\,\beta_{\boldsymbol{p}}$ are the coefficients for p covariates and $\boldsymbol{X_{1i},\ldots,X_{Pi}}$ are the covariates for ith individual. $\boldsymbol{h_0(t)}$ is the baseline hazard rate at time t and $\boldsymbol{h_i(t)}$ is the hazard rate of the ith individual at time t. The ratio of the hazard rate to baseline hazard is constant over time. The matrix notation for the Cox model is given by the formula (2).

$$h(t\mid X)\;=h_0(t)\;e^{(\beta\prime\,X)} \tag{2}$$

The hazard ratio is related to the covariate and only time dependent parameter is baseline hazard function. Cox (1972) [32] has used partial likelihood which is helpful in parameter estimation without requirement for calculating the baseline hazard. Considering $T_i$as the survival time of the ith individual, δ_i as the censoring indicator and X_i as a px1 covariates vector with β' as a transpose of px1 coefficient vector, the partial likelihood of the i$^{\text{th}}$ individual can be written by formula (3).

$$l_i(\beta) = \frac{h(T_i \mid X_i)}{\sum_{l \in R(T_i)} (T_i|X_l)} = \frac{\boldsymbol{h_0(t)}\, \boldsymbol{e^{(\beta' X_i)}}}{\sum_{l \in R(T_i)} \boldsymbol{h_0(t)}\, \boldsymbol{e^{(\beta' X_l)}}} = \frac{\boldsymbol{e^{(\beta' X_i)}}}{\sum_{l \in R(T_i)} \boldsymbol{e^{(\beta' X_l)}}} \tag{3}$$

The partial likelihood could be estimated without baseline hazard rate. The $R(T_i)$ are risk sets. All the individuals which are at risk at each time T_i will be summed together in the denominator of the partial likelihood. To derive the individuals which are at risk at time $t = T_i$. It is required to order all the individuals with their survival time in ascending order (from i = 1 to n). After ordering them by the survival time, then at $t = T_1$, all the n observations are at risk. The summation of $\boldsymbol{e^{(\beta' X_l)}}$ ($\boldsymbol{l}$=1 to n) for all individuals will be in denominator of the $l_1(\beta)$. Tied events are the observations with the exact same survival time, so that they get the same rank after sorting them. Neglecting the tied observations, for $t = T_2$ only the observation T1 is passed away and it is no longer at risk. So, all the observations except the first one will be included in the risk set. It continues like that until at observation $t = T_n$ only the nth observation is at risk and it will be considered in the risk set. The partial likelihood for all n observations could be written by formula (4).

$$l\,(\beta) = \prod_{i=1}^{n} \{\frac{\boldsymbol{e^{(\beta' X_i)}}}{\sum_{l \in R(T_i)} \boldsymbol{e^{(\beta' X_l)}}}\}^{\delta_i} \tag{4}$$

It is the multiplication of all $l_i(\beta)$ from (i = 1 to n | $\delta_i = 1$). If i$^{\text{th}}$ observation is not censored then $\delta_i = 1$otherwise $\delta_i = 0$. If $\delta_i = 0$ then $l_i(\beta)$ = 1.

For simplicity the log-partial likelihood is used in the formula for the estimation of the parameters. The equation for log-partial likelihood will be turned from the multiplication to summation since the natural logarithm of multiplication is equivalent with the summation of all logarithms. Log {a * b} = log {a} + log{b} and for the division Log {a / b} = log {a} - log{b}. Hence after taking the logarithm from formula (4) it will be as (5) which is the log-partial likelihood.

$$LL\,(\beta) \;=\; \sum_{i=1}^{n} \;\delta_i[\beta' X_i \;- log\{\sum_{l\,\epsilon\, R(T_i)} \;\; \boldsymbol{e}^{(\boldsymbol{\beta}'\,\boldsymbol{X_l})}\}] \qquad (5)$$

To find the maximum log-partial likelihood, derivative of the (5) is taken. Setting the equation (6) equal with zero the parameters could be estimated by finding the roots of the equation (6).

$$\partial LL\,(\beta)/\partial\beta = \sum_{i=1}^{n} \;\delta_i[X_i \;- \frac{\sum_{l\,\epsilon\, R(T_i)} \;\boldsymbol{X_l e^{(\beta' X_l\,)}}}{\sum_{l\,\epsilon\, R(T_i)} \;\boldsymbol{e^{(\beta' X_l\,)}}}] \;=\; 0 \qquad (6)$$

Parameter estimation could be done by using Newton-Raphson algorithm for maximum likelihood estimation. However, González-González et al (2008) [35], have mentioned that Newton-Raphson algorithm will not converge in the solution of maximum partial likelihood estimation of cox model when there is collinearity (collinearity is when some of the covariates are highly correlated with each other) in the design parameter matrix. For high dimensional data like the data in this study which has a number of parameters much more than number of observations (p >> n), dimension reduction is required. To do parameter selection in such cases one way is to use Lasso regression (Tibshirani, 1996) [36]. Lasso regression uses the L_1 norm penalty term. Xu et al, (2010) [37] claim that there is no obvious difference for selecting regularization penalty (0 < q < 1), but $L_{1/2}$ penalty is more efficient compared with L_0. Tibshirani, (1997) [39] has used lasso regression for cox model by minimization of the log partial-likelihood with subjecting the coefficients to specific constraints. $\sum \;|\beta_j| \;\leq s$. Hence using the penalized model, the coefficients could be estimated by the minimization of the partial log-likelihood. Estimation of the parameters in cox model with $L_{1/2}$penalty could be done by minimizing formula (7):

$$\beta \;=\; argmin\{\,LL(\beta) \;+\; \lambda \sum_{j\,=\,1}^{p} \;|\beta_j|^{1/2}\} \qquad (7)$$

where $LL(\beta)$ is the log partial likelihood. λ is the tuning parameter and $L_{1/2}$ norm is used for penalty term ($penalty(\beta) \;=\; \sum_{j\,=\,1}^{p} \;|\beta_j|^{1/2}$)

Regularization parameter of $L_{1/2}$penalty with new half thresholding presented by Liang, *et al.* (2016) [34] will be used in this study. coordinate descent algorithm will be used. Gui and Li (2005) [38] have shown that with choosing appropriate tuning parameters and implementing cross validation the thresholding operator used in the $L_{1/2}$ regularization will be converged.

3.2.2 Accelerated Failure Time (AFT)

In the Accelerated Failure Time is a linear model for survival analysis. In this model the logarithm of the response variable (time) is linearly related with the covariates X_i. The formula for the AFT model is given in formula (8).

$$h(t_i) \; = \; \beta_0 \; + \; \beta' \, X_i + \varepsilon_i \qquad (8)$$

here $h(t_i)$ is not hazard function and instead it is the logarithm of the time (for log-normal AFT model). h here is transform function which transfers the time to log time. Since the dataset include censored observations, the ordinary least squares method is not applicable for the parameter estimation. To solve this issue the outcome (survival time) for the censored observation can be replaced (imputed) by an appropriate method. Liang, *et al.* (2016) [34] have used a mean imputation method to impute the censored data by appropriate survival times. The mean imputation method employs the Kaplan-Meier estimator to estimate the survival function (Kaplan and Meier, 1958) [40]. The idea is that for each observation i if the observation is not censored $h(t_i)$ will not be imputed and if it was censored it will be imputed by the summation of $h(t_r) \, * \, \Delta\hat{S}(t_r)$ for all $t_r > t_i$ divided by $\hat{S}(t_i)$ where $\Delta\widehat{S}(t_r)$ is one step of difference in the survival function of Kaplan-Meier estimator which is $\Delta\widehat{S}(t_r) \; = \widehat{S}(t_r)$-$\hat{S}(t_{r-1})$ and $\hat{S}(t_i)$ is the Kaplan-Meier estimate for survival function at time t_i. The mean imputation could be seen in formula (9):

$$h(t^*{}_i) = \delta_i h(t_i) + (1 - \delta_i)\{\hat{S}_i\}^{-1} \textstyle\sum_{t_{(r)} > t_i} h(t_{(r)}) \Delta\hat{S}(t_{(r)}) \qquad (9)$$

in the formula (9) $h(t^*{}_i)$ is the imputed log survival time. δ_i is binary equal to 0 for censoring observation and equal to 1. For non-censoring. for each $h(t_i)$ if the log survival time is censored then it will be imputed by formula (9), by the summation of all survival difference multiplied with $h(t_r)$ for all $t_r > t_i$ and divided by the $S(t_{\ i})$. This imputation formula process is shown in Figure 3.1 for asymptotic censored t_i.

