December 2016

Stage-Specific Predictive Models for Cancer Survivability

Elham Sagheb Hossein Pour
University of Wisconsin-Milwaukee

Stage-Specific Predictive Models for Cancer Survivability

by

Elham Sagheb Hossein Pour

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Computer Science

at

The University of Wisconsin–Milwaukee
December 2016

Abstract

Stage-Specific Predictive Models for Cancer Survivability

by

Elham Sagheb Hossein Pour

The University of Wisconsin–Milwaukee, 2016
Under the Supervision of Professor Rohit J. Kate

Survivability of cancer strongly depends on the stage of cancer. In most previous works, machine learning survivability prediction models for a particular cancer, were trained and evaluated together on all stages of the cancer. In this work, we trained and evaluated survivability prediction models for five major cancers, together on all stages and separately for every stage. We named these models joint and stage-specific models respectively. The obtained results for the cancers which we investigated reveal that, the best model to predict the survivability of the cancer for one specific stage is the model which is specifically built for that stage. Additionally, we saw that for every stage of cancer, the most important features to predict survivability, differed from other stages. By evaluating the models separately on different stages we found that their performance differed on different stages. We also found that evaluating the models together on all stages, as was done in past, is misleading because it overestimates performance.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to sincerely thank my advisor Prof. Rohit J. Kate, who has mentored me tirelessly. I express my deepest appreciation for his continuous support of my research activities, for his patience, motivation, and encouragement.

Besides my advisor, I would like to thank all my thesis committee members: Prof. Hossein Hosseini and Prof. Jun Zhang for their invaluable suggestions and guidances.

I want to thank the Computer Science department at University of Wisconsin-Milwaukee for supporting me during my career, situating me in environments favorable to learn and the academic freedom to discover.

Last but not least, I wish to thank my mother, Tayyebeh and my father, Reza for their kindness, support, and the encouragement, and also my family, Sara and Ahmad for shaping the world around me and for supporting my dreams.

Chapter 1

Introduction

1.1 Background and Problem Statement

Cancer as one of the top leading causes of mortality worldwide is a generic name given to a collection of related diseases that stimulate some of the body's cells start to divide without stopping [3], and the rapid extension of those abnormal cells violates their usual boundaries, leading to metastasizing in which the malformed cells spread across other organs. This process is the major cause of death from cancer diseases all across the world, accounting for about 8.2 million deaths in 2012 [48]. There have been more than hundred types of cancer, including breast, liver, stomach, prostate, colorectal, lung, and brain purposes to name a few. After heart diseases, cancer was the second leading cause of death in the United States in 2014 while 22.52% of total deaths were because of cancer [2], [34]. From 2009 to 2013, concerning the United States, the most common cancer incidences were related to female breast cancer, prostate cancer, lung & bronchus cancer, colon and rectum cancer, and corpus & uterus; NOS respectively [1]. Figure 1.1 shows the top five cancers with highest incidence rates beside the stomach cancer (since it is among the set of cancers in this study) in the United States from 2009 to 2013 [1].

Given the fact of the high rate cancer mortality and the number of people who are involving with such a disease, the study of cancer and its survivability rates has been a longstanding research in the biomedical literature [12], [14], [18], [21], [42], [49], [52]. ***Cancer survival rates*** indicate the percentage of cancer patients who survive a certain kind of cancer for a specific period of time. For a particular cancer, the n-year survival rate is the percentage of cancer patients who live at least n years after being diagnosed with the cancer. For example, in the United State, from 2006 to 2012 the overall 5-year survival rate for bladder cancer was 77.5% [4], meaning that of all those people who have bladder

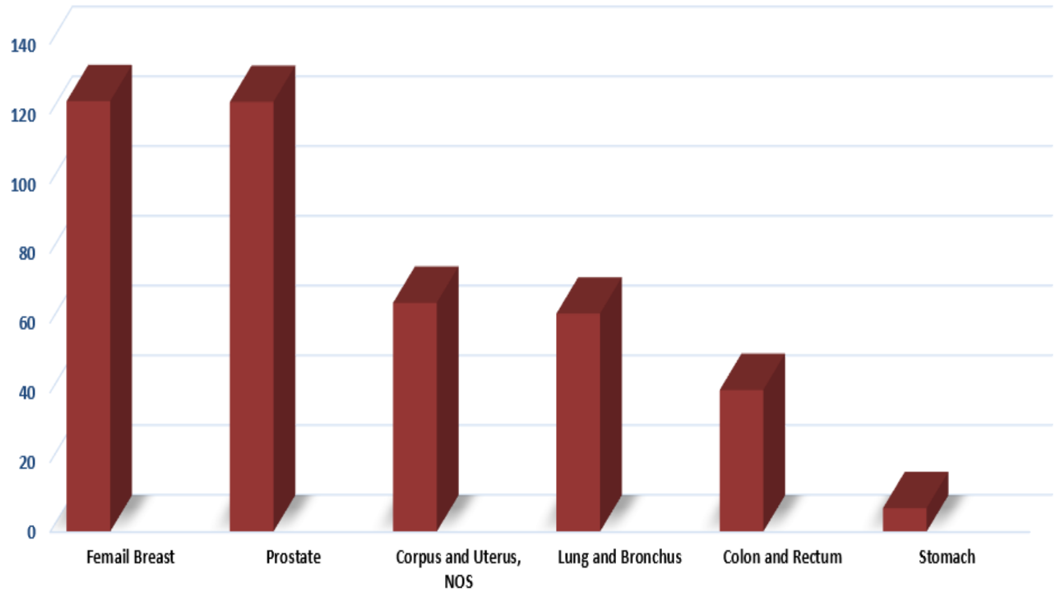

Figure 1.1: Age-adjusted invasive cancer incidence rates for the 6 major cancers in the United States: 2009-2013. The vertical axis units are expressed per 100,000 persons age adjusted to the 2000 US standard population [1]. Age-adjusted rate is a rate related to some events that became standard for a particular population. Therefore, this form of rate makes it possible to compare the rates for populations with different ages distribution in a fairly manner [1].

cancer, 78 of every 100 are living five years after diagnosis. Analysis of cancer survival data and its associated results is necessary to evaluate cancer treatment programs and to watch for unusual changes. Researchers and scientists all around the world are developing cancer treatment methods based on its general survivability rate, hoping the new treatments may lead to a better prognosis for patients. To this end, an accurate prediction of cancer survivability became very important. A reliable prediction of cancer survivability truly enables physicians to make well-organized systematic treatments, improving the quality of care that healthcare provides to the cancer patients. Also accurate prediction models help them to make more informed decisions to treat patients. For example, they may choose new medications or more aggressive therapies for patients with less hope of survival.

1.2 Literature Review

In this section we will discuss the current knowledge and progress so far in the computational methods developed for cancer survivability prediction. In the last few years, several machine learning approaches have been utilized for caner survivability prediction [10], [16], [27], [35]. While some of them employed only clinical and genomic data which is not widely available for the research community, the rest used publicly available datasets ***(e.g., SEER dataset)*** [6], [7] which included tens of thousands of records. A drawback of these datasets is that they do not usually contain genomic and clinical information specific to patients.

