CS229 Project Final Report:
Pancreatic cancer prognosis using clinical and radiomic data

Arash Jamalian (arashj) *

June 10, 2020 †

1 Abstract

According to World Health Organization Globocan database [6], more than 56000 Americans developed pancreatic cancer in 2018. Unfortunately most patients are not diganosed in the earlier stages when the tumor can safely be resected. For patients that are in borderline resectable or locally advanced stages, prognosis can play an important role in identifying the risk factors and assessing treatment options. In this project, we run survival analysis on pancreatic cancer dataset provided by Stanford Cancer Center. We augment the dataset with Synthetic Minority Over-sampling Technique (SMOT) and train Cox proportional hazard model on clinical, radiomic and combination of both feature sets. We will study the model performance in grouping high and low risk patients and measure the prognastic value of clinical and radiomic features in predicting survival days. We present regularization and feature removal techniques for reducing the model variance and identify top features most contributing to hazard scores [1].

2 Introduction and Related Work

In this project, our goal is to run survival analysis on pancreatic cancer patients dataset and assess the prognastic value of clinical and radiomic features. Speicifcally we will develop regression models using clinical, radiometric and combination of both feature sets and evaluate the model performance in predicting survival days for each of these cases.

Survival analysis on cancer patients has been widely covered in the past few decades by applying variety of machine learning techniques including regression models, bayesian networks, decision trees, and deep neural networks. Most of the studies have been focusing on supervised learning, and make strong assumptions on the underlying relationship between covariates and the survival risk of the patients. The most widely used one is the Cox [3] proportional hazard assumption that require the hazard ratio between patients to remain constant over time.

On feature guided studies, [12] presents a deep learning framework for extracting radiomic features that are then input into a tree regression based model for making survival prediction. Two survival pipelines, one with CNN based segmentation and random forest survival prediction, and the other with FCN segmentation and XGBoost survival prediction are implemented and evaluated using 10-fold cross validation on BraTS training dataset.

Two studies looked at extending Cox proporional hazard model in a deep neural network framework [9] DeepSurv implements a deep neural network with fully connected hidden layers and a linear output layer predicting the log risk function. The network is trained by optimzing the log Cox partial likelihood function and L2 regularization. [2] Cox-nett neural netowork has one hidden layer with tanh activation function and uses Cox regression as output layer. This model is trained by optimizing the partial log likelihood function with a ridge regularization term. Overfitting is a potential problem with these models and both studies utilize dropout regularization on their hidden layer to address this issue.
DeepHit model [10] takes a different approach and looks at learning the survival probability distribution directly. This model does not make any proportional hazard assumption and allows for patients covariates to change over time. The model architecture also provide a framework for studying multiple events of interest each having their own sub neural network with a common base network.

3 Dataset and Features

MRI images from 80 patients are processed using Stanford Quantitative Image Feature Extraction (QFIP) pipeline to extract radiomic image features. Each patient case has 900 radiomic features including first order statistics (voxel intensities within region defined by mask), shape based (2D and 3D size and shape of region of interest), gray level co-occurrence matrix (co-occurring pixel gray values over the image), gray level zone matrix (describing describing the count of zones at different gray levels), and gray level run length matrix (describing number of runs with different gray levels). Dataset also includes clinical data on patient's gender, KPS, date of diagnosis, last follow update, whether patient survived as well as tumor characteristics such as resectability criteria and stage. Date of birth, race, and ethnicity of patients was missing and we could not source them in time for this project.

*Department of Computer Science, Stanford University, arash.jamalian@gmail.com
†Please keep the report private for possible inclusion in a future biomedical journal publication.
[1]Code for the project uploaded to gradescope. No public access.

Figure 1 visualizes survival days and tumor's greatest dimension for different cancer stages. Survival days were calculated from date of diagnosis to date of death or last follow up date (for right censored cases).

