Oral Cancer (SR-BMC-Cancer 2022)

June 6, 2022

Abstract

Predicting the severeness of the disease of a patient could have a significant impact on the methods and treatment plans that should be done to cure a patient. Data gathered from clinical treatments and other studies can be used in order to use for the purpose of survival analysis. These data are usually high-dimensional, censored and contain missing information which brings about the need for methods which can overcome these challenges. Machine learning techniques because of their potential to estimate complex functions and overcome those challenges have recently received remarkable attention. In this paper, we run experiments and use different classes of machine learning models that have been developed for this purpose. We discuss about these classes and compare some instances of them. Also, we analyse the importance of features in our analysis.

1 Introduction

Oral cancer accounted for 2% of newly diagnosed cancers in 2020. In that year, more than 370,000 people worldwide were diagnosed with oral cancer, and over 170,000 death were caused by that (1). A little bit more than half of the patients who are diagnosed with oral cancer survive more than five years (2). These facts indicate that the death rate of oral cancer is significant and the time for treatment is short, which shows the importance of effective medical treatments against that. Therefore, identifying important prognostic factors in the survival of oral cancer patients and predicting their survivals can remarkably help to use more appropriate and necessary treatments for each individual patient (3). A closer follow-up and more aggressive treatment plan could be more beneficial for patients who have a shorter predicted survival time.

Survival analysis in the context of cancer is the study of the time between the diagnosis and the death from cancer (4). There has been a long-standing desire to accurately predict the survival of patients. Classic methods were first proposed for survival analysis. These methods had statistical perspectives and were used for estimating survival function or hazard function. For example, the Kaplan-Meier estimator which uses the event times to estimate the survival function $S(t)$ which expresses the probability that a patient given from the

distribution of patients survives after time t is a classical method which is being widely used in survival analysis. Another famous classic method is Cox PH which assumes that the hazard function of a patient is proportionally related to their explanatory variables, $x_1, x_2, ..., x_d$, via some baseline hazard that in the Cox PH model is defined by

$$h(\tau|x) = h_0(\tau) \exp(\beta^T x)$$

where h_0 is the non-negative baseline hazard function and $\beta = [\beta_1, ..., \beta_d]^T$ where $\beta_i \in \mathbb{R}$ are coefficients to be fit. The value of β is found by maximum likelihood estimation of the partial likelihood (14). The important point about classic models such as Cox PH is that they are easily interpretable and fast to fit with low complexity (15).

Recently, machine learning methods that are able to learn and extract important information from high-dimensional data have been employed to serve the purpose of survival analysis. There are many different algorithms in this regard (22). There are generally four classes of machine learning models for this purpose (15). First class consists of random forests which fit many simpler models called decision trees, and then averages the prediction of those trees to get the final prediction. This process is called 'Bagging' in literature. Decision trees are a common model in machine learning and have the advantage of being simple to implement and highly interpretable. Decision trees use split rules to create a series of splits based on features in the tree that leads to terminal nodes called leaves. In survival analysis, the splitting rules for decision trees have to handle censored data but there are many different splitting rules proposed for decision trees in this context (16). One of those main splitting rules is based on hypothesis testing (log-rank test), and one other famous rule is using likelihood-based measures (16). The second class consists of gradient boosting machines. Boosting strategy in machine learning comes from the idea that one can make a powerful learner based on a committee of weak learners. The difference between this class of models with random forests is that in random forests all of the components (decision trees) are trained independently while in gradient boosting we train components sequentially and their training is based on previous components and their performance and the resulting model is a linear combination of components (15). In these models, the selection of loss function and weak learners is important. One of the main choices for loss function is the partial log-likelihood (14) which wqs discussed above. The third class of machine learning models is support vector machines. Most of these models try to find a function $g(x) = \beta^T x + \beta_0$ that is able to predict the death time for patients. Some other tends to find a ranking of patients while others prefer to predict the hybrid of these two. Finally, we have neural networks which are capable of estimating complex functions. Some neural network models in survival analysis tend to discretize time and propose prediction on a limited number of points while some others provide continuous predictions and discretizing events is no longer needed. Definition of loss functions and how the model is going to predict desired functions is essential in developing neural networks. There are

many different neural networks models proposed in recent years (18, 19, 20, 21). Comparing different methods and analyzing their performance on datasets has been studied in recent years (17).