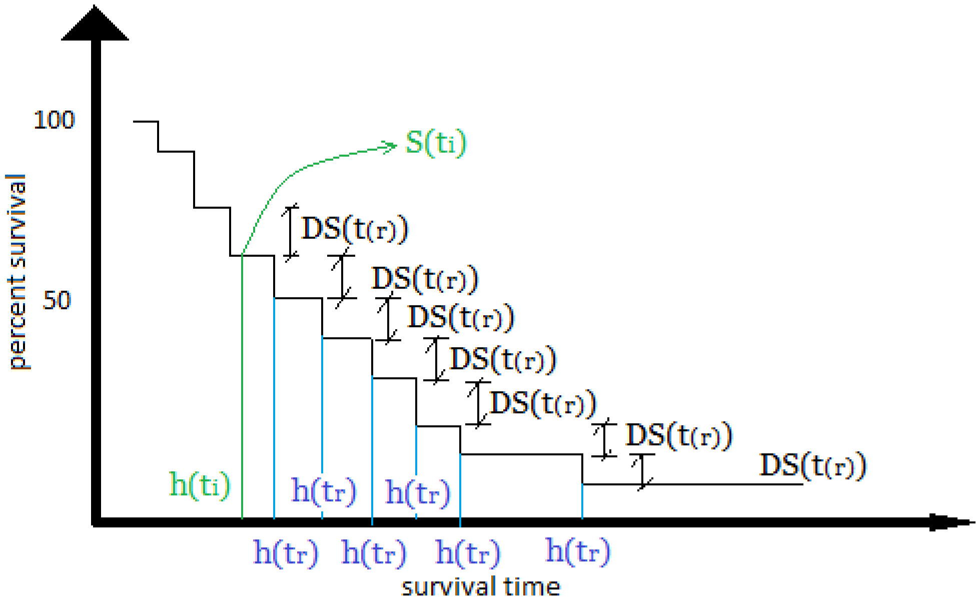

Figure 3.1. mean imputation shown for Kaplan-Meier survival function [39]

3.2.3 Kaplan-Meier estimator

Kaplan-Meier estimator is a non-parametric method for estimation of survival function. The estimator for S(t) is the probability that the survival time is more than t and it could be calculated by formula (10).

$$\hat{S}(t) \;=\; \prod_{i:\, t_i \leq t}\; 1 - \frac{d_i}{n_i} \tag{10}$$

where d_i is the number of deaths happened at time t_i and n_i is the number of observations who survived up to time t_i. Hence if there are in total 10 observations in the study and at t = 1, two of them are dead then S(t = 1) = 1 - 2/10 = 80%. If at t = 2, three observation are dead then S(t = 2) = (1-2/10)*(1 -3/8) = 0.5. If at t = 3, one observation is censored and 1 were dead then S(t = 3) = (1-2/10)*(1-3/8)*(1-1/5)= 0.4. If at t = 4, two were dead then S(t = 4) = (1-2/10) * (1-3/8) * (1-1/5) * (1-2/3) = 0.13. As it was seen the censoring event was not counted in the denominator in the calculation of the survival up to time t_4.

3.2.4. Semi-Supervised Learning

The methodology presented by Liang et al, (2016) [34] has been used for semi-supervised learning algorithm. The algorithm workflow is presented in Figure 3.2. In the preprocessing step the missing values of the data were removed, and the data was normalized to have mean 0 and 1 standard deviation. Then as is shown in the scheme (Figure 3.2) the completed data was selected as those data which are non-censored. After separating the censored data from the non-censored data (completed data), the $L_{1/2}$ regularized cox model was fitted on the completed data. Then the completed data was classified as “low risk” and “high risk”. by doing prediction using the selected parameters of the cox $L_{1/2}$ regularized model. Then after separating the high-risk data from low risk. Kaplan-Meier estimator was calculated for each of these classes separately, so that two survival functions are available, one for low risk and one for high risk data. Then based on the classes specified for all censored data the mean imputation method was used to impute the censored survival time as explained in the previous section.

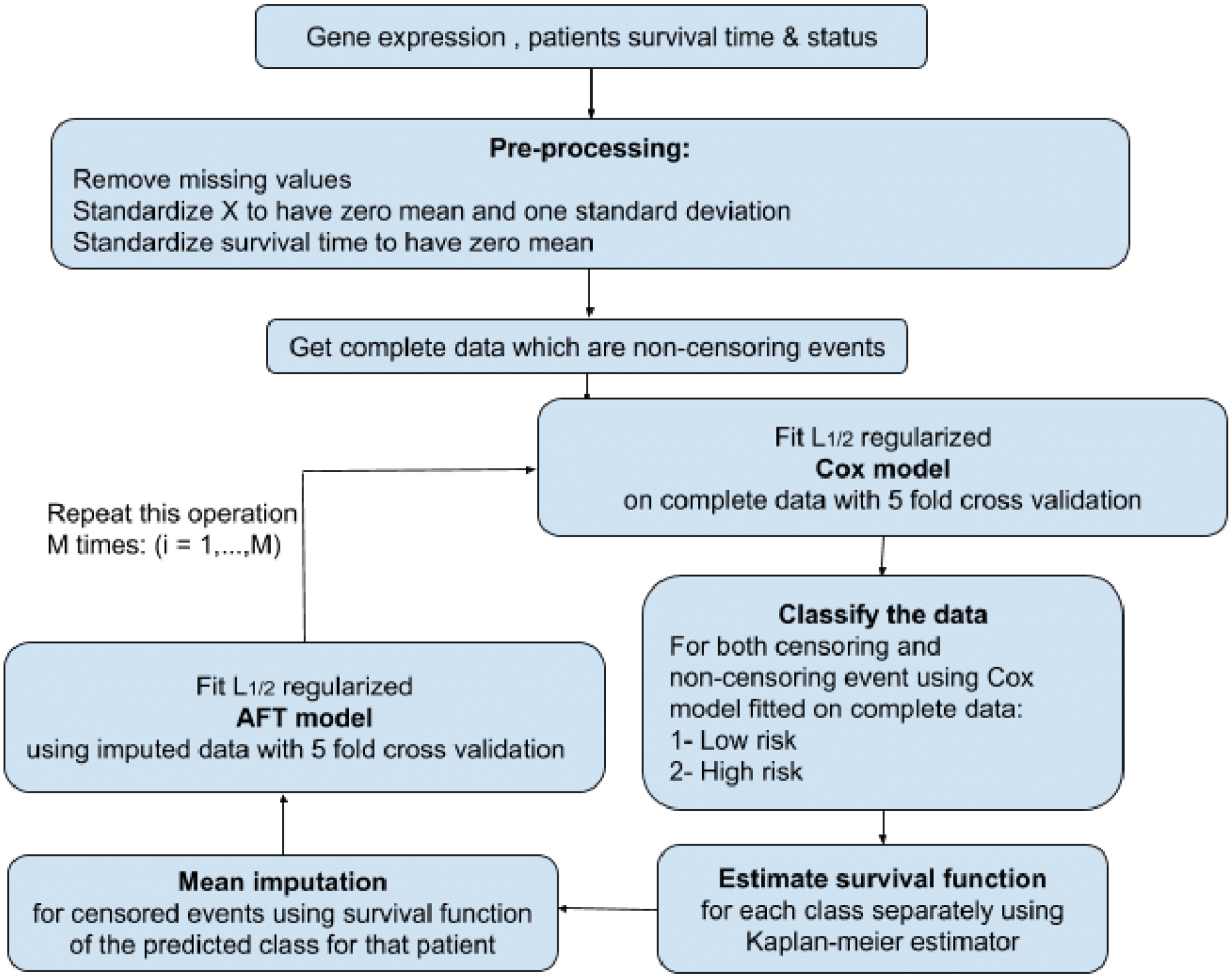
Gene expression , patients survival time & status
Pre-processing:
Remove missing values
Standardize X to have zero mean and one standard deviation
Standardize survival time to have zero mean
Get complete data which are non-censoring events
Fit L1/2 regularized
Cox model
on complete data with 5 fold cross validation
Repeat this operation
M times: (i = 1,...,M)
Classify the data
For both censoring and
non-censoring event using Cox
model fitted on complete data:
1- Low risk
2- High risk
Fit L1/2 regularized
AFT model
using imputed data with 5 fold cross validation
Mean imputation
for censored events using survival function
of the predicted class for that patient
Estimate survival function
for each class separately using
Kaplan-meier estimator

Figure 3.2. workflow for semi-supervised learning (Liang et al, 2016) [34]

After imputation, a $L_{1/2}$ regularized AFT model was fitted to the data. Then by predicting the survival time, the survival time of the censoring was checked to see whether it is correct or not. As stated by Liang et al, (2016) [34] those predictions which were even less than censoring time were considered as error in the estimation and was excluded from the training model. The operation of semi-supervised learning is repeated several times until the censoring time be replaced with a proper value and remain censored time in the data as low as possible.