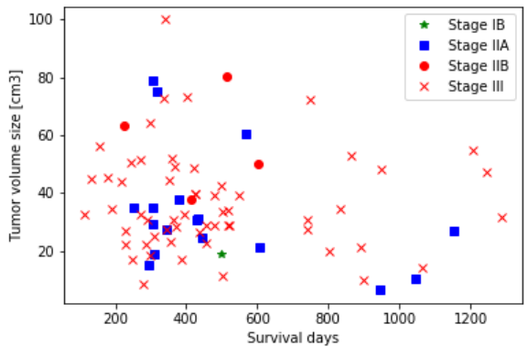


Figure 1: Clinical feature visualization for different cancer stages

4 Methods

We start with modeling selected clinical features with a Naive Bayes classifier. For this generative model, we divide the cohort into two groups with labels $y = 1$ for high risk and $y = 0$ for low risk patients. The clinical features are discretized and represented with multinomial feature vector x_d where each d represents a different clinical feature taking values in $\{1, \ldots, V\}$. V is the number of descritization buckets for feature d. We denote F for number of clinical features selected. The graph below represents our generative model:

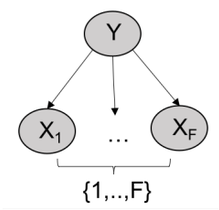


Figure 2: Naive Bayes Classifier Model

For each patient i, the likelihood of the data can be expressed as:

$$p(D_i) = p(y^i) \prod_{d=0}^{F} p(x_d^i | y^i)$$

For each clinical feature d, we can learn the model parameters $\phi_{dk} = p(x_d = k|y)$ independent of other clinical features. We use Maximum Likelihood estimates with Laplace smoothing on the training set $\{(x^i, y^i); i = 1, \ldots, n\}$ to learn the parameters:

$$\phi_{dk|y=1} = \frac{1 + \sum_{i=1}^{n} 1\{x_d^i = k \wedge y^i = 1\}}{V + \sum_{i=1}^{n} 1\{y^i = 1\}}$$

$$\phi_{dk|y=0} = \frac{1 + \sum_{i=1}^{n} 1\{x_d^i = k \wedge y^i = 0\}}{V + \sum_{i=1}^{n} 1\{y^i = 0\}}$$

$$\phi_y = \frac{\sum_{i=1}^{n} 1\{y^i = 1\}}{n}$$

Next, we will run non-parametric and parametric survival analysis on clinical and radiomic features. The survival function $S(t)$ is the probability that the random variable survival time T be bigger than specified time t:

$$S(t) = P(T > t) \tag{1}$$

This function can be naively interpreted as the proportion of population surviving after time t. The probability density function can be obtained by differentiating $S(t)$:

$$f(t) = -\frac{dS(t)}{dt} \tag{2}$$

We can also define hazard function h(t) as the instant probability of patient dying at time instance t given that that patient survived to time t. Mathematically:

$$h(t) = \lim_{\Delta T \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta T \to 0} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)\Delta t} = \frac{f(t)}{S(t)} \tag{3}$$

To estimate survival function (Eq 1) we start with Kaplan-Meier Estimate given by:

$$S_{KM}(t) = \prod_{i;t_i<t} \frac{n_i - d_i}{n_i} \tag{4}$$

This estimator does not take into account input features and only depends on n_i: number of patients at risk at time t_i and d_i: number of events at time t_i. This estimate can naively tell us the percentage of population surviving after a given point in time.

Next we test proportional hazard assumption on our model by fitting the data to Cox proportional Hazard model. This model takes into account the covariates of the patients (X_i) each influencing the patient risk through β_i parameters. The hazard function for proportional Hazard model can be expressed as [5]:

$$h(t, X) = h_0(t) exp\Big(\sum_{i}^{p} \beta_i X_i\Big) \tag{5}$$

To learn the parameters of this model, we can use partial likelihood function [5]:

$$L(\beta) = \prod_{t_i uncensored} \frac{exp\Big(\beta^T X_i\Big)}{\sum_{j \in R(t_i)} exp\Big(\beta^T X_j\Big)} \tag{6}$$

$$\beta_{ML} = \underset{\beta}{\operatorname{argmax}}\, L(\beta) \tag{7}$$

where $R(t_i)$ is the set containing patients at risk at t_i. To validate Cox proportional hazard model we applied 10-fold cross validation with Harrell's Concordance Index (C-index) scoring [8]. 10-fold cross validation was used because of small number of training examples (80 patients). For a pair of patients, C-index [7] measures, how often the model is predicting higher hazard score for the patient that actually has lower survival days:

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{1}[T_j < T_i]\mathbb{1}[\eta_j > \eta_i]}{\sum_{i,j} \mathbb{1}[T_j < T_i]} \tag{8}$$

where η_j is patient j hazard score, and T_j is patient j time to event. Higher C-index values (closer to 1) represents better model prediction.

5 Experiments and Results

Given small number of patient cases, we decided to use Synthetic Minority Oversampling Technique (SMOTE) [1] to augment the dataset. Specifically, from Figure 1 we observed that the high risk region (less than 1 year) only has few stage IIB and stage III cancer cases present. To apply SMOT, first we had to classify the patients into low and high risk cohorts using survival days as threshold. We experimented with 240, 270, 300, 330, 360 survival days. For each case we trained a Naive Bayes classifier on clinical features: Stage Grouping, Chemotherapy Time, Tumor Width, NCCN Resectability Criteria, and Radiation Therapy. The model diagram is shown in Figure 2 where X random variables are dicretized clinical features and Y is a Bernouli random variable representing low or high risk. To learn the model parameters, we used maximum likelihood estimate with Laplace smoothing. Using 10-fold cross validation, we compared the classifiers' performances and observed that the model trained with 270 days threshold data gives the best balanced performance. Table 1 shows the balanced performance for each of the Naive Bayes classifier models.

High/Low Risk Survival Days Threshold	Balanced Accuracy
240 days	0.53
270 days	**0.69**
300 days	0.61
330 days	0.57
360 days	0.61

Table 1: 10-fold cross validation balanced accuracy with Naive Bayes Classifier models

The SMOTE technique was implemented with imbalanced-learn Python toolbox [11] which randomly selects one of 5 nearest neighbors of minority class example data and generate a synthetic example between the two examples feature space. For feature space we used all the clinical and radiomic features. The effect of this dataset augmentation is visualized in Figure 3 with high-risk region of the dataset populated with more synthetic stage IIB and stage III patients.

With the augmented dataset, we moved on to regression survival analysis and implemented Cox proportional hazard models on clinical and radiomic feature sets using lifelines library [4]. The radiomic dataset had 900 features present in each case with many of the features highly correlated. For our first model, we trained on all 900 radiomic features and applied L2 regularization. Lifelines provides a penalizer term in Cox fitter API that control L1 and L2 regularization using the following equation:

$$\frac{1}{2}\text{penalizer}\Big((1 - \text{l1\_ratio})||\beta||_2^2 + \text{l1\_ratio}||\beta||_1\Big) \tag{9}$$

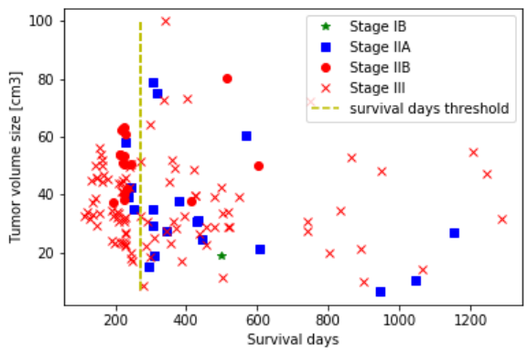


Figure 3: Clinical feature visualization after data augmentation by SMOT

We experimented with l1_ratio and observed that for a fixed penalizer, the model's concordance index improves with lower l1_ratio. The model did not converge at all when l1_ratio=1 (L1 regularization only). We picked penalizer=0.1 and l1_ratio=0 (L2 regularization only) for the rest of the trainings.