Generally speaking, machine learning is used to build prediction models by using training data or past experiences in a particular domain. Machine learning is basically applied to four major paradigms, such as classification, clustering, regression, and rule extraction [39]. Classification methods try to assign labels to new unlabeled data, while clustering algorithms divide data into different groups based on similarities or a set of structural measures [51]. Regression statistically approximates the underlying associations between variables and is used in modeling, while rule extraction algorithms are employed to explore propositional rules or relationships between attributes in the data. Support vector machines (SVM) [15], artificial neural networks (ANN) [26], naïve Bayes [36], logistic regression [28], decision trees [45] are the most popular machine learning algorithms which have been utilized for cancer survivability prediction. While In 2000, Zupan et al. [53] applied two well-know machine learning strategies including naïve Bayes and decision trees on a dataset including 1055 instances of prostate cancer patients to make survivability predictive models, and showed that both models were able to produce promising results using such a dataset, in 2005, Delen et al. [19] employed ANN, decision trees, and logistic regression to design a group of prediction models utilizing a large dataset including more than 200,000 cases. For the evaluation purposes, they used 10-fold cross-validation strategy to measure the performance of the three prediction models. Their experimental results indicated that the decision tree (C5) was able to offer the best predictive model with 93.6% accuracy on the holdout sample. ANN was the second with

91.2% accuracy and the logistic regression models came to be the worst of the three with 89.2% accuracy using the proposed dataset.

In 2005, Chang et al. [13] utilized decision trees and a data set to confirm that a wound-response gene expression signature was a powerful predictor of clinical outcome in different patients with early stage breast cancers. Together with their other obtained experimental results on advanced breast, gastric, and lung cancer, that discovery reinforced the concept that a gene expression program associated to the physiological response to a wound was frequently activated in common human epithelial tumors, increasing risk of cancer advancement and metastasis. In 2006, Jonsdottir et al. [30] developed a predictive model for breast cancer survivability and assessed 5 year outcome of an incidence of cancer using data mining algorithms. They utilized almost one hundred datasets containing a set of different features (e.g., age, an indicator if the tumor was found in a medical examination, an indicator if the nodes are palpateble or suspicious, pathologic primary tumor size, etc.) along with naïve Bayes, decision trees, logistic regression, and a meta algorithm which was able to combine results from other classification algorithms. While the AUC values for the proposed strategies ranged from 76% to 85%, one limitation was investigating only small number of available instances in the datasets. In 2006, Bellaachia et al. [11] examined the prediction of survivability rates of breast cancer using data mining algorithms. They applied naïve Bayes, back-propagated neural network, and C4.5 decision tree algorithms on a publicly available SEER dataset which included 151,886 instances. They found that C4.5 algorithm has had a better performance than two other data mining techniques. In 2006, Cruz et al. [17] extensively reviewed and compared the general performance of several machine learning algorithms that are being applied to different cancer prediction and prognosis, particularly identified a number of trends concerning the types of machine learning algorithms being employed, the types of training data being integrated, and the types of cancer diseases being well analyzed along with and the performance of these methods in predicting cancer survivability.

In 2016, Liang et al. [37] studied lung cancer survivability analysis using a semi-supervised machine learning algorithm including Cox proportional hazards regression model (Cox) [25] and accelerated failure time model (AFT) [40], and reported the proposed semi-supervised algorithms was more appropriate tool for survival analysis in clinical cancer research. In 2016, Kate et al. [33] used three different machine learning strategies to predict breast cancer survivability separately for every stage. They compared the proposed methods with the traditional joint models built for all the stages, evaluated the models separately for every stage and together for all the stages. The results obtained by their study showed that the most appropriate model to predict breast cancer survivability for a specific stage was the model trained for that particular stage. Readers interested in recent advances in building cancer survivability models are referred to [9], [20], [31], [38], [43], [46], [50] for other techniques.

1.3 Motivations and Objectives

While using a broad range of machine learning algorithms and training strategies have been well studied in the past, the use of different cancer stages either for training a prediction model or for model evaluation have been limited so far. Cancer stage refers to the extent of the cancer. Cancer incidences are basically assigned stages based on tumor size and/or the vastness of spread, accordingly survivability rates varies across the stages. There are several cancer staging systems in use. While such a system categorizes cancers to be in Stage 0, Stage I, ..., and Stage IV along with further subcategories, another system namely TNM (tumor, node, metastasis) tries to categorize cancers based on the status of tumor, node, and metastasis [22]. Since we are going to apply our proposed method on the SEER dataset, in this work we used the staging system of that data which includes four stages: in-situ, localized, regional and distant. Table 1.1 illustrates these stages in further detail [5].

For the most cancers, the survivability rates differ significantly according to the stage of the cancer. For instance, for the breast cancer, while the survivability rate for in-situ stage is 99.42%, it is 36.17% for distant stage, and survivability rate for all stages is 92.04%. We obtained these numbers from the subset of the SEER dataset which we

Table 1.1: Cancer stages in the SEER dataset.

Stage	Description
In-situ	Abnormal cells are existent but they have not spread to nearby tissue
Localized	Cancer is limited to the place where it started. There is no any sign that it has spread
Regional	Cancer has spread out to nearby lymph nodes, organs, or tissues
Distant	Cancer has spread out to distant parts of the body

used in this work. It is clear that the survivability prediction for in-situ stage is so much different from distant stage. The research study Kate et al. [33] found that the models trained with every stage separately (called **stage-specific** model) could often provide better survivability prediction than the model which is trained with all stages (called **joint model**). The work investigated the idea for only the breast cancer. In this study, we are going to investigate this idea on other cancers to determine whether the stage-specific models are able to provide better prediction than joint model or not. Additionally, we are interested in seeing whether the most important features to predict survivability are different for different stages.

Hence the objective of the proposed research is to:

- Examine and build survivability prediction models for five major cancers, trained on all stages (joint models) and separately for every stage (stage-specific models).
- Analyze whether the best model to predict the survivability of the cancer for one specific stage is the model which is specifically built for that stage.
- Investigate the importance order of features to predict survivability for every stage of cancer.
- Find reasons for the differences between stage-specific and joint models.
- Compare the performance of different machine learning methods for building cancer survivability predictive models.

1.4 The Main Contributions

Our goal is to utilize machine learning algorithms to design and develop stage-specific survivability predictive models for five different cancers: lung & bronchus, breast, colon, corpus uteri, and stomach cancers. We first selected the five most common cancers with no missing feature values from the SEER dataset. We dropped prostate cancer and included the next most common cancer, because of prostate cancer has a very high survivability rate which makes predicting survivability almost trivial and because it had very few incidences available for the distant stage. We compared cancer survivability prediction models trained on all stages (joint model) and trained separately on every stage (stage-specific model). We illustrate which features are most indicative of survivability across different stages. We show performance differences when the models are evaluated separately for different stages. We also compare the performance of different machine learning techniques on building predictive models for cancer survivability on different cancer stages.

Chapter 2

Materials and Methods

In this section, we explain the different computational approaches involved in our proposed stage-specific predictive models for cancer survivability. We shall begin with the dataset and machine learning algorithms which we used, then, we will explain our approach.

2.1 Dataset

For the current research study, we used SEER Cancer Dataset [6], [7] which is a publicly and freely available dataset collected as part of National Cancer Institute's Surveillance, Epidemiology, and End Results program. It is a dataset of cancer incidences in the United States, which is collected from geographical areas which represent 28% of the US population (according to the US Census 2010), and is updated every year.

The available features for cancer data in this dataset are more general, and it does not include the genome and/or clinical information, but since it is widely available and it covers millions of cancer instances, it has been considered as a very valuable data source for cancer research.

The SEER dataset can be easily obtained through the Internet. To obtain the data, an applicant need to sign up a data use agreement letter available through the website. Once the user sends a signed agreement letter, they provide the user, access to the dataset. The latest version of SEER dataset which was used in the current thesis was released in April 2016 and included cancer incidences form 1973 to 2013 to a total of 9.18 million incidences.

2.2 Definitions

Following are the definitions of two important terms used in this work.