3.2.5 Coordinate descent algorithm for L½ regularized Cox

In coordinate descent algorithm all the parameters remain fixed and only one parameter is free. The minimum of the log-partial likelihood for that parameter will be solved. For all parameters β_j, j = 1 to p, in each loop one coefficient will be free and other parameters are constant (the last updated value from the previous loop). Then $\beta_j$will be estimated. In this way the problem reduces from system of equation to one equation at each loop. Hence, using Newton-Raphson algorithm for solving the equation with one parameter will not require to calculate the Hessian matrix (which is the Jacobian of the gradient of the log-partial likelihood). The flowchart for the algorithm presented is as below:

==

Coordinate descent algorithm for cox L½ penalized model:

1- set all coefficients as zero. $\beta_1, \beta_2, \ldots, \beta_p = 0$

2- Loop from $\beta_1$to β_p:

calculate the 1st and 2nd order derivative of the partial log likelihood + L½ penalty as below:

first derivative:

$$\partial LL\,(\beta)/\partial\beta_1 = \sum_{i=1}^{n} \delta_i [X_{i1} - \sum_{l\,\epsilon\,R(T_i)} X_{l1} \boldsymbol{e}^{\sum_{j\neq 1} \beta_j X_{lj}} \boldsymbol{e}^{\boldsymbol{\beta}_1 X_{l1}} / \sum_{l\,\epsilon\,R(T_i)} \boldsymbol{e}^{\sum_{j\neq 1} \beta_j X_{lj}} \boldsymbol{e}^{\boldsymbol{\beta}_1 X_{l1}}]$$

$$+ \frac{\boldsymbol{\beta_1/|\beta_1|}}{\boldsymbol{2\sqrt{|\beta_1|}}} \boldsymbol{\lambda} = \boldsymbol{0}$$

2nd derivative:

$$\sum_{i=1}^{n} -\delta_i [\sum_{l \,\epsilon\, R(T_i)} X^2{}_{l1} \boldsymbol{e}^{\sum_{j \neq 1} \beta_j X_{lj}} \boldsymbol{e}^{\boldsymbol{\beta}_1 X_{l1}} \sum_{l \,\epsilon\, R(T_i)} \boldsymbol{e}^{\sum_{j \neq 1} \beta_j X_{lj}} \boldsymbol{e}^{\boldsymbol{\beta}_1 X_{l1}}$$

$$- (\sum_{\boldsymbol{l \,\epsilon\, R(T_i)}} X_{\;l1} \boldsymbol{e}^{\sum_{j \neq 1} \beta_j X_{lj}} \boldsymbol{e}^{\boldsymbol{\beta}_1 X_{l1}})^{\mathbf{2}}] / (\sum_{l \,\epsilon\, R(T_i)} \boldsymbol{e}^{\sum_{j \neq 1} \beta_j X_{lj}} \boldsymbol{e}^{\boldsymbol{\beta}_1 X_{l1}})^{\mathbf{2}}$$

$$-\lambda \sqrt{|\beta_1|} / 4 |\beta_1|$$

3- Newton-Raphson method: $\beta_{\;1(m+1)} = \beta_{\;1(m)} - \{\partial LL\,(\beta)/\partial\beta_1\} \,/\{\partial^2\, LL\,(\beta)/\partial\beta^2{}_1\}$

4- repeat 3 until converge and go to step to 2 choosing next β and repeat for all β_j from j = 1 to p

==

This procedure is complicated, and this method may have convergence issues especially because of the first and 2nd order derivative of the penalized part. Tibshirani, (1997) [39] and Cheng et al, (2014) [41] have shown that in the cox penalized model the loss function could be written as formula (11):

$$L\,(\beta)\ = argmin(\tfrac{1}{n}\{y_i - \textstyle\sum_{k\neq j}^{n}\ \beta_k X_{ik} - \beta_j X_{ij}\}^2\ + \lambda|\beta_j|^{1/2}) \qquad (11)$$

This formula (11) is easier for taking the derivative and for the solution using coordinate descent algorithm. since instead of log-partial likelihood of cox model the mean squared error is placed. In each loop only $\beta_j$is variant and other parameters remain as the last updated values. By using this loss function and considering, $\eta = \beta' X$, $u = -\partial l(\beta)/\partial\eta$, $A =\ -\partial^2 l/\partial\eta\partial\eta^T$ and $z = \eta + A^{-1}u$ then Tibshirani, (1997) [39] has shown that one term of taylor expansion for loss function of cox model will be as:

$$(z - \eta)^T\, A\, (z - \eta)$$

After that Cheng et al, (2014) [41] has used this loss function for penalized cox. they have used also $\hat{z} = (A^{1/2})\, z$ and $\hat{X} = (A^{1/2})\, X$. Their algorithm for coordinate descent by just replacing the new-half thresholding operator of Liang et al (2016) [34] will be as below:

===

Coordinate descent algorithm for cox L½ penalized model:

1- set all coefficients as zero. $\beta_1, \beta_2, \ldots, \beta_p = 0$ and λ = 0 and setting $m = 0$

2- calculate η(m), u(m) , A(m) , $\hat{X}(m)$ and $\hat{z}\,(m)$ from the current values of β(m)

3- minimizing $(z(m) - \hat{X}\,\beta(m))^T(z(m) - \hat{X}\,\beta(m)) + \lambda \sum_{j=1}^{p} |\beta_j(m)|^{1/2}$by repeating the following:

looping over j = 1, 2, … , p until $\beta(m)$ not changes

then calculate $\omega_j = \sum_{i=1}^{n} \hat{X}_{ij}(\hat{z}_i(m) - \sum_{k \neq j} \beta_k(m)\, \hat{X}_{ik})$and $\varphi_\lambda = arccos(\lambda/8 * (|\omega|/3)^{-3/2})$ th

$$\beta_{j(m)} = New\_half(\omega_j, \lambda) = 2/3\, \omega_j(1 + cos(\frac{2(\pi - \varphi_\lambda(\omega_j))}{3}))\ if\ |\omega_j| > \frac{\sqrt[3]{54}}{4}\,(\lambda)^{2/3}\ otherwise\ \beta_j = 0$$

set m = m + 1 and repeat step 2 and 3 until $\beta_{j(m)}$ converges.

===

This algorithm works fine in dealing with sparsity issues. (Cheng et al, 2014) [41].

for evaluation of the fitted model in the survival analysis, Integrated Brier score and Concordance index have been used.

3.2.6. Integrated Brier score

The Brier score changes over time and the integrated Brier score is the integral over the brier score from the minimum time to maximum time in the data.

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\widehat{S}(t \mid X_i)^2 \, 1(t_i \leq t \wedge \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1 - \widehat{S}(t \mid X_i))^2 \, 1(t_i > t)}{\hat{G}(t)}\right] \tag{12}$$

where $\widehat{S}(. \mid X_i)$ is the survival time estimated for patient i and $\hat{G}(.)$ is the Kaplan-Meier estimator for the censored data. the integrated brier score then is given by formula (13).

$$IBS = \frac{1}{max(t_i)} \int_{t=0}^{t=max(t_i)} BS(t)dt \tag{13}$$

lower values of IBS show better models. For a random model, the IBS is around 0.25. So, for a proper model it is expected to have IBS less than this value.

3.2.7. Concordance Index

Concordance index is interpreted as the fraction of the events in which their predicted survival is in correct order of the survival times. The formula for concordance index is presented in (14).

$$CI = \frac{\sum_i \sum_j 1(f_i < f_j \wedge \delta_i = 1)}{\sum_i \sum_j 1(t_i < t_j \wedge \delta_i = 1)} \tag{14}$$

where t is the survival time and f is the survival function. The values of CI are between 0 to 1. Higher values show better models. Values close to 1 show very good model and values close to 0.5 show random model.

Chapter 4

Results and Discussion

In this chapter the semi-supervised learning method is used to do the survival analysis for predicting the hazard rate. As a first step, data simulation is performed to check the semi-supervised learning method in variable detection using simulated data. In the second step, the methodology is checked using two real datasets. Censoring data is treated, and the mean imputation method is used for the censoring data to predict the risk class for each patient whether they are censored or not. The analysis is done on two real datasets of DLBCL (2002) [30] and Colorectal cancer dataset [31].

4.1. Simulation study

Survival data was simulated using the same method as Bender et al (2005) [42]. The simulation of the features is done as independently normal distribution with setting the correlation between the parameters to zero. Also, multivariate normal distribution data was simulated by setting the correlation between the parameters to 0.3. The survival time was simulated using Gompertz distribution with setting shape parameter $\alpha = 0.1$ and scale parameter $\omega = 1.0$. The simulation procedure is as follow:

1- There are 1000 simulated variables. The coefficients (β) for 990 of them are zero and only 10 of them are considered as prognostic genes and their coefficient was set as nonzero. At each simulated data 10 out of 1000 coefficients were randomly selected and set equal to 1 and the rest of the coefficients are set to zero.

2- randomly generate $\gamma_{i0}, \gamma_{i1}, \ldots, \gamma_{ip}$ (i=1, 2,...,n) for n observations and p parameters, from standard normal distribution which has zero mean and standard deviation equal to zero. Then using these independent random variables, the parameters X are simulated by adding a correlation between them. Correlations used are $\rho = 0$ and 0.3. The parameters were simulated using: $X_{ij} = \gamma_{ij}\sqrt{1-\rho} + \gamma_{i0}\sqrt{\rho}$. So that depending on the correlation coefficient the parameters will be simulated as correlated or independent parameters.

3- survival time was simulated using Gompertz distribution with $\alpha = 0.1$, $\omega = 1.0$. The survival times are calculated using: $y_i = 1 - \frac{\alpha * log(U)}{\omega * exp(\beta X)}$, where U is a random number from uniform distribution between 0 to 1, β are the coefficients and X are the simulated parameters.

4- The censoring times were simulated as random parameters by setting a rate for the number of censoring. 40% of the data were censored for each simulation. Hence for 40% of the survival times the survival time was replaced by random numbers. We have used uniform distribution between 0 to average value of the survival for these 40% of observations. Then the survival time for each of these 40% of the observations were taken as minimum of the simulated y and the censored time.