(9) Different types of input can lead to different methods. In some studies, images of patients have been used as inputs and machine learning methods for image processing such as convolutional neural networks which are able to extract important features from images by applying convolution layers on images have been hired to extract notable information from images and then the methods described above have been used for survival prediction. There have been many studies which benefit from deep convolutional neural networks for survival prediction (10, 11). With the advancement of new technologies such as next-generation sequencing, omics data has become available for a comprehensive assessment of a patient's condition. Therefore, in some other studies, omics data such as gene expression, DNA methylation, miRNA expression, and copy number variations have been the inputs of the models (12). For example (13) has developed a deep learning method to apply on multi-omics data for breast cancer analysis. (23) has used integration of histopathological images and genomic data and has proposed a biologically interpretable deep learning model for this integration. In some other studies, tabular information which expresses different features of a patient which are related to the study has been given to survival analysis models.

There are also some related studies where different models were evaluated. For example, (4) studied data collected from 255 oral cancer patients and trained Cox PH, random survival forest, and DeepSurv on that. They reported training and test accuracy of those models based on c-index measurement. In their study, DeepSurv led to the best test error. Moreover, (5) analyzed data from patients who were diagnosed with head and neck cancers and used random survival forests to study the importance of each prognostic factor in their analysis. They also compared different machine learning techniques for survival prediction and compared those models based on c-index and AUC. They found out that in their study, RSF performed the best. Some others have proposed novel models for prediction. For example, (6) designed a new neural network and evaluated its performance on lung cancer patients, and showed that their model outperformed Cox Net and Cox PH in c-index criterion on that dataset.

We have gathered oral cancer patients data in .... from ... to .... In this study, we run different models and compare their performance based on c-index, integrated brier score. we use random survival forests and gradient boosting survival as well as classic models such as Cox PH. Also, some famous survival analysis methods based on neural networks have been explored in this paper. Additionally, we have analyzed the interpretability of our models based on our features.

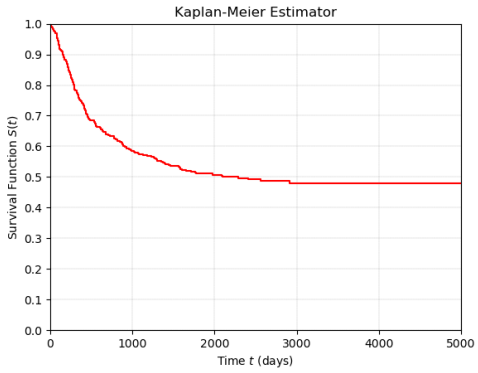


Figure 1: Estimation of survival function $S(t)$ with Kaplan-Meier estimator.

2 Method

2.1 Data

We gathered the information of 526 patients who were diagnosed with oral cancer and were being treated at ... hospital. The clinical characteristic of our patients is depicted in Table 2. For each patient, we have 22 features. The Kaplan-Meier survival curve of our analyzed patients is depicted in Figure 1. Before training models, the entire dataset was split into training/validation set with 75% of subjects and test set with 25% of subjects. We imputed the missing values in our dataset using k-nearest neighbors (kNN) with $k = 10$.

2.2 Models

2.2.1 Random Survival Forest

As we discussed in the introduction section, we have explored the performance of random forests on our dataset. Random survival forests have many hyper-parameters such as number of decision trees, maximum depth of the trees, maximum number of features selected by estimators, and other hyper-parameters. We have used 5-fold cross-validation to choose the best hyper-parameters. This model also gives us the importance of different features on the prediction of the model.

2.2.2 Gradient Boosting Machine

We have also evaluated gradient boosting machines in our analysis. This model also has many hyper-parameters such as learning rate, loss function, number of estimators, and others. Like the random survival forest training, we have used 5-fold cross validation to tune models and get the best hyper-parameters. We have also extracted features' importance with this model.

2.2.3 Cox PH

As one of the most frequently used methods for survival analysis, we have also analyzed the classic Cox PH model. Hazard function predicted by this model can be used to compute the survival function of patients.

2.2.4 Deep Networks

We have analyzed the DeepSurv, which is a deep feed-forward neural network that predicts the log-risk function of a given patient (7). The DeepSurv tries to find an appropriate model to minimize the negative log Cox partial likelihood. In addition, we used the Logistic Hazard model (8) which is a model that discretizes event times and predicts the hazard function of a patient with neural networks. PC-Hazard(8) is another deep-learning method that we have run on our dataset and compared its accuracy with other methods. Dropout has been used for the generalization of these models.
In order to tune these models, we have chosen a subset of training data as aa validation set. Different architectures and learning rates have been evaluated on the validation set to find the best possible of each mentioned model.