- **Survivability Rate**: According to the National Cancer Institute (NCI), the survivability rate of cancer indicates that, from 100 persons diagnosed with a particular type of cancer, how many of them live for more than n years. And usually the number of years to investigate the survival rate is 5 years (n=5).
- **Survived vs. Not Survived**: In respect to the survivability rate definition, if a patient with specific cancer, lives more than 5 years after his/ her diagnosis, then the patient can be considered as survived. To consider a patient who is diagnosed with a specific cancer as not survived, this patient should have died within 5 years of the diagnosed date and the cause of death of that person must be the same cancer with which he/she was diagnosed. So, if a patient was diagnosed with a specific cancer and died after 2 years because of any other reason, this patient could not be considered either as survived or not survived. Therefore, the information of such patients would not be used for the survivability cancer research.

2.3 Machine Learning Methods

In the following sections, we first discuss the concept of information gain, then we will briefly review as set of classification algorithms including logistic regression, naïve Bayes, decision tree, and cost-sensitive classifiers. Finally, we will explain our proposed approach to build cancer survivability predictive models.

2.3.1 Information Gain

Information gain [39] is a fundamental concept in machine learning which can be used to see order of importance of the features. While the information gain is specifically used in decision tree algorithm, it can also work as a feature selection tool for other methods.

Information gain statistic for a feature indicates its importance in the prediction of the class.

2.3.2 Logistic Regression

Linear regression is one the machine learning methods that is used to model continuous value functions. A popular type of generalized linear regression is called logistic regression which models the probability of the variable being predicted as a linear function of a group of predictor variables. The logistic regression is used for binary classification when the output variable of a model is specified as a categorical binary [32].

2.3.3 Naïve Bayes

Naïve Bayes is a classification method based on probability theory. In order to estimate joint probability distribution of the features and output, it makes a naïve assumption that all the features are conditionally independent of each other given the output. Along with this assumption it uses Bayes theorem to compute probability of the output given the features in terms of the probability of the features given the output which is easier to estimate using the training data. Naïve Bayes is computationally a very fast machine learning method [24].

2.3.4 Decision Tree

Decision tree is among one the most popular classification models in machine learning. Decision tree is a rule based method in which a tree structure is learned from the training data where each node represents a test on a feature value, and each branch denotes a result of the test and the terminal shows the classes. The ID3 and C4.5 are among widely used decision tree algorithms in machine learning community [32].

2.3.5 Cost-Sensitive Classification

For an unbalanced data, a machine learning method can classify most of the instances in the majority class to maximize the accuracy, without doing anything associated to the minority class, which can cause misleading results. In such dataset to maximize the accuracy in both majority and minority classes, we can use cost sensitive classifier which penalizes misclassifying the minority class [33] according to a user-specified weight.

This algorithm has a cost matrix which lets the users or developers assign some penalty for miss classifying classes. The algorithm does not determine which cost works best, therefore, the investigator has to find the best cost to assign empirically which is typically done through internal cross-validation within the training data. The cost sensitive algorithm employs a classifier (e.g., decision tree, logistic regression, etc.) as its internal classification algorithm.

2.3.6 Probability or Confidence

All machine learning algorithms we used, give us a probability or confidence for each given instance which indicate how much they are confident about putting that instance in one specific class. These confidences are used to plot ROC Curve [29]. We will talk about ROC later in the next section.

2.3.7 Model Evaluation

There are several evaluation measures, including accuracy, f-measure, precision, recall, sensitivity, specificity, AUC, etc. which can be used to evaluate prediction models. There are advantages and disadvantages of using them, but in general, most of them can not indicate the overall model performance. For example, the accuracy only checks the correct classification on test data which could be misleading. Let's consider a scenario where we have 95% of data belonging to one majority class, if a classifier just classifies all the data in this class, then the final accuracy would be 95% without doing anything regarding the minority class and this misclassification will not be fairly represented in the accuracy of the model. However, the AUC (Area Under the ROC Curve) metric can fairly reflect the performance of the model even in such situations.

ROC (Receiver Operating Characteristic curve), is a curve between a models true positive rate (or sensitivity which is the fraction of positive instances which model classified correctly as positive) versus false positive rate (or 1-specificity which is the fraction of negative instances misclassified as positive by model). For every decision threshold for the confidence of the model in classifying instances we obtain a true positive rate and a false positive rate. Thus by varying the decision threshold from 0 to 1, one can obtain an entire range of true positive rates and false positive rates which when plotted on a graph is called an ROC curve. One of the noticeable property of ROC curve is that it is independent of the class distribution. It means that, if the distribution of positive and negative instances changes in the dataset, its value does not change [44].

Area under this curve (AUC) is then used to indicate performance with a single number. AUC is a very popular metric to evaluate the performance of classifiers. We used AUC as a measure to evaluate and compare the performance of the models. It can be helpful to indicate that the value of AUC ranges from 0.5 to 1. A random classifier has a 0.5 AUC and the perfect classifier has 1 AUC. A higher AUC shows better performance for a classifier.

10-Fold cross validation

In this evaluation methodology the available data is randomly divided into k equal size folds, and each time the model is trained with k-1 folds and tested remaining fold, and this process is repeated for k times each time using a different fold for testing. The final performance is reported by taking average of the k metrics obtained from the k folds. k=10 is the standard and the most common value used for k.

2.4 Methodology

In this section, we will explain our methodology to build cancer survivability predictive models.

2.4.1 Data Preparation

Our data preparation had two different stages as follows:

The SEER Dataset Subset Selection

The **first step** in the data preparation was to pick up a subset of the SEER dataset for the proposed study. We needed to select a subset related to a specific cancer in a specific period of time. In doing so, the need was to choose a group of patients who have diagnosed by a specific cancer in a desired period of time. The “Primary Site” and “Year of diagnosis” features in the SEER helped to do pick such records. For investigating 5 years survivability, we needed at least, 5 more years data ahead of every diagnosed records to check the patient vital status. So, since the last data available in SEER covers up to 2013, we limited our instances to the instances having diagnosed date in years 2004 to 2008. We did not select the data before 2004, because many SEER fields where changed in 2004, and in addition cancer rates changed over time (some of the features which are used in this study, just have value for the instances after 2004 in the SEER).

Cancer Patients Tagging

The **second step** here was to tag patients as survived or not survived, or ignore the instance (while the cause of death of a patient is different from the cancer of the study or the patient died after 5 years of living with the cancer). In doing so, we utilized three different SEER dataset features including “vital status recode”, “survival months” and “cause of death” to develop the following rule which assist us to tag the patients:

```
if ("survival months">=60 and "vital status recode"="alive" )
  {
    tag the patient as "survived";
   }
 else if ("survival months"<60 and "cause of death"="the cancer of study")
  {
    tag the patient as "not survived";
    }
 else ignore this patient instance;
```

For those patients who have been diagnosed of cancer more than once, we picked up their last diagnosis to have them in the data subset. “patient Id number”, “month of diagnosis”, and “year of diagnosis” or “SEER record number” are the features in the dataset that we used to determine the last diagnosis in case there were more than one.

2.4.2 Features

For building a cancer survivability predictive models, it is crucial to use informative features. We used 18 features of the SEER dataset to build survivability predictive models. We picked these features because they have been used in previous works in the literature [8], [11], [18], [33], [41].

The features which should be used are illustrated in Table 2.1. We needed the instances with valid feature values and also be categorizable as either survived or not survived, and with diagnosis date between 2004 to 2008. In doing so, we checked out the entire SEER dataset using such criteria and eventually the following cancers received the largest valid records respectively: Breast, Colon, Lung & Bronchus, Prostate, Corpus Uteri and Stomach. Among those cancer diseases, prostate has had a very high survival rate (98.9%) which makes building survivability prediction models for that cancer not very useful, therefore, this study focused on the rest of them. There were enough instances for breast and colon cancers even after excluding ones with missing features, but for the rest, we included the instances with missing features.

Table 2.1: The SEER dataset features used to make a predictive model.