After training proportional hazard model on all 900 radiomic features, we identified 410 feature coeficients with large error band (>1). Given the small number of patient cases compared to large number of radiomic features, we decided to remove these 410 radiomic features from our training set to reduce variance in the model. The model's C-index trained on the remaining 490 radomics features was 0.76 as presented in Table 2.

For clinical data, we picked 12 features from clinical dataset to train the proportional hazard model. Same penalizar was used for both clinical and radiomic training for consistency. The 12 clinical features were Radiation Therapy, NCCN Resectability Criteria, Chemotherapy Time, Tumor location at the organ, Stage Grouping, Primary Tumor, KPS, Gender, and tumor dimensions.

Next, we used the combination of the clinical and radiomic features to train a new proportional hazard model. We evaluated the models using 10-fold cross validation with C-index scoring method Eq(8). Table 2 presents the C-index values for all three models.

Datasets for Cox Regression	Number of features	Concordance index
Clinical Features	12	0.67
Radiomic Features	490	0.76
Clinical and Radiomic Features	502	0.82

Table 2: 10-fold cross validation concordance index with Cox Proportional Hazard Models

In addition, with the proportional hazard model trained on clinical and radiomic features, we computed the partial hazard scores (baseline hazard not included) and took median score to divide patients into low and high risk groups. We then plotted separate Kaplan-Meier curves shown in Figure 4 for each group. As expected, the high risk group show lower survival probabilities over the whole duration of the study. Interestingly, the high risk curve shows survival probability dropping significantly before 270 days which was our Naive Bayes classifier best survival threshold. We also ran logrank test [4] on the two groups and observed the survival difference between these two groups is statistically significant (test statistic: 151.59, $p < 0.05$).

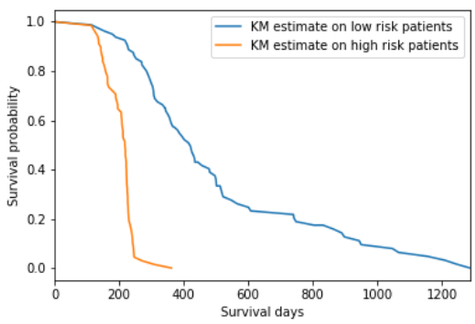


Figure 4: Kaplan Meier survival curves for high and Low risk patients

We also looked at how different features are impacting the survival predictability. First we reviewed the trained model coefficients and identified the top 15 features with the highest coefficients as shown in 5. Chemotherapy Time and Primary Tumor assessment (T2/T3/T4) have the highest prognostic values. Next, using Lifelines library [4], in Figure 6 we plotted survival curves for varying Chemotherapy and Cancer Stage covariates while keeping other covarites constant .