2.3 Evaluation Metrics

We have evaluated discussed models on two different criteria. One of them is c-index, which is defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs. c-index equals to 0.5 is the average accuracy of a random model, and c-index equals to 1 refers to the best death ranking of patients.
Brier score is another metric that evaluates the performance of a model given its predicted survival functions for different patients at a particular time. Integrated Berier score computes the average of Berier score on a given time period. A random model can get a Berier score equals to 0.25, and lower Berier score shows that a model predicts better.

3 Result

Table 1 shows the accuracy of models on our dataset. All of the features have been used to train and test models. Our experiments show that among neural networks logistic hazard model has outperformed others and also random survival forests have shown great results. Figure 2 shows the predicted survival function of the logistic hazard and random survival forest for ten patients in test data. Among those patients, both models have identified patient numbers 3 and 8 as high-risk patients whose survival function decreases quickly while survival function of patients 0, 4, and 7 in both models show that they are not in danger and their survival function is significantly high even after a long period of time. Random survival forests and gradient boosting models give us

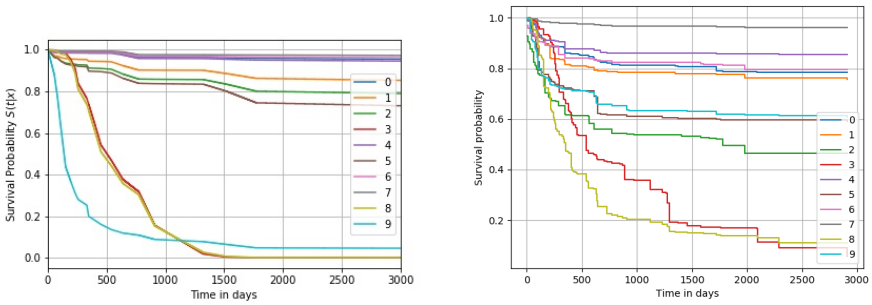
Models	Train		Test	
	C-index	IBS	C-index	IBS
RSF	0.91	0.05	0.83	0.10
GBS	0.92	0.05	0.81	0.08
Cox PH	0.82	0.11	0.78	0.12
Logistic Hazard	0.92	0.05	0.81	0.11
	0.78	0.15		
DeepSurv	0.83	0.07	0.74	0.12
	0.81	0.10		
PC Hazard	0.86	0.06	0.80	0.12
	0.79	0.10		

Table 1: Performance of models on our dataset. Integrated berier score has been evaluated in a 5-year period. All numbers have been round to two decimal points.

the importance of features. Figure 3 shows the importance of features in our study. This figure tells us that features such as Age, radiation therapy, surgery, pathological level, and recurrence are important features in both of the models. Also, histological stage and chemotherapy are two features which have a significant impact on GBS and RSF respectively. We have also analyzed important features with our best neural network model -logistic hazard- by evaluating the accuracy of a model trained by only a subset of features. Subsets of features which are able to train a model with a significant accuracy show us that those features can be enough in our analysis while subsets of features which are not capable of training a model to reach a good accuracy show that some of important features are missing in those subsets. By precisely analyzing the accuracy of models trained on different subsets, we can gain information about features that play a significant role in the accuracy of the logistic hazard model.

In order to evaluate different subsets, we have used two different techniques. The first method evaluates 1000 random subsets of length 15 of features and trains the model based on only the features in those subsets. Figure 4a shows the distribution of accuracy (C-index) in these 100 experiments. We have put a threshold of C-index equals to 0.80 and Figure 4b shows the probability that each feature does not appear in a subset with accuracy greater than the threshold on validation data. We can deduce that features with low removal probability in this figure have significant role in the accuracy of the model while features with high removal probability indicate that they are not important in the performance of the model. In our analysis, we can infer that undertaking chemotherapy, surgery, age, and undertaking recurrence surgery have the minimum probability which indicates that their existence is important for the sake of prediction.