Feature name	**Data type**	**Description**
Marital Status at DX	Nominal	Marital status of the patient
Race/Ethnicity	Nominal	Race of the patient
Sex	Nominal	Gender of a patient
Age of Diagnosis	Numeric	Age of the patient at diagnosis
Primary Site	Nominal	It specifies the site in which the primary tumor has emerged (this feature is also can be used to categorize the cancer types)
Histologic Type ICD-O-3	Nominal	It describes the microscopic form of the primary tumor
Behavior Code ICD-O-3	Nominal	It can describe the primary tumor as either benign or malignant, noninvasive or invasive
Grade	Nominal	It is for categorizing the shape of the tumor and its speed of spread
Reginal Nodes Positives	Numeric	The exact number of examined regional lymph nodes, containing metastases
Reginal Nodes Examined	Numeric	The exact number of regional lymph nodes which were removed and examined
CS Tumor Size	Numeric	Indicating size of Tumor in mm
CS Extension	Nominal	Regarding the extension of the tumor
CS Lymph Nodes	Nominal	Describes how the lymph nodes are involved
CS Mets at DX	Nominal	Provide information about metastasis at distant
RX Summ-Surg Prim Site	Nominal	Describes the surgical routine which is used to remove or/and destroy involved tissue of the primary site in the first step of treatment
RX Summ-Radiation	Nominal	Shows which radiation therapy approach has used in the first step of primary site treatment
Summary stage 2000 (1998+)	Nominal	There are now 4, In-situ, Localized, Regional, Distant stages, showing the spread of the cancer
Sequence Number-Central	Nominal	Indicates the sequential number of all in situ, benign, malignant and borderline primary tumors that can be reported during the living time of a patient

We also excluded male patient instances for breast cancer because they contribute only 1% of all the incidences [47]. Corpus uteri also is a cancer related to female organs. Therefor, “gender” feature was not used in building breast cancer and corpus uteri models. “Sequence number-central” feature also was used just for building the lung & bronchus, colon and corpus uteri. We will discuss about its reason later.

Tables 2.2 to 2.6 show stage-wise survivability distribution for five different cancers obtained from the SEER dataset. One can see that the survivability rate in In-situ stage is high and the number of records in this stage are few. Therefore, like the previous research study [33], we excluded this stage from all of our investigations.

By using the information gain statistic, we recognized the importance of the “Reference Number-Central” feature for lung & bronchus, colon and corpus uteri cancers . This feature was used previously to build lung cancer survivability model [8]. Therefore, for these cancers we used this feature, but for the rest, we did not use this feature to make predictive models.

2.4.3 Predictive Models

In every experiment in this study, we utilized 10-fold cross validation. For each cancer, we randomly shuffled the order of available instances and divided the data in 10 equal folds, such that the distribution of the survived and not survived and also every stage in the 10 folds were equal (stratified). For the stage-specific models, we used the same folds, just by keeping the data regarding every stage and ignoring the rest of the records. We used this strategy to be able to perform a fair and meaningful comparison between models. By looking at the records in the tables 2.2 to 2.6, it is clear that we were dealing with unbalanced data in all 5 cancers. One can see how the number of survived and not survived are different from each other at every stage and/or in all stages together. For classifying such unbalanced data, we used cost-sensitive classifier by considering its ability to apply penalty for any misclassification. To know which cost we should use to make a model for every fold, we assigned different costs (0.25, 0.5, 1, 2, 4, 6, ..., 18, 20) for misclassification of minority class (either survived or not survived). And for every cost, we did 5 internal cost cross validation through the available training data and eventually picked the cost which gave the best AUC average. We used this cost for building the model for its fold. We employed naïve Bayes, logistic regression, and decision tree (AD3) machine learning algorithms as an internal classifier for the cost-sensitive algorithm. Therefore, for every cancer, we first built models trained with all available data for that cancer by doing 10-fold cross validation and three different machine learning algorithms, and called these models as joint models. Then, we employed the same folds by keeping only the instances of one stage every time and doing 10-fold cross validation using the same three machine learning methods to build the stage-specific models. For expressing the performance of every model, we utilized the AUC average obtained through the 10-fold cross validation.

To compare the models, we reevaluated the joint models for every stage separately. We tested the joint models by 10 folds, including those instances that only belong to one specific stage at a time, and recorded the average of obtained AUC as the joint model performance for that specific stage. We did this process for all 3 stages. For stage-specific models, besides obtaining results for their each individual stages we obtained their results of each stages to also obtain results on all stages together as was done in [33]. We named these results (which are obtained from combination of every three stage-specific models) as combined model results.

2.4.4 Test Bed and Experimental Setup

For all experimental results presented in this section, we used 64-bit Windows 8 operating system on a PC with 2.20 GHz Intel Dual core CPU, 4MB cache and 8GB of RAM. All parts of the system were developed by Java2SE 8 and Weka data mining library (version 3.6.13) [23] which has been freely available to the research community.

Table 2.2: The **breast cancer** stage-wise survivability distribution obtained from the SEER dataset. We utilized these data records to build and also analyze predictive cancer survivability models.

	Total incidences	**Survived**	**Not survived**	**Percent survived**
All stages	174518	160623	13892	92.04%
In-situ	10106(5.79%)	10047	59	99.42%
Localized	106390(60.96%)	102737	3653	96.57%
Regional	55340(31.71%)	46872	8468	84.69%
Distant	2682(1.54%)	970	1712	36.17%

Table 2.3: The **lung & bronchus cancer** stage-wise survivability distribution obtained from the SEER dataset. We utilized these data records to build and also analyze predictive cancer survivability models.

	Total incidences	**Survived**	**Not survived**	**Percent survived**
All stages	183033	24358	158675	13.31%
In-situ	80(0.04%)	26	54	32.50%
Localized	27809(15.19%)	13190	14619	47.43%
Regional	42256(23.09%)	8414	33842	19.91%
Distant	112888(61.68%)	2728	110160	2.42%

Table 2.4: The **colon cancer** stage-wise survivability distribution obtained from the SEER dataset. We utilized these data records to build and also analyze predictive cancer survivability models.

	Total incidences	**Survived**	**Not survived**	**Percent survived**
All stages	61858	40200	21658	68.99%
In-situ	256(0.41%)	247	9	96.48%
Localized	21307(34.45%)	19036	2271	89.34%
Regional	29050(46.96%)	19452	9598	66.96%
Distant	11245(18.18%)	1465	9780	13.03%

Table 2.5: The **corpus uteri cancer** stage-wise survivability distribution obtained from the SEER dataset. We utilized these data records to build and also analyze predictive cancer survivability models.

	Total incidences	**Survived**	**Not survived**	**Percent survived**
All stages	36820	33731	3089	91.61%
In-situ	514 (1.40%)	512	2	99.61%
Localized	28127 (76.39%)	27381	746	97.35%
Regional	6616 (17.97%)	5373	1243	81.21%
Distant	1563 (4.24%)	465	1098	29.75%

Table 2.6: The **stomach cancer** stage-wise survivability distribution obtained from the SEER dataset. We utilized these data records to build and also analyze predictive cancer survivability models.

	Total incidences	**Survived**	**Not survived**	**Percent survived**
All stages	18286	5283	13003	28.89%
In-situ	179 (0.98%)	140	39	78.21%
Localized	4945 (27.04%)	3245	1700	65.62%
Regional	5962 (32.61%)	1645	4317	27.59%
Distant	7200 (39.37%)	253	6947	3.51%

Chapter 3

Results and Discussion

To analyze our proposed predictive cancer survivability models, several experiments were performed. The joint versus stage-specific predictive models analysis results are shown in Section 3.1. In Section 3.2, we measured and analyzed the order of importance of the features. Finally, in Section 3.3 we further compared the performance of three machine learning methods to make survivability predictive models.