This method was used to simulate survival time and gene expression considering 40% of the observations as censored observations. The number of patients was taken as N = 100, 200 and 300, number of parameters as 1000 and correlation coefficient as 0 and 0.3. Censoring events is set to 0 for censored observation and 1 for uncensored observations. For each of these values 50 times data was simulated and the Cox model, AFT model, semi-supervised cox model and semi-supervised AFT model was fitted on them to find the prognostic genes in the dataset. There are 10 prognostic genes in each of these simulated data since only 10 coefficients are set to be nonzero. The data is high dimensional because the number of parameters is much more than the number of observations (P >> N). Hence a proper method is required to do parameter selection to find appropriate parameters within these 1000 parameters in the simulated data. For each of the simulated data and each of these four methods for parameter selection (Cox, AFT, semi-supervised Cox, semi supervised AFT) the analysis was repeated 50 times. The number of correctly selected parameters were calculated. Also, for each of them the total number of selected parameters were counted. The number of correctly selected parameters were considered as the average of the number of correctly selected parameters over 50 repetitions. The precision of each model was considered as division of average correctly selected parameters by average of total selected parameters over 50 repetitions.

4.1.1 Cox method

Cox proportional hazard method was used to fit each one of the simulated data. Since the data is high dimensional (P >> N), the $L_{1/2}$ penalty was used to reduce the number of parameters and select the prognostic genes. There are 10 prognostic genes in each simulated data. The cox method was fitted on each of them with 5-fold cross validation. The tuning parameter λ was selected as the value which has the highest Concordance Index (CI) in the testing data over these 5-fold cross validations. Then using the best tuning parameter value (λ), the number of correctly selected parameters, the total number of selected parameters and precision of the cox method was computed for each of the 50 repetitions. Then, the average value of these measures are calculated. Table 4.1 shows the results of the simulation for Cox model using different parameters N = 100, 200 and 300 and $\rho\ =\ 0$ and 0.3.

Table 4.1. Simulation results for cox proportional hazard model with $L_{1/2}$ penalty.

Cor.	Size	Single COX		
		Correct	Selected	Precision
$\rho = 0$	100	4.60 (1.63)	24.10 (2.25)	0.191 (0.250)
	200	7.64 (1.43)	30.98 (2.03)	0.247 (0.128)
	300	8.42 (0.99)	31.62 (1.73)	0.266 (0.135)
$\rho = 0.3$	100	3.88 (1.99)	29.36 (2.97)	0.132 (0.248)
	200	7.32 (1.43)	41.94 (2.25)	0.175 (0.115)
	300	8.14 (1.14)	39.40 (2.24)	0.207 (0.102)

It could be seen in Table 4.1 that for higher correlation between the parameters or in other words using non-independent parameters, the efficiency of the cox model is reduced a bit, since for N = 100, 200 and 300 the number of correctly selected parameters in $\rho = 0$ are 4.6, 7.64 and 8.42 out of 10 in the average of 50 simulated data. While for $\rho = 0.3$, the average number of correctly selected parameters are 3.88, 7.32 and 8.14 respectively which are a bit less than independent parameters. The precision of the model for average correctly selected parameters to total selected parameters is 0.19, 0.24 and 0.26 for $\rho = 0$ while it is a bit lower in the correlated parameters $\rho = 0.3$. The precision is 0.13, 0.17 and 0.2 for correlated parameters for N = 100, 200 and 300, respectively. The values inside the parentheses show the standard error of the average value for the number of correctly selected parameters, the total number of selected parameters and the precision.

4.1.2 AFT method

The simulated data was tested using the AFT method. For the AFT method also the parameter selection was done using $L_{1/2}$ penalty. The best value of tuning parameter λ was found by implementing 5-fold cross validation. The value which returns the highest concordance index (CI) in the prediction of survival of the testing data. The results of the simulation for AFT method were also computed as the average over 50 repetitions. The results of parameter selection for AFT model with $L_{1/2}$ penalty is presented in Table 4.2.

Table 4.2. Simulation results for AFT model with $L_{1/2}$ penalty.

Cor.	Size	Single AFT		
		Correct	Selected	Precision
$\rho = 0$	100	3.22 (1.57)	46.36 (2.51)	0.069 (0.037)
	200	5.66 (1.41)	49.94 (2.46)	0.113 (0.058)
	300	7.28 (1.34)	36.38 (1.99)	0.200 (0.121)
$\rho = 0.3$	100	3.04 (1.51)	46.94 (2.62)	0.065 (0.040)
	200	5.96 (1.34)	44.08 (1.91)	0.135 (0.064)
	300	7.24 (1.51)	36.48 (1.76)	0.198 (0.071)

The results of simulation for AFT method are about less than cox method for both correlation coefficients. The results show that in AFT method also, the efficiency reduced by using correlated parameters when compared with independent parameters. Results show that for N = 100, 200 and 300 the average number of correctly selected parameters are 3.22, 5.66 and 7.28 for $\rho = 0$ while for correlated parameters using N = 100, 200 and 300 that was 3.04, 5.96 and 7.24 which is a bit lower in N = 100 and 300. but it is not lower in N = 200. The correlated parameters have less precision than the independent parameters except for N = 200. The total number of selected parameters is not much different between independent and correlated parameters except for N = 200 which is a little less (44 < 49) value for correlated parameters.

4.1.3 Semi-supervised Cox method

As illustrated in the methodology Chapter 3(section 3.2.3), semi-supervised learning method was used for training and parameter selection for each of the simulated data. It is an iterative method which obtains and replaces the censoring survival time after each loop. The number of iterations used for each of simulated data and in each of 50 repetitions, 3 times iteration were done in which the cox model was fitted on complete data. The risk rate was predicted for each observation. Then the Kaplan-Meier estimator was used to find empirical distribution for high and low risk data separately. After that the censoring observations were replaced using the average imputation method as explained in Chapter (section 3.2.2). Using the imputed data, the cox model was fitted using $L_{1/2}$ penalty and the tuning parameter λ was estimated by using 5-fold cross validation. This operation was iterated 3 times until the results of the model were more realistic by considering the imputed censoring times instead of censored time. This method shows a good improvement compared with the Cox method. The analysis was repeated 50 times for each simulated data. The results for the semi-supervised cox model are presented in Table 4.3. The results of semi-supervised learning show a very high accuracy in detection of the correct parameters. For N = 200 and 300 almost all 10 parameters were detected correctly when the correlation between the parameters is zero. The average total number of selected parameters for N = 200 and 300 are 25.9 and 25.8, which means 38% and 39% of the total selected parameters are detected correctly.

Table 4.3. Simulation results for semi-supervised cox model with $L_{1/2}$ penalty.

Cor.	Size	Semi COX		
		Correct	Selected	Precision
$\rho = 0$	100	5.98 (1.76)	24.50 (0.91)	0.269 (0.132)
	200	9.96 (0.20)	25.98 (0.33)	0.387 (0.038)
	300	10.00 (0.00)	25.84 (0.43)	0.393 (0.049)
$\rho = 0.3$	100	3.70 (1.63)	25.64 (0.49)	0.144 (0.058)
	200	5.36 (1.71)	25.08 (0.43)	0.213 (0.062)
	300	6.96 (1.85)	23.96 (0.48)	0.289 (0.067)

For the correlated parameters, the value of the number of correctly selected parameters is only higher in N = 300. But the precision is higher in correlated parameters for N = 100, 200 and 300 compared with the cox method. The precision of semi-cox is 14%, 21% and 28.9% while the precision was 13%, 17%, 20% in the cox method.

4.1.4 Semi-supervised AFT method

In the semi-supervised AFT method, the same procedure was followed as explained in semi-supervised cox method, that after imputation of censored data using the average imputation method. An AFT method with $L_{1/2}$ penalty is fitted. The semi-supervised AFT method was executed using 3 iterations for imputation of censoring and then parameter selection. The results of the semi-supervised AFT method for each simulated data are presented in Table 4.4. The results show a very high value in the number of correctly selected parameters for independent data with 9.82 and 10 out of 10 prognostic parameters were detected correctly in the average of 50 repetitions for N = 200 and 300, respectively.

Table 4.4. Simulation results for semi-supervised cox model with $L_{1/2}$ penalty.

Cor.	Size	Semi AFT		
		Correct	Selected	Precision
$\rho = 0$	100	3.04 (2.54)	11.00 (1.52)	0.276 (0.256)
	200	9.82 (0.66)	24.30 (1.82)	0.510 (0.238)
	300	10.00 (0.00)	22.06 (1.54)	0.557 (0.240)
$\rho = 0.3$	100	2.06 (1.35)	18.66 (1.50)	0.124 (0.094)
	200	4.72 (1.95)	24.40 (1.54)	0.208 (0.086)
	300	7.00 (1.96)	26.74 (1.64)	0.301 (0.137)

The total number of selected parameters for N=200 and N=300 is 24 and 22, respectively with a precision of 51% and 55.7% respectively. The semi-AFT method does not show an improvement in the number of correctly selected correlated parameters, the precision of the semi-AFT is higher than AFT method for correlated data with precision of 12%, 20.8% and 30.1% for N = 100, 200 and 300 respectively while in AFT the precision was 6.5%, 13.5% and 19.8% for N = 100, 200 and 300, respectively.