The second method is to use exhaustive search to remove two features from the set of all features and then train the model based on the remaining features. We have explored all the possible combinations of removing two different features



(a) Survival function predicted by Logistic Hazard

(b) Survival function predicted by RSF

Figure 2: Survival function ($S(t|x)$) of ten patients in test data. Both model are approximately giving the same result. Patients with very high and very low risk in both models are same.

from our data. Figure 5a shows the distribution of C-index in all of those experiments. Figure 5b indicates the probability that a feature does not appear in a subset capable of training the model with an accuracy of at least 0.81 on test data. The knowledge we can infer from these figures is consistent with what we have discovered about our features so far. For example, it emphasizes that undertaking chemotherapy is important for survival prediction.

4 Discussion

5 Conclusion

In this study, we discussed about traditional methods and machine learning classes for survival analysis. We explored some random forests, gradient boosting machines, and deep learning models and compared the accuracy of these models under c-index and integrated berier score criteria. Random survival forests and logistic hazard model achieved the best results. We have also tried to interpret our models and understood the effect of features of our dataset on our models. Features such as Age, undertaking radiation therapy, undertaking surgery, undertaking chemotherapy, histological, and recurrence have shown significant impact in our study.

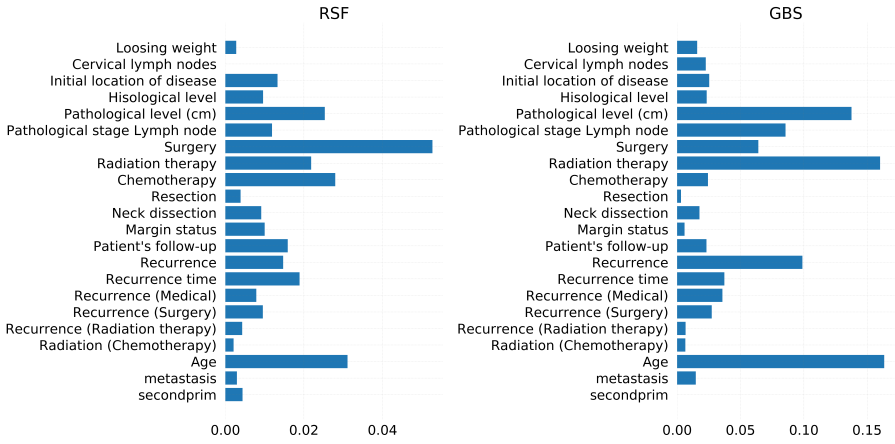


Figure 3: Relative importance of features in gradient boosting model and random survival forest.

6 Appendix

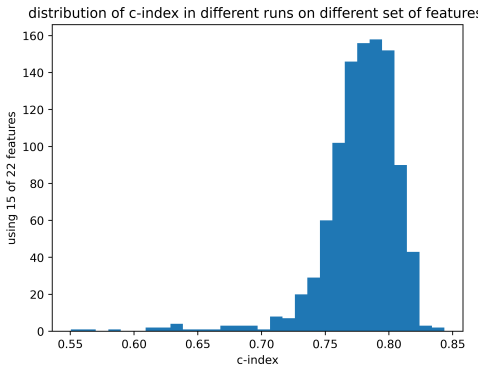
Information about our features and the values that they can take can be found in Table 2. Some other information about the hyper-parameters of models we have used to evaluate on our dataset is also given in Table 3 *TODO (Also get info for description of the meaning of these features*.

Loosing weight (0/1)	Cervical lymph nodes (0/1)	Initial location of disease Categorical	Histological level Categorical
Pathological level (cm) Numerical	Pathological stage Lymph node Numerical	Radiation therapy (0/1)	Resection (0/1)
Surgery (0/1)	Chemotherapy (0/1)	Neck dissection (0/1)	Margin status (0/1)
Patient's follow-up (0/1)	Recurrence (0/1)	Recurrence time (Categorical)	Recurrence (Medical) (0/1)
Recurrence (Surgery) (0/1)	Recurrence (Radiation therapy) (0/1)	Radiation (Chemotherapy) (0/1)	Death (0/1)
Age Numerical	metastasis (0/1)	secondprim (0/1)	- -