3.1 Joint versus stage-specific predictive models

The AUC average obtained from the models trained with the all cancer stages (joint models) and also trained with each stage separately (stage-specific models), are shown in the Tables 3.1 to 3.5.

Where a model performance is statistically significantly better (with p value less than 0.05) compared to its corresponding paired value (the value at the same row and for the same machine learning technique), we denoted it in bold in the table. Two-tailed paired t-test was used for statistical significance testing.

Table 3.1: AUC obtained from joint models and stage-specific models for **breast cancer**. Bold numbers are statistically significantly better (p<0.05; two-tailed paired t-test) compared to the corresponding value at the same row and for the same machine learning algorithm in the table.

	Naïve Bayes		Logistic Regression		Decision Tree	
	joint	stage-specific	joint	stage-specific	joint	stage-spacific
all-stages	0.828	**0.843**	0.846	0.847	**0.846**	0.838
Localized	0.758	**0.768**	0.769	**0.774**	0.773	**0.779**
Regional	0.759	**0.778**	0.792	0.789	0.781	**0.788**
Distant	0.654	**0.707**	0.700	0.714	0.664	**0.710**

Table 3.2: AUC obtained from joint models and stage-specific models for **lung & bronchus cancer**. Bold numbers are statistically significantly better ($p<0.05$; two-tailed paired t-test) compared to the corresponding value at the same row and for the same machine learning algorithm in the table.

	Naïve Bayes		Logistic Regression		Decision Tree	
	joint	stage-specific	joint	stage-specific	joint	stage-spacific
all-stages	0.907	**0.911**	0.924	0.925	0.918	0.909
Localized	0.845	**0.856**	0.865	0.866	0.840	**0.861**
Regional	0.789	**0.806**	0.835	0.836	0.816	**0.830**
Distant	0.764	**0.821**	0.822	0.826	0.803	**0.819**

Table 3.3: AUC obtained from joint models and stage-specific models for **colon cancer**. Bold numbers are statistically significantly better ($p<0.05$; two-tailed paired t-test) compared to the corresponding value at the same row and for the same machine learning algorithm in the table.

	Naïve Bayes		Logistic Regression		Decision Tree	
	joint	stage-specific	joint	stage-specific	joint	stage-spacific
all-stages	0.853	0.862	0.861	0.868	**0.864**	0.838
Localized	0.713	**0.750**	0.741	**0.764**	0.751	**0.755**
Regional	0.736	**0.758**	0.762	**0.778**	0.739	**0.756**
Distant	0.730	**0.772**	0.754	**0.788**	0.752	**0.778**

Table 3.4: AUC obtained from joint models and stage-specific models for **corpus uteri cancer**. Bold numbers are statistically significantly better ($p<0.05$; two-tailed paired t-test) compared to the corresponding value at the same row and for the same machine learning algorithm in the table.

	Naïve Bayes		Logistic Regression		Decision Tree	
	joint	stage-specific	joint	stage-specific	joint	stage-spacific
all-stages	0.912	0.874	**0.909**	0.891	**0.910**	0.848
Localized	0.824	**0.835**	0.818	0.812	0.815	0.814
Regional	0.809	**0.818**	0.818	0.813	0.787	0.794
Distant	0.762	**0.821**	0.792	0.775	0.784	**0.805**

Table 3.5: AUC obtained from joint models and stage-specific models for **stomach cancer**. Bold numbers are statistically significant ($p<0.05$; two-tailed paired t-test) to the corresponding value at the same row and for the same machine learning algorithm in the table.

	Naïve Bayes		Logistic Regression		Decision Tree	
	joint	stage-specific	joint	stage-specific	joint	stage-spacific
all-stages	0.904	0.852	**0.919**	0.791	**0.927**	0.778
Localized	0.855	0.854	0.863	**0.873**	0.894	**0.902**
Regional	0.760	**0.797**	0.800	**0.805**	0.788	**0.814**
Distant	0.772	**0.831**	0.818	0.836	0.844	0.849

From the above five tables, it can be seen that:

- For all the stages, the stage-specific naïve Bayes models were always statistically significantly better than the joint models (except for stomach localized).
- For breast, colon and lung & bronchus cancers, the stage-specific decision tree models were statistically significantly better than the joint models.
- The joint models did not perform statistically significantly better than the stage-specific models for any stage (in the all last three rows).
- The performance of most models are generally worse on the distant stage. This is most likely because there was less data available for that stage for training the models. This indicates that if more data is collected for distance stage it may improve performance. We would not have made this observation if we had not evaluated results separately for every stage.
- The performance when evaluated on all the stages together is always better than the performance when evaluated on each stage separately whether for joint models or stage-specific models. This is counter-intuitive because one would normally expect performance on all stages to be around average of the performance on individual stages. But as was pointed out in [33], the different survivability rates of different stages lead to an over-estimation of performance when evaluated on all stages together. For example, by simply calling all localized instances to survive and others to not survive one can get a reasonably good performance when evaluated on all the stages together. Thus by simply biasing instances in early stages to more likely survive and latter stages to less likely survive one can artificially boost the performance of evaluation on all stages together [33]. A machine learning model can easily learn such a bias from data. Thus evaluating a model on all stages together overestimates its performance and is misleading. For example, boosting up the confidence of survivability prediction of the patients in localized cancer stage, because of the other patients are in the distant stage (with lower survivability rate) is not wrong, but it cannot add any meaningful information to the survivability

prediction of the instances in the localized stage. Hence to get a fair idea about a models performance it is best to evaluate it separately for every stage [33] and not on all stages together as was done in most of the previous work.

- As we discussed earlier (Section 2.4.3), the number of training data for joint models is significantly larger than the stage-specific models, and by building the stage-specific models with less training data, we can have faster training times for those machine learning algorithms which have higher than linear training time complexity.
- All above results were according to the thoughts and results which was discussed in the recent work performed for breast cancer [33], and we can see they are still correct for 4 more additional cancers mentioned in this study.

3.2 Features importance order in different models

To investigate the differences between the joint and the stage-specific predictive models, we checked the order of used features according to their importance in predicting the class. This was measured in terms of the information gain statistic. The results obtained for this experiment are shown in the Figures 3.1 to 3.5.

From all these figures (Figures 3.1 to 3.5), we can conclude the following facts:

- All the figures, clearly show that, every stage has different order of important features and this fact approves the idea of building stage-specific predictive models for cancer survivability instead of joint predictive models.
- The importance order of used features differ from one cancer to another cancer.
- For breast cancer, “tumor size” is between the top three features in all related figures which indicates its high contribution to predict the survivability in this cancer. For lung & bronchus and stomach cancers, “site-specific surgery” is in between the top two features. This emphasizes the importance role of this feature to predict the chance of survivability in these cancers. It is also shown that how survivability predictive models can help the physicians to decide the best site for surgery. Also