After implementing semi-supervised learning method for each sample size N = 100, 200 and 300 and for each $\rho = 0$ and 0.3, higher proportion of the data could be classified as high risk and low risk, since the censoring data is reduced after implementing the mean imputation method. If we consider 6 simulated data as: {a: N = 100, $\rho = 0$. b: N = 200, $\rho = 0$, c: N = 300, $\rho = 0$, d: N = 100, $\rho = 0.3$, e: N = 200, $\rho = 0.3$ and f: N = 300, $\rho = 0.3$ } then the proportion of the high risk, low risk and censored observations before and after implementing the semi-supervised learning method could be seen in Figure 4.1.

As we can see in the Figure 4.1, 40% of the data were censored in the simulated data. However, after implementing semi-supervised learning method only a few of them remained as Censored and most of them could be classified as high risk and low risk observations for all sample sizes and correlation coefficients. For small sample size (N=100), the proportion which remains as uncensored is a little more compared with others. The proportion of the data which remained as censored is around 5%.

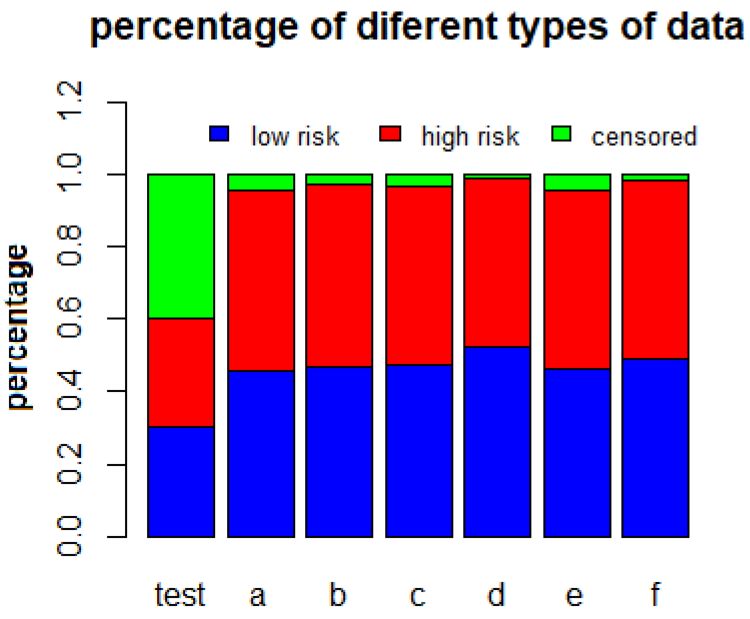

Figure 4.1. proportion of High risk, low risk and censored data after semi-supervised learning

4.1.5 Semi-supervised learning method on Test Data

After training on simulated data, semi-supervised learning was applied on test data. For this reason, other data were simulated with N = 200 for testing. To simulate the data for testing, same coefficients, correlation and the same Gompertz distribution were used for each data which was simulated for training. Then the model of the training data was used to predict the risk in the test data. The procedure is as explained below:

1- making a semi-supervised learning model for each of training simulated data, for N = 100, 200, 300 with $\rho = 0$ and 0.3.

2- predicting the risk for the test data which was simulated with N = 200 and the same coefficients as the training data, as low risk and high risk.

3- making a semi-supervised learning model with the test data and classifying the test data into low risk and high risk and those which remain as censored.

4- The risk values of the testing data which were predicted in step 2 will be compared with the risk values of step 3 to find the number of correctly classified as high risk, low risk and the error rates.

The procedure explained above was used to test the semi-supervised learning method.

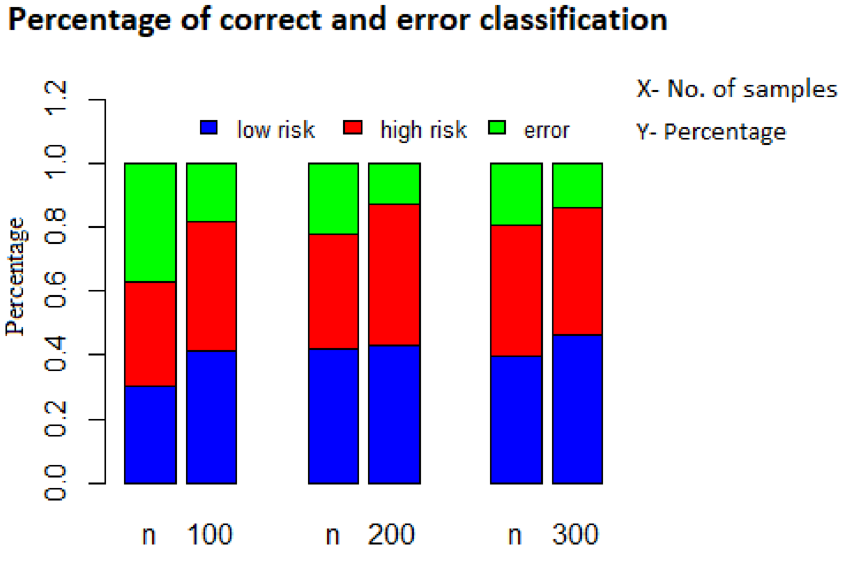

Figure 4.2. Percentage of error and correct classification obtained using semi-supervised learning method

The results for the test data are presented in Figure 4.2. It can be seen that increasing the sample size, the error rate is reduced from around 40% and 20% in N = 100 and $\rho = 0$ and 0.3, respectively to around 20% and 10% in N = 200, respectively. For N = 300 the error rate is around 20% and 15% for $\rho = 0$ and 0.3, respectively.

4.2 Survival analysis for Real data

4.2.1 Results of analysis for DLBCL (2002) dataset

Semi-supervised learning analysis was done using DLBCL (2002) dataset. 102 out of 240 observations in this data are right censored (42.5%). The survival analysis was done using semi-supervised learning to classify the observations into low risk and high risk. As the first step, like in the simulation study, the number of selected parameters from 7399 genes present in this dataset, are calculated using a single cox model, single AFT model, semi-supervised cox model and semi-supervised AFT model. The number of selected parameters was found to be lower using a semi-supervised learning method compared with single cox or AFT method. Figure 4.3. Shows the number of selected parameters in each of these four methods. The number of selected parameters in semi-cox and semi-AFT is about 40. While in single cox around 50 parameters are selected and in single AFT around 60 parameters are selected. In the study done by Liang et al, (2016) [34] the number of selected parameters for DLBCL (2002) data were as 42, 28, 60 and 54 for single cox, semi-cox, single AFT and semi-AFT. The results we get is close to what they obtained as the minimum number of selected parameters in semi-cox model by our method was around 40 parameters. Single AFT around 60 parameters are selected. In the study done by Liang et al, (2016) [34] the number of selected parameters for DLBCL (2002) data were as 42, 28, 60 and 54 for single cox, semi-cox, single AFT and semi-AFT.

The results we get is close to what they obtained as the minimum number of selected parameters in semi-cox model by our method was around 40 parameters.

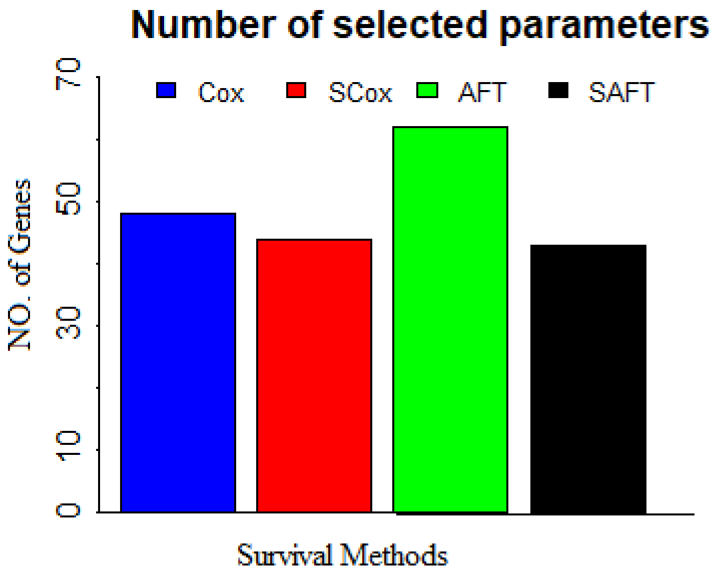

Figure 4.3. Number of selected parameters in single model and semi-supervised learning methods.

While 42.5% of the samples were censored observations, after implementing the semi-supervised learning method only 3% of the observations remained as censored and the risk for the rest of observations could be predicted as low risk or high risk. Figure 4.4. Shows the bar plot for censoring before and after implementing the semi-supervised learning method.

The results for classifying the observations in low risk and high risk is close to the results presented by Liang et al, (2016) [34] after implementing semi-supervised learning they get 50% of the data as low risk and 46% as high risk and 3.3% percent remained as censored, but in our implementation, we get around 65% of the data as low risk, 38% as high risk and 3% remained censored.