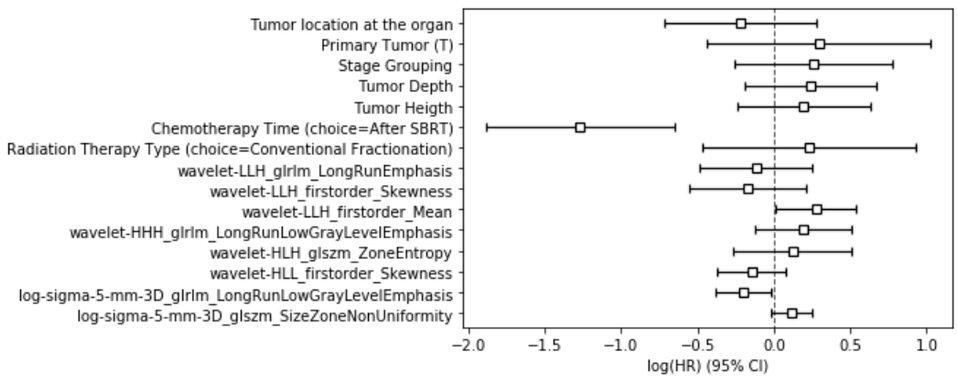
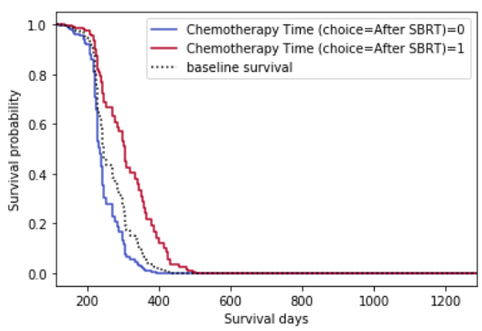
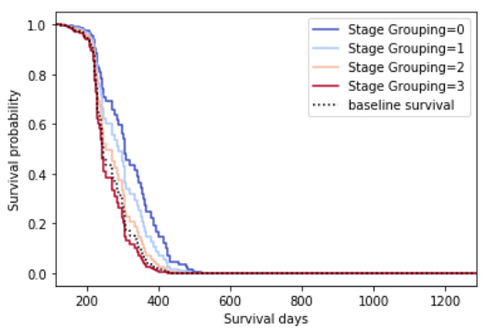


Figure 5: Cox proportional hazard model top 15 feature coefficients with error bands



(a) Chemotherapy Time: Unchecked:0, Checked:1



(b) Stage Grouping: Stage IB:0, Stage IIA:1, Stage IIB:2, Stage III: 3

Figure 6: Kaplan Meier survival curves with changing covariate

6 Conclusion and Future Work

We implemented Cox proportional hazard model to predict survival days for pancreatic cancer patients. Given the low number of training examples, we presented SMOT as an effective technique for augmenting the dataset. The challenge with applying SMOT was finding the right threshold for classifying the patients. Comparing Naive Bayes model performances gave us a robust way to automatically select the best threshold for classification. The Cox proportional hazard model identified Chemotherapy and Tumor assessment as key indicators on patient survival analysis. We also showed that adding radiomic features to regression analysis can improve the survival model performance compared to analysis done on clinical feature set alone.

Going beyond Cox proportional hazard model, we can apply neural networks to identify latent features that can better predict the patients hazard risk. We can also take a different approach and look at how we can learn the survival probability distribution of the patients directly.

7 Contributions

Arash Jamalian worked on this project with guidance from Dr. Haruka Itakura Assistant Professor of Medicine (Oncology) at Stanford University. Pancreatic cancer dataset was provided by Stanford Cancer Center under NDA agreement. Pyradiomics files were generated using Stanford Quantitative Image Feature Extraction(QFIP) pipeline.

References

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[2] Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. PLoS computational biology, 14(4):e1006076, 2018.

[3] David R Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, 1972.

[4] Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Paul Zivich, Ben Kuhn, Mike Williamson, sean reed, AbdealiJK, Andrew Fiore-Gartland, Deepyaman Datta, Luis Moneda, Gabriel, Daniel WIlson, Alex Parij, and Arturo Moncada-Torres. Camdavidsonpilon/lifelines: 0.24.6, May 2020.

[5] Frank Emmert-Streib and Matthias Dehmer. Introduction to survival analysis in practice. Machine Learning and Knowledge Extraction, 1(3):1013–1038, 2019.

[6] International Agancy for Research on Cancer, 2018.

[7] Stephane Fotso et al. PySurvival: Open source package for survival analysis modeling, 2019–.

[8] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. Jama, 247(18):2543–2546, 1982.

[9] Jared L Katzman, 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(1):24, 2018.

[10] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[11] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17):1–5, 2017.

[12] Zeina A. Shboul, Mahbubul Alam, Lasitha Vidyaratne, Linmin Pei, Mohamed I. Elbakary, and Khan M. Iftekharuddin. Feature-guided deep radiomics for glioblastoma patient survival prediction. Frontiers in Neuroscience, 13:966, 2019.