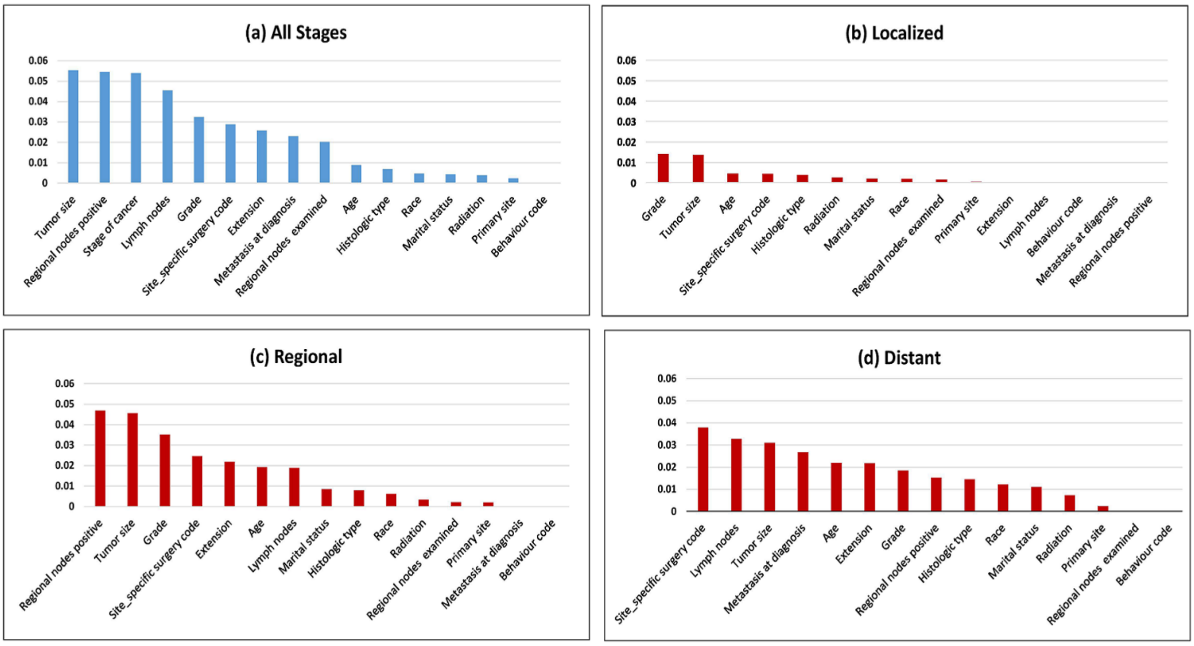

Figure 3.1: Information gain statistic for all the used features for **breast cancer** predicting survivability on all stages together and on the different stages separately.

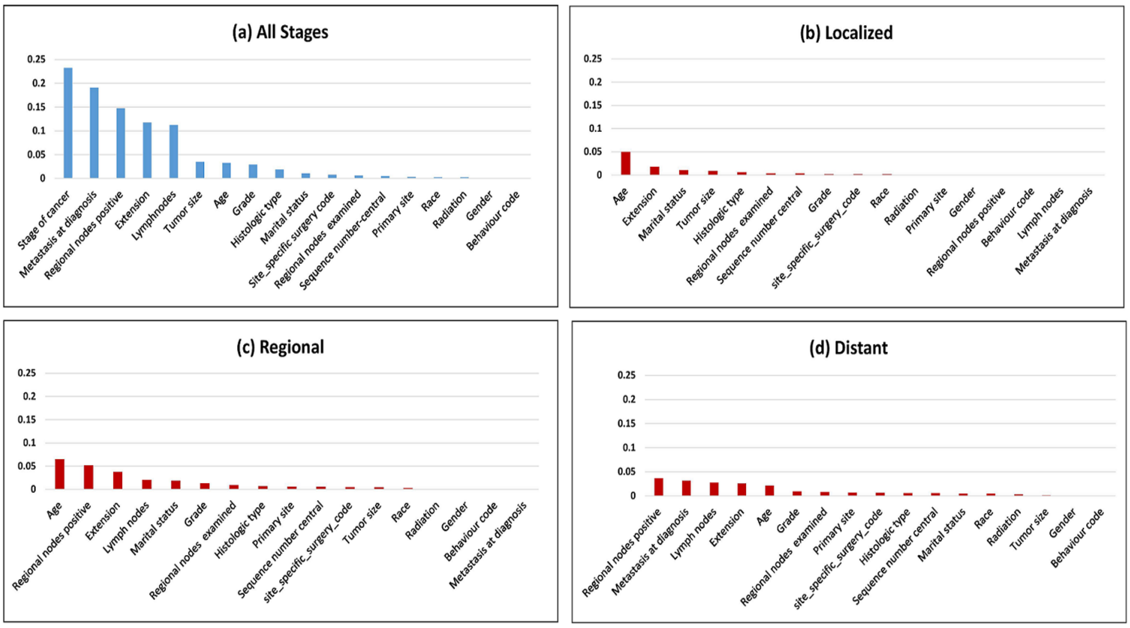

Figure 3.2: Information gain statistic for all the used features for **colon cancer** predicting survivability on all stages together and on the different stages separately.

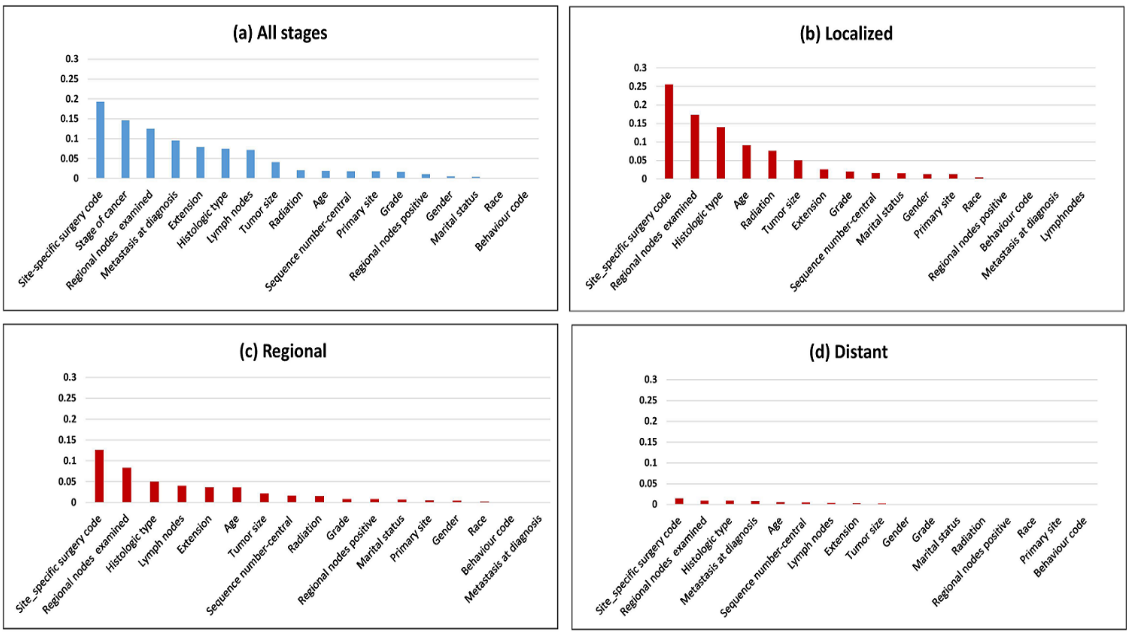

Figure 3.3: Information gain statistic for all the used features for **lung & bronchus cancer** predicting survivability on all stages together and on the different stages separately.