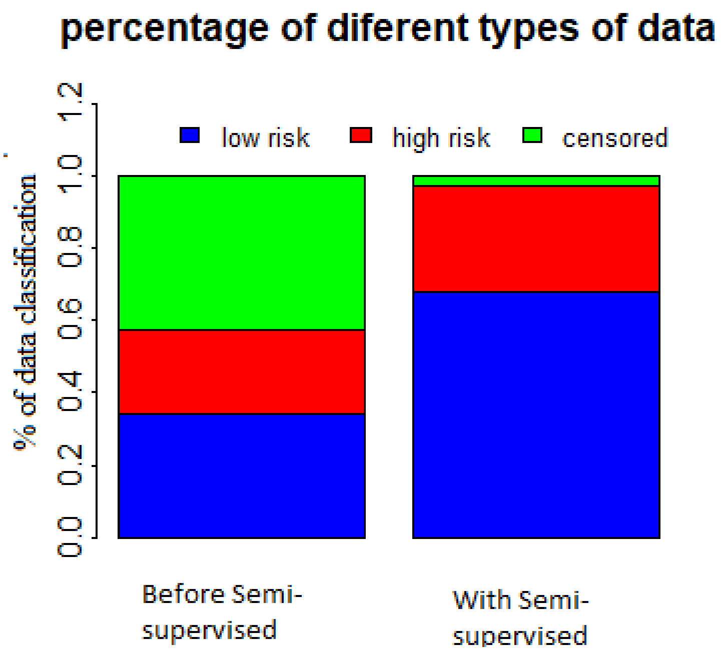

Figure 4.4. Percentage of low risk, high risk and censored observations before and after semi-supervised learning.

Evaluation of the fit for these survival methods was done by using concordance index and integrated brier score (IBS). The concordance index (CI) for these four models is presented in Figure 4.5. The concordance index (CI) for the semi-supervised learning method is quite higher compared with single cox and single AFT. The value of CI for semi cox and semi-AFT is around 0.75.

While for single cox and single AFT the CI is less than 0.6. Higher values for CI show better fit. The results of CI are close to the CI calculated by Liang et al, (2016) [34] in which they get CI: 0.62, 0.68, 0.65 and 0.74 for single cox, semi-cox, single AFT and semi-AFT respectively. The CI which we get in semi-supervised learning is a bit higher compared with the value found by Liang et al, (2016) [34].

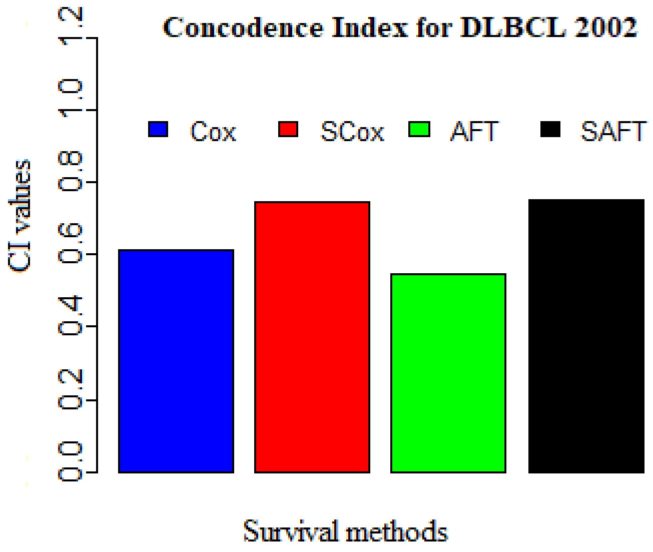

Figure 4.5. Concordance index (CI) for single cox, AFT and semi supervised learning cox and AFT method.

The predicted survival function was evaluated by calculation of integrated brier score (IBS). The Integrated Brier Score (IBS) for four methods of cox, semi-cox, AFT and semi-AFT is presented in Figure 4.6. It can be seen that the IBS for single cox and single AFT is around 0.22 which is only a bit less than 0.25. The IBS of 0.25 is derived for random models and do not show a good fit for Survival function.

For the IBS, the lower values are showing better fit. The value of IBS for semi-cox and semi-AFT is less than 0.1 which shows a good improvement compared with single cox and single AFT. The IBS we get using semi-supervised learning is less than 0.1 while in the analysis performed by Liang et al, (2016) [34] the IBS for DLBCL2002 was computed were 0.12 and 0.13 for semi-cox and semi-AFT, respectively.

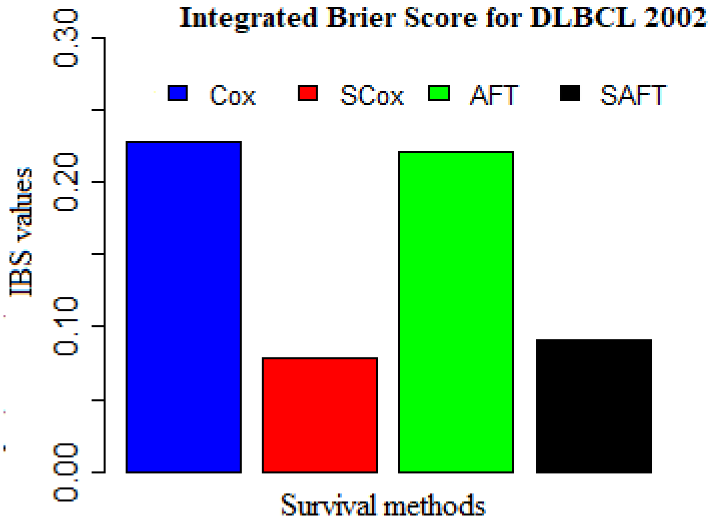

Figure 4.6. Integrated brier score (IBS) for DLBCL 2002 data using four methods

This shows that the predicted survival function is a bit better in the current analysis compared with the previously done analysis and these results may vary with the latest or different datasets.

4.2.2 Results of analysis for Colorectal Cancer dataset

Another real data was used to test the semi-supervised learning method. This data is also high dimensional (p >> N) 54,675 gene expressions are included. The data includes 177 patients, and the censoring is much more in this data (58.8%). So, 104 patients out of 177 are censored and only 73 (41.2%) of the patients are uncensored. In this data most of the observations are censored and this data is very big which requires much more time for training.

The semi-supervised learning methods with single cox and single AFT models were used to do parameter selection. The number of selected parameters for each of these four models are presented in Figure 4.7.

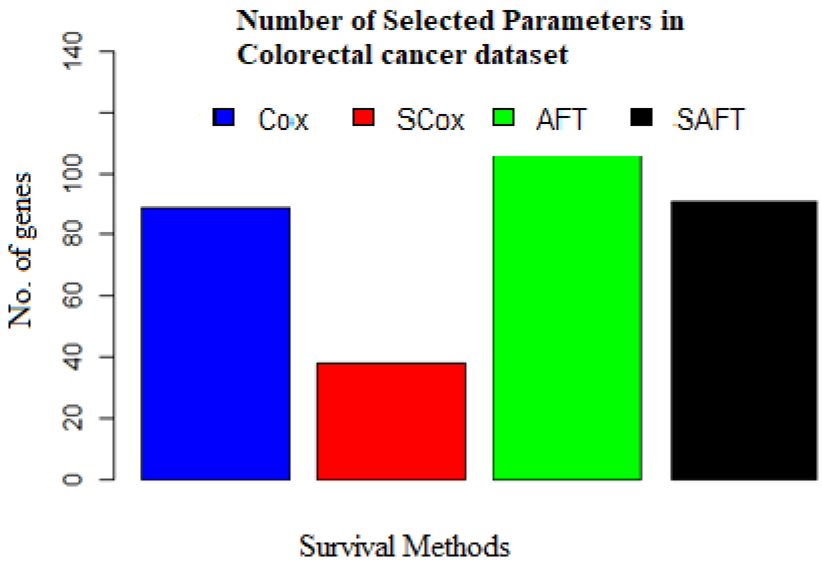

Figure 4.7. Number of selected parameters in colorectal data using four methods

While in single cox, single AFT and semi-aft more than 80 parameters were selected by the models, in the semi-cox model less than 40 parameters were selected.

In this data the proportion of censoring observations were more than the uncensored observations. After implementing a semi-supervised learning method, the proportion of censoring reduced from 58.8% to less than 40%. Figure 4.8 shows the proportion of the censoring, low risk and high-risk observations before and after implementing the semi-supervised learning method.

In the presented dataset by (Smith et al, 2010) [31] the colorectal data include a column for grade. The dataset grade was 134 (75.7%) as medium, 27 (15.25%) as low and 16 (9%) of them as high. The cross tabulation shows that from 134 medium grade 81 of them are censored and 53 of them are uncensored.

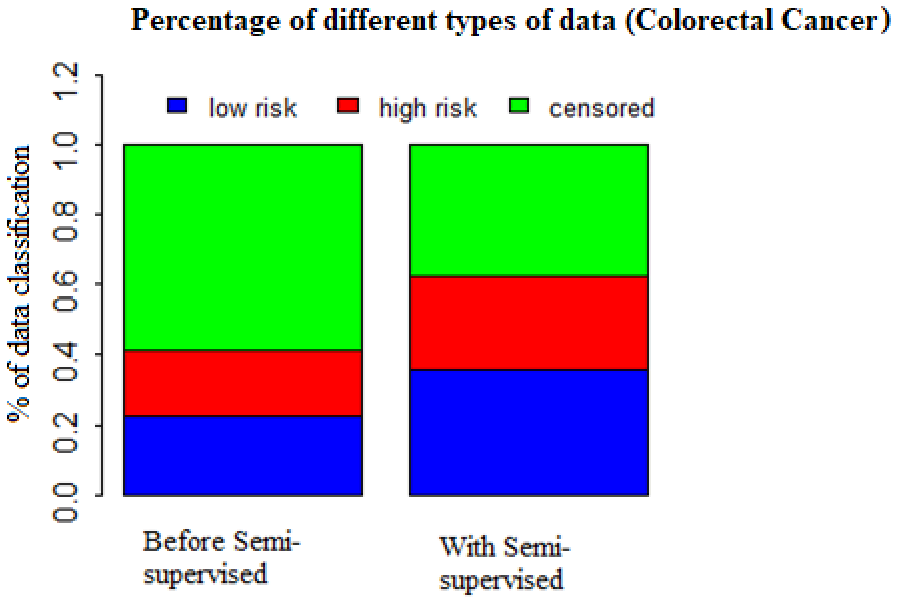

Figure 4.8. Proportion of high risk, low risk and censoring before and after semi-supervised learning

Hence from 104 censored observations. 81 of them are in medium grade and only 23 of them are in high and low grade. From the uncensored data only 20 of them are in low and high grade. Table 4.5. Shows the cross tabulation of high, low, and medium grade verse censored and uncensored observations.

