Development of a Predictive Model for Survival Time for Breast Invasive Carcinoma

Supraja Battula
battula3@wisc.edu

Lexi Luo
qluo38@wisc.edu

Emil Walleser
walleser@wisc.edu

1. Introduction

Breast cancer is one of the most common cancer found in women, accounting for about 250,000 new diagnosis each year [7]. One in eight women in the United Stated will develop breast cancer in their life time, with the most common one being Invasive breast cancer, which describes breast cancer spread to surrounding normal breast tissue with the possibility to metastasis to other organ tissue [5]. Although the mortality rate of such cancer have decreased with better treatments, breast invasive carcinoma still remains the second most common cause of cancer death in women, and many patients experience recurrent diseases.

With better prediction for survival time, both overall survival and diseases free survival, physicians may better allocate time for proper treatments, thus prolong the life of a patients. Therefore, the purpose of this project is to develop a predictive model for survival time for breast invasive carcinoma in hope to better the treatment decision and life expectancy for the patients.

1.1. Previous Study

Previous work from Berger et al. discovered five prognostic subgroups utilizing 16 important molecular characteristics in 2,579 TCGA gynecological (OV, UCEC, CESC, and UCS) and breast cancers and suggested a decision tree based on six clinically assessable criteria that classifies patients into the categories; however, no analysis was done examining the survival time [1].

Furthermore, study done by Pang el al. created a breast cancer survival prediction model based on The Cancer Genome Atlas (TCGA). The Pang group focused on the genome of such cancer and the prediction model is composed of seven prognostic biomarker genes [3].

1.2. Method and Planning

Data cleaning will be done first to exclude any incomplete data from the data set. We will then explore the data to better understand the structure of the data and features in the data. We plan to focus on mutation data, RPPA (protein expression level) data, as well as patients' demographic data to a predictive model for survival time. For model selection, we plan to experiment with different preprocessing options and models: KNN, Decision trees, Regression analysis, etc. to determine the optimal model, and cross-validation will also be done for verification.

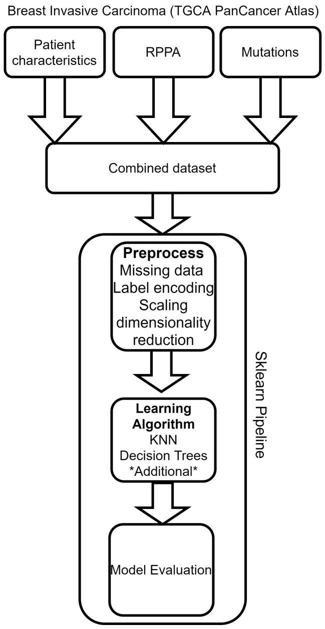

Figure 1. Example illustrating how to get BibTeX references from Google Scholar as a 1-column figure.

2. Motivation

2.1. Academic Purpose

Working on this project allows us to put what we have learnt in class into practice. We believe the topic we selected is impactful, and the aim of this project of developing a predictive model for survival time is achievable. We will investigate several methods both within and outside the classroom as we work to develop such predictive model. As a result, we may extend our machine learning understanding.

2.2. Greater Implication

Previous survival prediction model for invasive carcinoma have largely focused on genome of such cancer; however, due to post translational modifications and epigenetic changes, genes may not appropriately translate to functioning proteins. Therefore, for this study, we will focus on the RPPA protein expression data, hoping to develop a more reliable predictive model for survival time for breast invasive carcinoma patients. Certain protein may also become therapeutic targets for future treatments to improve survival time.

3. Evaluation

Our project would be considered successful if we could generate an accurate prediction model for Breast Invasive Carcinoma survival time. This model could be implemented either as a classification (ie. survival time greater than 1 year), regression (ie. survival time in months), or ordinal regression (ie. survival time 6 months, 12 months, 18 months). We will evaluate each of our prediction models(Decision Trees, KNN, and others) using cross-validation along with any feature preprocessing options such as (principal component analysis, Synthetic Minority Oversampling Technique). For classification, we will also evaluate different metrics such as false positives and false negatives [2][4]. These additional metrics will help us understand if our model is optimistically or pessimistically predicting survival times.

An important consideration for any model with any model being implemented in a healthcare setting is model interpretability [6]. Therefore we will also take into consideration if our selected model is a "black box" or a human can view the model and explain what features are associated with improved survival time.

If this model was successful we could apply similar methods to additional cancer datasets individually or evaluate several cancer types simultaneously.

4. Resources

We will be using a dataset from cBioPortal which provides various cancer genomics datasets. The dataset we are using was generated by The Cancer Genome Atlas (TCGA) PanCancer Atlas, a cancer genomics program that characterizes normal and cancer samples across many cancer types. The dataset has information about 1,084 patients with different breast invasive carcinomas. The dataset includes data regarding mutated genes, mutation count, RPPA(protein expression), overall survival status, diagnosis age, sex, race, and tumor type.

Computations will be done on our own personal computers, but if needed we will consider using external computer resource. We will be primarily be using Python through Jupyter Notebook and Scikit Learn.

5. Contributions

We will all equally contribute to all parts of the project. Each member in our group has various strong suits that will allow us to successfully complete our project. Lexi is very knowledgeable in cancer research, Emil has previous experience working with various machine learning models, and Supraja has experience coding using Python and statistical analysis. The code production will be split equally among the group. We will all help write the code and build the machine learning model. For writing the report, Lexi will be responsible for the introduction and conclusion, Emil will be responsible for methods of the model, and Supraja will be responsible for results. All in all, we will ensure that everyone does their part and no member does more than anyone else.

Clearly indicate what computational and writing tasks each member of your group will be participating in.

References

[1] A. C. Berger, A. Korkut, R. S. Kanchi, A. M. Hegde, W. Lenoir, W. Liu, Y. Liu, H. Fan, H. Shen, V. Ravikumar, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, 33(4):690–705, 2018.

[2] P. R. Harper. A review and comparison of classification algorithms for medical decision making. *Health Policy*, 71(3):315–331, 2005.

[3] L. Liu, Z. Chen, W. Shi, H. Liu, and W. Pang. Breast cancer survival prediction using seven prognostic biomarker genes. *Oncology letters*, 18(3):2907–2916, 2019.

[4] A.-M. Šimundić. Measures of diagnostic accuracy: basic definitions. *Ejifcc*, 19(4):203, 2009.

[5] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu. Risk factors and preventions of breast cancer. *International journal of biological sciences*, 13(11):1387, 2017.

[6] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083, 2020.

[7] E. J. Watkins. Overview of breast cancer. *Journal of the American Academy of PAs*, 32(10):13–17, 2019.