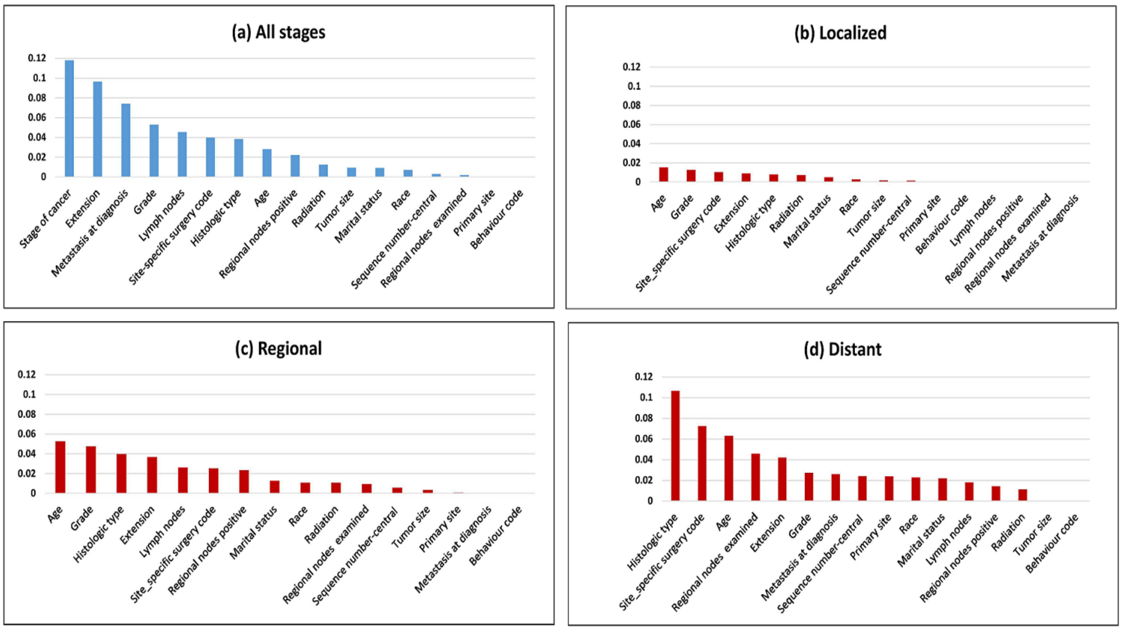

Figure 3.4: Information gain statistic for all the used features for **corpus uteri cancer** predicting survivability on all stages together and on the different stages separately.

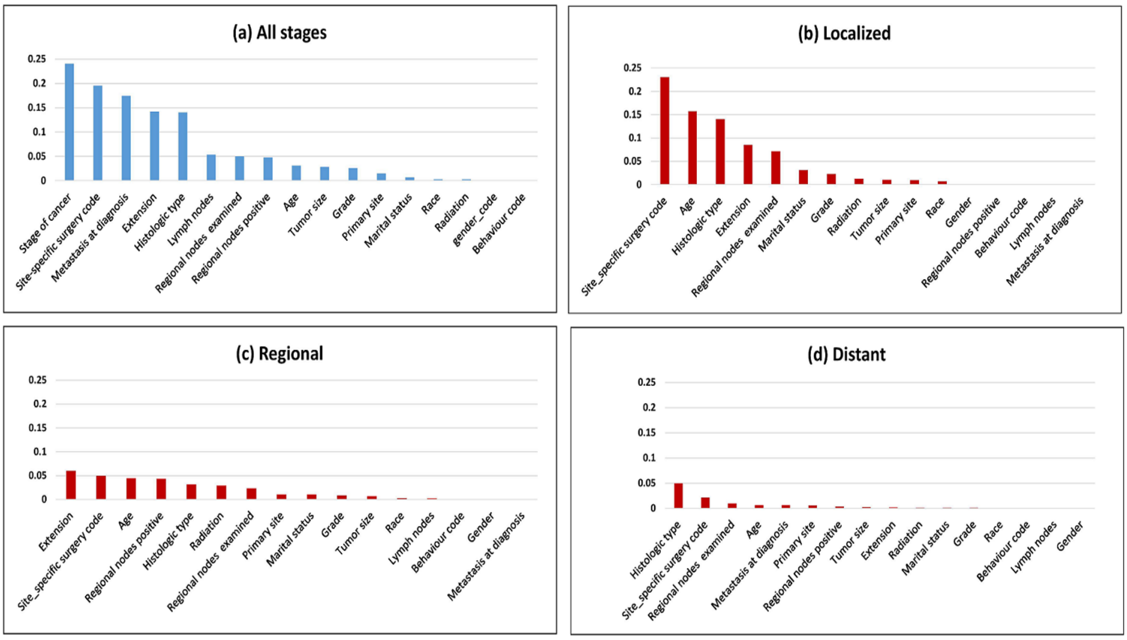

Figure 3.5: Information gain statistic for all the used features for **stomach cancer** predicting survivability on all stages together and on the different stages separately.

"age" can be seen as an important factor in these cancers.

- Metastasis at diagnosis was ranked 2-7 in all five cancers on distant stage (ranked 4, 2, 4, 7, 5 for breast, colon, lung & bronchus, corpus uteri, and stomach cancers respectively) and it shows its importance to predict the survivability of the patients having cancer in distant stage which is compatible with clinical sense of metastasis.
- There are other important facts which can be seen in these figures, but it must be investigated by experts in cancer treatment, so they can interpret the figures from clinical side appropriately.

3.3 Comparative study of machine learning methods

As it can been seen in the Tables 3.1 to 3.5, we have tried 3 popular machine learning algorithms to build joint and stage-specific survivability models. As we discussed earlier, these algorithms are the most popular techniques to build survivability models. We compared the performance of stage-specific models, created with these three algorithms to figure out which algorithm gives a model with better performance.

Table 3.6: Algorithms comparison result for **breast cancer** survivability predictive models.

	Comparison result
Localized	AD3 >Logistic regression >Naïve Bayes
Regional	Logistic regression >Naïve Bayes, AD3 >Naïve Bayese
Distant	No statistically significant differences found between used algorithms

Table 3.7: Algorithms comparison result for **colon cancer** survivability predictive models.

	Comparison result
Localized	Logistic regression >AD3 >Naïve Bayes
Regional	Logistic regression >Naïve Bayes, Logistic regression >AD3
Distant	Logistic regression >Naïve Bayes

Table 3.8: Algorithms comparison result for **lung & bronchus cancer** survivability predictive models.

	Comparison result
Localized	Logistic regression >AD3 >Naïve Bayes
Regional	Logistic regression >AD3 >Naïve Bayes
Distant	Logistic regression >AD3

In the Tables 3.6 to 3.10, one can see the results of these comparisons. For some models, we did not find any statistically significant difference (p value less than 0.05). We considered two-tailed paired t-test to compare the corresponding values (values at the same row but obtained from different machine learning techniques). The results are denoted by using the "greater than" symbol (>). Therefore, where the algorithm is statistically significantly better than the other algorithm, we expressed it greater than the other.

From the Tables 3.6 to 3.10, it can be seen that logistic regression and AD3 generally do better performance on localized and regional stages in building survivability prediction models than naïve Bayes for cancers in the study (except for corpus uteri cancer). In general, there is not much difference between them on distant stage instances.

Table 3.9: Algorithms comparison result for **corpus uteri cancer** survivability predictive models.

	Comparison result
Localized	Naïve Bayes >AD3, Naïve Bayes >Logistic regression
Regional	Naïve Bayes >AD3, Logistic regression >AD3
Distant	Naïve Bayes >Logistic regression

Table 3.10: Algorithms comparison result for **stomach cancer** survivability predictive models.

	Comparison result
Localized	AD3 >Logistic regression >Naïve Bayes
Regional	AD3 >Logistic regression >Naïve Bayes
Distant	No statistically significant differences found between used algorithms

Chapter 4

Conclusion and Outlook

In past, only joint survival prediction models were built by training machine learning methods on all the stages together. They were also evaluated together on all the stages. In [33] stage-specific models were built for breast cancer survivability and it was reported that joint models offer no advantage over stage-specific models for predicting breast cancer survivability and are often worse. It was also reported that evaluating on all stages together, as was always done in past, leads to an overestimation of performance.