Here in the data before semi-supervised learning the fraction of low, high and censoring observations is 21.2%, 20% and 58.8%, respectively. After semi-supervised learning, the fraction of low risk increased to 37%, high risk increased to 25% and the censoring decreased to less than 40%. In Table 4.5, the low grade is more than high grade for uncensored. However, the majority of patients are categorized as medium grade.

Table 4.5. cross tabulation of censored and uncensored observations verse grades of colorectal data

Grade	Censored	Uncensored	summation
High	12	4	16 (9%)
Low	11	16	27 (15.3%)
Medium	81	53	134 (75.7%)
summation	104 (58.8%)	73 (41.2%)	177 (100%)

The quality of the fit was evaluated by using concordance index (CI) and integrated brier score (IBS). The concordance index for the four methods is presented in Figure 4.9.

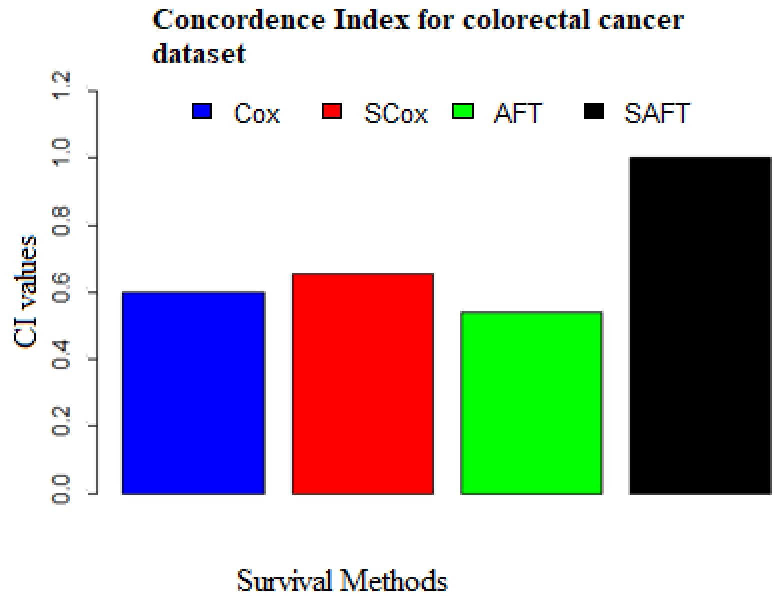

Figure 4.9. Concordance index for single cox, single aft, semi-cox, and semi-aft models for colorectal dataset.

Looking at the values of CI for the four models, semi-supervised AFT has the highest CI (0.95) compared with other models. The CI for semi-cox is around 0.63. The CI for single cox model is 0.6 and the CI for single AFT is 0.55.

The integrated brier score (IBS) for these four methods is also calculated. The IBS measure for single cox and single AFT are around 0.2 and 0.21, respectively. The IBS for semi-AFT is close to 0.16 and for semi-cox the IBS is the lowest which is around 0.12. The results of IBS show that the best survival function is estimated using a semi-supervised cox method, since it has the lowest IBS compared with the other methods. The value of CI for semi-AFT was surprisingly high which is probably due to overfitting to the dataset after some iteration in the imputation of the censored data.

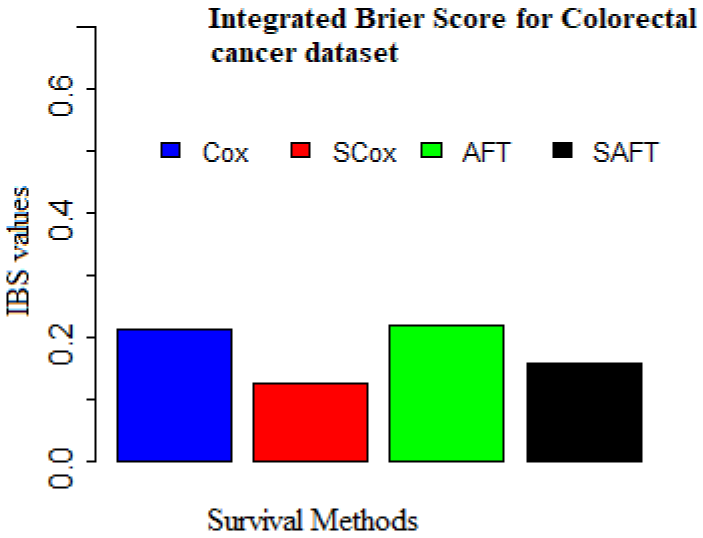

Figure 4.10. Integrated brier score (IBS) for four methods in colorectal dataset.

The integrated brier score (IBS) in Figure 4.10 shows that the semi-cox model performed best compared with other methods in the survival analysis of the colorectal data. This data set had a high proportion of censoring observation, as after semi-supervised learning methods there remained much censoring (40%) in the dataset. Neglecting these 40% of the censored data, the sample size is small and close to 100. We saw in the simulation study that the precision of the method for parameter selection for low sample size N = 100 was not as high as higher sample size like N = 200 and 300.

Chapter 5

Conclusions and Future Work

5.1. Conclusions for simulation study

Survival analysis on high dimensional data (P >> N) were studied in this study. The first phase of the study was related to data simulation. Several data with different sample sizes and correlation coefficients between the parameters were simulated. The survival analysis was done using a single cox model, a single AFT model, a semi-supervised cox, and semi-supervised AFT model. The results of simulation analysis showed that the efficiency of the methods for parameter selection in higher sample size is more compared with lower sample size. For the semi-supervised learning in both semi-cox and semi-AFT the results were promising using independent parameters with zero correlation between them. In the results of simulation study for correlated parameters it was observed that a better performance was not achieved for semi-supervised learning (both semi-cox and semi-AFT) for number of correctly selected parameters compared with single cox and single AFT. However, the precision of parameter selection was higher in both semi-cox and semi-AFT compared with single cox and single AFT. Hence for the correlated parameters the results were not as good as those obtained by Liang et al, (2016) [34] except for the precision of the semi-AFT which were higher in the current study. But for uncorrelated parameters the results of the semi-cox and semi-aft method was promising and more than what was obtained by Liang et al, (2016) [34] for both average number of selected parameters and the precision of the parameter selection. For this study 3 times iterations were performed for iterative procedure of fitting cox model on completed data, classifying the data, then data imputation and using AFT or cox model with $L_{1/2}$ penalty. More iterations lead to better results.

5.2 Conclusions for Real Data

5.2.1 DLBCL 2002

Four methods of single cox, single AFT, semi-cox and semi-AFT were used for parameter selection, censored data imputation and prediction of the classes for the DLBCL 2002 data. The number of selected parameters was lowest in the semi-cox model. After implementing a semi-supervised learning method only 3% of the observations remained as uncensored and the rest could be classified into low risk and high risk. The results show that semi supervised learning (both semi-cox and semi-AFT) performed better than other methods. The concordance index (CI) was the highest for semi-cox and semi-AFT models which show that the number of concordant observations is the highest in the semi-supervised learning models compared with other methods. The integrated brier score (IBS) shows the lowest value in the semi-cox model which shows that semi-cox is slightly better than semi-AFT.

5.2.2 Colorectal cancer data

The four models were used on the colorectal data also. The colorectal data was big data containing many gene expressions (54675) and most of the patients were censored (58.8%). After implementing a semi-supervised learning method, the number of censoring reduced by around 18.8% but still it was high (40%). The number of selected parameters were lowest in the semi-cox model and highest in single AFT model. The results of the semi-supervised learning model showed a very high value in the concordance index (CI) of semi-AFT which is probably due to overfitting to the data (since around 85 parameters) were remaining in the semi-AFT model. After semi-AFT, semi-cox had the highest concordance index (CI). In this data set again the semi-cox performed the best compared with other methods by having the lowest integrated brier score (IBS) compared with other methods. The number of selected parameters were also lowest in the semi-cox model.

5.3 Suggestions for future study

In this study we have implemented the semi-supervised learning on the dataset with small sample size with low number of observations as uncensored and very big number of gene expression. The analysis for this dataset is time consuming hence it requires use of GPU for fitting the semi-supervised learning method on the whole dataset and do some iterations for the semi-supervised learning method to reach a more accurate data. Implementing proper cross validation is required for estimating the concordance index but the difficulty is there that the sample size is small, and data is so big. Hence the cross validation could be unstable and requires many iterations to a proper result to avoid overfitting to the data. Hence using a GPU and sparse matrices could be helpful in increasing the performance of the algorithm. For this reason, python packages like TensorFlow are the easiest and most straight forward approach for programming.