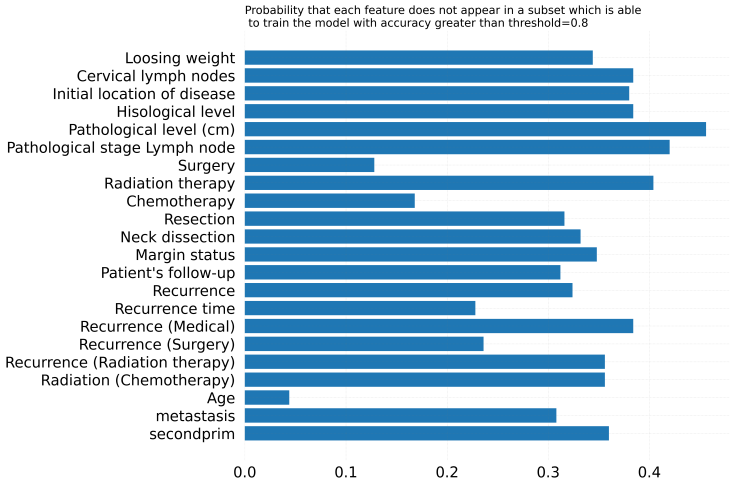
Table 2: List of features we have gathered in our study. type of data is written under the name of each feature.

7 Reference

1. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries
2. Global epidemiology of oral and oropharyngeal cancer
3. Deep learning-based survival prediction of oral cancer patients
4. Long-term cancer survival prediction using multimodal deep learning
5. Machine Learning Incorporating Host Factors for Predicting Survival in Head and Neck Squamous Cell Carcinoma Patients
6. SurvNet: A Novel Deep Neural Network for Lung Cancer Survival Analysis With Missing Values
7. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network
8. Continuous and Discrete-Time Survival Prediction with Neural Networks
9. A Theoretical and Methodological Framework for Machine Learning in Survival Analysis
10. Deep convolutional neural networks for imaging data based survival analysis of rectal cancer
11. Image-based Survival Analysis for Lung Cancer Patients using CNNs
12. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis
13. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer
14. Partial Likelihood
15. A Theoretical and Methodological Framework for Machine Learning in Survival Analysis
16. A review of survival trees
17. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction
18. RNN-SURV: a Deep Recurrent Model for Survival Analysis
19. Time-to-Event Prediction with Neural Networks and Cox Regression
20. Deep Extended Hazard Models for Survival Analysis
21. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks
22. Machine Learning for Survival Analysis: A Survey
23. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data

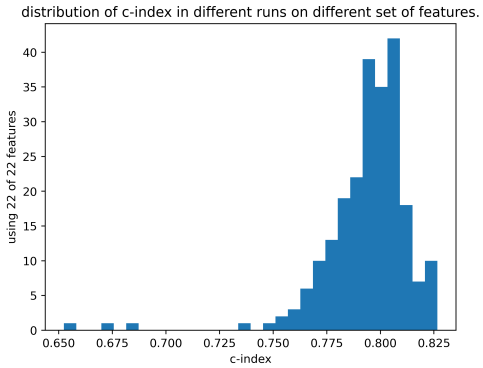


(a) C-index distribution

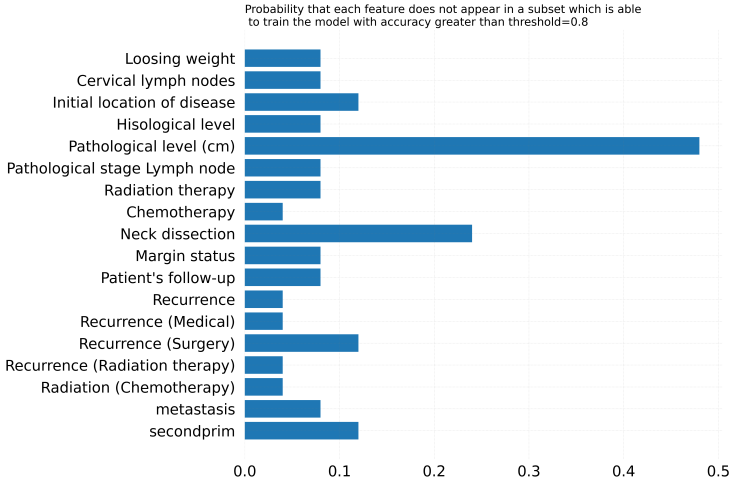


(b) Probability of appearance of features

Figure 4: (a) shows the distribution of C-index when logistic hazard model is trained on a random subset of features containing 15 elements. (b) depicts the probability that each feature does not appear in a subset with good accuracy (> 0.80).



(a) C-index distribution



(b) Probability of appearance of features

Figure 5: (a) shows the distribution of C-index when logistic hazard model is trained on all subsets containing 20 elements. (b) depicts the probability that each feature does not appear in a subset with good accuracy (> 0.81).