In this study, we investigated whether the above observations made about breast cancer in [33] also generalizes to other cancers. To this end, we built joint and stage-specific survivability predictive models for five cancers - breast, colon, lung & bronchus, corpus uteri, and stomach. We used SEER dataset along with three machine learning algorithms - naïve Bayes, logistic regression, and decision tree to make the survivability predictive models. According to the obtained results, joint models do not provide better performance than stage-specific models, and in most cases, their performance are statistically significantly worse than the stage-specific models. We also saw that the order of importance of features, as determined using information gain statistic, is different for each stage which further supports building separate models for every stage instead of building joint models. Based on these results, we recommend building separate survivability predictive models for separate stages. We also found that the evaluation of models on all stages together is misleading, since it tends to overestimate the performance. Therefore, to see the real performance of a model, we recommend evaluating its performance for each stage separately. Hence we found that the observations reported in [33] for breast cancer also generalize to other cancers.

We also saw that there is no specific pattern to suggest using a specific algorithm to make accurate predictive models for cancer survivability, but all the algorithms used here can give a reasonable performance. However we note that naïve Bayes requires much less computational time and the models generated with logistic regression and AD3 were better for most cancers. We also observed that the performance of most models was generally worse on the distant stage. In future, more training data of that stage may help improve the performance.

Bibliography

[1] Centers for disease control and prevention-united states cancer statistics (uscs). 2016. https://nccd.cdc.gov/uscs/toptencancers.aspx.

[2] Medical news today. 2016. http://www.medicalnewstoday.com/articles/282929.php.

[3] National cancer institute. 2016. https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[4] National cancer institute-seer stat fact sheets. 2016. http://seer.cancer.gov/statfacts/html/urinb.html.

[5] National cancer institute-seer stat fact sheets. 2016. https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[6] Seer data. 2016. http://seer.cancer.gov/data/.

[7] *Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission.* 2016.

[8] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary. Lung cancer survival prediction using ensemble data mining on seer data. *Scientific Programming*, 20(1):29–42, 2012.

[9] E. A. Asare, L. Liu, K. R. Hess, E. J. Gordon, J. L. Paruch, B. Palis, A. R. Dahlke, R. McCabe, M. E. Cohen, D. P. Winchester, et al. Development of a model to predict breast cancer survival using data from the national cancer data base. *Surgery*, 159(2):495–502, 2016.

[10] D. M. Ashley, S. Gupta, T. Tran, L. Wei, P. K. Lorgelly, D. M. Thomas, S. B. Fox, and S. Venkatesh. Machine-learning prediction of cancer survival: A prospective study examining the impact of combining clinical and genomic data. In *ASCO Annual Meeting Proceedings*, volume 33, page 6521, 2015.

[11] A. Bellaachia and E. Guven. Predicting breast cancer survivability using data mining techniques. *Age*, 58(13):10–110, 2006.

[12] C. A. Bertelsen, A. U. Neuenschwander, J. E. Jansen, M. Wilhelmsen, A. Kirkegaard-Klitbo, J. R. Tenma, B. Bols, P. Ingeholm, L. A. Rasmussen, L. V. Jepsen, et al. Disease-free survival after complete mesocolic excision compared with conventional colon cancer surgery: a retrospective, population-based study. *The Lancet Oncology*, 16(2):161–168, 2015.

[13] H. Y. Chang, D. S. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sørlie, H. Dai, Y. D. He, L. J. van't Veer, H. Bartelink, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3738–3743, 2005.

[14] C.-P. Chen. Predicting medical expenditure and survivability of the lung cancer patient by bayesian network. 2016.

[15] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[16] J. A. Cruz and D. S. Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 2006.

[17] J. A. Cruz and D. S. Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 2006.

[18] D. Delen, G. Walker, and A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

[19] D. Delen, G. Walker, and A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

[20] D. Dooling, A. Kim, B. McAneny, and J. Webster. Personalized prognostic models for oncology: A machine learning approach. *arXiv preprint arXiv:1606.07369*, 2016.

[21] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, et al. Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(D1):D805–D811, 2015.

[22] L. Frederick, D. L. Page, I. D. Fleming, A. G. Fritz, C. M. Balch, D. G. Haller, M. Morrow, et al. *AJCC cancer staging manual*, volume 1. Springer Science & Business Media, 2002.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[24] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques.* Elsevier, 2011.

[25] F. E. Harrell Jr. Cox proportional hazards regression model. In *Regression Modeling Strategies*, pages 475–519. Springer, 2015.

[26] S. Haykin. Neural networks, a comprehensive foundation. 1994.

[27] M. Hilario, A. Kalousis, M. Müller, and C. Pellegrini. Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics*, 3(9):1716–1719, 2003.

[28] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression.* John Wiley & Sons, 2004.

[29] N. Japkowicz and M. Shah. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press, 2011.

[30] T. Jonsdottir, E. T. Hvannberg, H. Sigurdsson, and S. Sigurdsson. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. *Expert Systems with Applications*, 34(1):108–118, 2008.

[31] K.-W. Jung, Y.-J. Won, C.-M. Oh, H.-J. Kong, H. Cho, D. H. Lee, and K. H. Lee. Prediction of cancer incidence and mortality in korea, 2015. *Cancer Research and Treatment*, 47(2):142–148, 2015.

[32] M. Kantardzic. *Data mining: concepts, models, methods, and algorithms.* John Wiley & Sons, 2011.

[33] R. J. Kate and R. Nadig. Stage-specific predictive models for breast cancer survivability. *International Journal of Medical Informatics*, 97:304–311, 2017.

[34] K. D. Kochanek, S. L. Murphy, J. Xu, and B. Tejada-Vera. Deaths: final data for 2014. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 65(4):1, 2016.

[35] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[36] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[37] Y. Liang, H. Chai, X.-Y. Liu, Z.-B. Xu, H. Zhang, and K.-S. Leung. Cancer survival analysis using semi-supervised learning method based on cox and aft models with l 1/2 regularization. *BMC medical genomics*, 9(1):1, 2016.

[38] L. Macyszyn, H. Akbari, J. M. Pisapia, X. Da, M. Attiah, V. Pigrish, Y. Bi, S. Pal, R. V. Davuluri, L. Roccograndi, et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology*, 18(3):417–425, 2016.

[39] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine learning: An artificial intelligence approach.* Springer Science & Business Media, 2013.

[40] P. M. Odell, K. M. Anderson, and R. B. D'Agostino. Maximum likelihood estimation for interval-censored data using a weibull-based accelerated failure time model. *Biometrics*, pages 951–959, 1992.

[41] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin. Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, 26(9):2194–2205, 2013.

[42] D. N. Pearlman, M. A. Clark, W. Rakowski, and B. Ehrich. Screening for breast and cervical cancers: the importance of knowledge and perceived cancer survivability. *Women & health*, 28(4):93–112, 1999.

[43] M. D. Podolsky, A. A. Barchuk, V. I. Kuznetcov, N. F. Gusarova, V. S. Gaidukov, and S. A. Tarakanov. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention*, 17(2):835–838, 2016.

[44] F. Provost and T. Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.", 2013.

[45] J. R. Quinlan. *C4. 5: programs for machine learning.* Elsevier, 2014.

[46] A. Silva, T. Oliveira, V. Julian, J. Neves, and P. Novais. A mobile and evolving tool to predict colorectal cancer survivability. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 14–26. Springer, 2016.

[47] A. C. Society. American cancer society: Cancer facts and figures 2016, 2016.

[48] B. W. Stewart and C. Wild. World cancer report 2014. international agency for research on cancer. *World Health Organization*, 505, 2014.

[49] D. G. Tang, L. Li, D. P. Chopra, and A. T. Porter. Extended survivability of prostate cancer cells in the absence of trophic factors: increased proliferation, evasion of

apoptosis, and the role of apoptosis proteins. *Cancer research*, 58(15):3466–3479, 1998.

[50] L. Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[51] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[52] H. M. Zolbanin, D. Delen, and A. H. Zadeh. Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, 74:150–161, 2015.

[53] B. Zupan, J. DemšAr, M. W. Kattan, J. R. Beck, and I. Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75, 2000.