The results for simulation study on the correlated data was a bit different from the one obtained by Liang et al, (2016) [34]. So, it is suggested to simulate more correlated data and investigate whether the performance of the semi-supervised learning is as good as uncorrelated data, or the efficiency of the model will be considerably reduced.

Another suggestion is to use data with supervised (observed) risks. It is helpful in better evaluation of the accuracy of the survival analysis using semi-supervised learning methods.

Another suggestion for a data which is big like the one which we used in our study (colorectal data) deep learning methods be implemented. Deep learning methods such as Artificial neural network connected with a survival layer which leads to cox proportional hazard loss function with and without including of the $L_{1/2}$ penalty for parameters selection seems to be useful and may lead to improvement in the results especially in the big and high dimensional datasets.

References

1. World Health Organization: https://www.who.int/
2. Darren R. Brenner, Hannah K. Weir, Alain A. Demers, Larry F. Ellison, Cheryl Louzado, Amanda Shaw, Donna Turner, Ryan R. Woods and Leah M. Smith; for the Canadian Cancer Statistics Advisory Committee. Projected estimates of cancer in Canada in 2020, CMAJ March 02, 2020 192 (9) E199-E205; DOI: https://doi.org/10.1503/cmaj.191292
3. Public Health Agency of Canada; Statistics Canada; Canadian Cancer Society; provincial/territorial cancer registries. Release notice - Canadian Cancer Statistics 2019. Health Promot Chronic Dis Prev Can. 2019 Sep;39(8-9):255. doi: 10.24095/hpcdp.39.8/9.04. PMID: 31517469; PMCID: PMC6756131
4. Global Cancer Observatory (GCO): Cancer Today, Cancer Factsheets. https://gco.iarc.fr/today/data/factsheets/populations/124-canada-fact-sheets.pdf
5. Canadian Cancer Society: 6 statistics that reveal the impact of cancer in Canada for 2020. https://www.cancer.ca/en/about-us/our-stories/6-statistics-about-cancer-in-canada-for-2020/?region=bc
6. Canadian Cancer Society: Childhood cancer statistics. https://www.cancer.ca/en/cancer-information/cancer-101/childhood-cancer-statistics/?region=on
7. Stages of Cancer, https://www.webmd.com/cancer/cancer-stages#1
8. Shankar Prinja, Nidhi Gupta and Ramesh Verma. "Censoring in clinical trials: Review of Survival analysis techniques." Indian J Community Med 35, no. 2 (2010): 217.

9. TG Clark, MJ Bradburn, SB Love and DG Altman. “Survival Analysis : basic concepts and first analysis.” British Journal of Cancer 89, 2003: 232

10. Swaminathan R and Brenner H, Statistical methods for cancer survival analysis, Vol. 2, 2011. https://survcan.iarc.fr/survivalchap2.php

11. Abadi, Alireza, Parvin Yavari, Monireh Dehghani-Arani, Hamid Alavi-Majd, Erfan Ghasemi, Farzaneh Amanpour, and Chris Bajdik. "Cox models survival analysis based on breast cancer treatments." *Iranian journal of cancer prevention* 7, no. 3 (2014): 124.

12. Manish Devgan, Gaurav Malik, Deepak Kumar Sharma, *Machine learning & Big data, Semi-supervised Learning, 2020,* doi: 10.1002/9781119654834.ch10

13. Xiaojin Zhu; Andrew Goldberg, *Introduction to Semi-Supervised Learning*, Morgan & Claypool, 2009, doi: 10.2200/S00196ED1V01Y200906AIM006.

14. Ecloudvalley digital technology co Ltd: “Types of machine learning methods” https://www.ecloudvalley.com/mlintroduction/

15. Abdulaziz Aflakseir and Parinaz Abbasi. Health Beliefs as Predictors of Breast Cancer Screening Behavior in a Group of Female Employees in Shiraz, *Iranian Journal of Cancer Prevention*, 2012; 3:124-129

16. Canadian Partnership Against Cancer: Canadian Strategy for Cancer Control. https://www.partnershipagainstcancer.ca/cancer-strategy/

17. Thongpim, Nattawut, Chidchanok Choksuchat, Tanan Bejrananda, and Sureena Matayong. "On Predicting Survival Opportunities for Prostate Cancer by COX Regression in PSU Patients Data." In *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI-CON)*, pp. 775-778. IEEE, 2020.

18. Chai, Hua, Zi-na Li, De-yu Meng, Liang-yong Xia, and Yong Liang. "A new semi- supervised learning model combined with cox and sp-aft models in cancer survival analysis." *Scientific reports* 7, no. 1 (2017): 1-12.

19. Wang, Qingyong, Liang-Yong Xia, Hua Chai, and Yun Zhou. "Semi-supervised learning with ensemble self-training for cancer classification." In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 796-803. IEEE, 2018.

20. Qiu, John X., Shang Gao, Mohammed Alawad, Noah Schaefferkoetter, Folami Alamudun, Hong-Jun Yoon, Xiao-Cheng Wu, and Georgia Tourassi. "Semi-Supervised Information Extraction for Cancer Pathology Reports." In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1-4. IEEE, 2019.

21. Kabir, Md Faisal, and Simone A. Ludwig. "Classification Models and Survival Analysis for Prostate Cancer Using RNA Sequencing and Clinical Data." In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 2736-2745. IEEE, 2019.

22. Huang, HaiHui, and Yong Liang. "A novel Cox proportional hazards model for high-dimensional genomic data in cancer prognosis." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019).

23. Nezhad, Milad Zafar, Najibesadat Sadati, Kai Yang, and Dongxiao Zhu. "A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer." *Expert Systems with Applications* 115 (2019): 16-26.

24. Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network." *BMC medical research methodology* 18, no. 1 (2018): 24.

25. Hirozawa, Tatsuki, Takeshi Yamada, and Hayato Ohwada. "New survival prediction system for terminal patients based on machine learning." In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* pp. 2756-2758. IEEE, 2018.

26. Yang, Cheng-Hong, Sin-Hua Moi, Fu Ou-Yang, Li-Yeh Chuang, Ming-Feng Hou, and Yu-Da Lin. "Identifying Risk Stratification Associated with a Cancer for Overall Survival by Deep Learning-Based CoxPH." *IEEE Access* 7 (2019): 67708-67717.

27. Barsainya, Aditya, Anusha Sairam, and Annapurna P. Patil. "Analysis and Prediction of Survival after Colorectal Chemotherapy using Machine Learning Models." In *2018 International Conference on Advances in Computing, Communications, and Informatics (ICACCI),* pp. 862-865. IEEE, 2018.

28. Cai, Zhuqing, Zhuliang Yu, Haiyu Zhou, and Zhenghui Gu. "The early stage lung cancer prognosis prediction model based on support vector machine." In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pp. 1-4. IEEE, 2018.

29. Imani, Farhad, Ruimin Chen, Conrad Tucker, and Hui Yang. "Random Forest Modeling for Survival Analysis of Cancer Recurrences." In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE),* pp. 399-404. IEEE, 2019.

30. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, López-Guillermo A, Grogan

TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM; (2002). Lymphoma/Leukemia Molecular Profiling Project. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.; 346(25):1937-47.

31. Smith JJ, Deane NG, Wu F, Merchant NB et al. (2010). Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. 138(3):958-68. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17536
32. Cox, D.R(1972). : Regression Models and Life Tables, Journal of the Royal Statistical Society, Series B, (Methodological),Vol 34, No. 2, 187-220.
33. Wei, L.J. (1992). The Accelerated Failure Time Mode: A Useful Alternative To the Cox Regression Model in Survival Analysis, Statistics in Medicine, 1871-1879.
34. Liang, Y., Chai, H., Liu, XY. *et al.* (2016). Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with L½ regularization. *BMC Med Genomics* 9, 11.
35. González-González, D. & Pina-Monarrez, M. & Torres-Treviño, L.. (2008). Estimation of Parameters in Cox's Proportional Hazard Model: Comparisons between Evolutionary Algorithms and the Newton-Raphson Approach. 513-523. 10.1007/978-3-540-88636-5_49.
36. Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso."Journal of the Royal Statistical Society,Series B58:267–288.
37. Xu ZB, et al. (2010). L½ regularization. Sci China. 40(3):1–11. series F.
38. Gui J, Li H. (2005). Penalized Cox regression analysis in the high- dimensional and low sample size settings, with applications to microarray gene expression data. Bioinformatics. 21(13):3001–8.
39. Tibshirani R. (1997). The lasso method for variable selection in the Cox model.

https://pubmed.ncbi.nlm.nih.gov/9044528/

40. Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457-481. https://web.stanford.edu/~lutian/coursepdf/KMpaper.pdf

41. Cheng, L., Yong L., Xin-Ze L., Kwong-Sak L., Tak-Ming C., Zong-Ben X., Hai Zhang. (2014). The L½ regularization method for variable selection in the Cox model, Applied Soft Computing, (14). 498-503,1568-4946.

42. Bender R, Augustin T, Blettner M. (2005). Generating survival times to simulate Coxproportional hazards models. 24:1713–23