

# Developing an Individualized Survival Prediction Model for Rectal Cancer

Ana Silva<sup>1</sup>, Tiago Oliveira<sup>1</sup>, Paulo Novais<sup>1</sup>, José Neves<sup>1</sup>

Algoritmi Centre/Department of Informatics, University of Minho, Braga, Portugal  
a55865@alunos.uminho.pt  
{toliveira,pjon,jneves}@di.uminho.pt

**Abstract.** This work presents a survivability prediction model for rectal cancer patients developed through machine learning techniques. The model was based on the most complete worldwide cancer dataset known, the SEER dataset. After preprocessing, the training data consisted of 12,818 records of rectal cancer patients. Six features were extracted from a feature selection process, finding the most relevant characteristics which affect the survivability of rectal cancer. The model constructed with six features was compared with another one with 18 features indicated by a physician. The results show that the performance of the six-feature model is close to that of the model using 18 features, which indicates that the first may be a good compromise between usability and performance.

## 1 Introduction

The most common cancer of the digestive system is colorectal cancer, also known as bowel cancer, which develops in the cells lining the colon and rectum [18]. About 70 percent of the colorectal cancer cases occur in the colon and about 30 percent in the rectum [9]. Although colon and rectal cancers are considered to be very similar pathologies, they have different associated genetic causes and different progressions according to distinct molecular pathways [21]. The work disclosed herein focuses solely on rectal cancer, the anatomic part where material called feces or stool is stored until it is expelled of the body through the anus. Machine learning (ML) methods have been widely applied in cancer research, due to their competence in identifying relevant information from complex datasets. An accurate survivability prediction helps physicians in effective and precise decision-making. Although there are some tools which provide survivability predictions for rectal cancer, none of them apply ML techniques in order to build evolving predictive models. Therefore, and following the previous work developed for colon cancer patients[17], the aims of this work are the following: i) to make an individualized prediction of the survivability of a rectal cancer patient in each year of the five years following treatment; and ii) to determine which features are the most important for survivability prediction of rectal cancer patients. The number of features can be crucial when available in a clinical

decision support tool, which is the end goal of the work. The number can determine the use or not of a application, taking into account the time to obtain an output (a prediction). The prediction model was developed using data from the Surveillance, Epidemiology, and End Results (SEER) program [13], the most complete cancer database in the world.

The paper is structured as follows. Section 2 describes related work in rectal cancer survival prediction. Section 3 provides the steps and machine learning methods used to develop the prediction model. The corresponding experimental results are disclosed and discussed in Section 4. Finally, Section 5 presents the conclusions drawn so far and future work considerations.

## 2 Related Work

Existing approaches to calculate rectal cancer patients survivability are regression-based. Wang et al. [20] developed nomograms to make an individualized prediction of the conditional survivability for rectal cancer patients. The estimate is valid when calculated after a certain period of time (months) passed since diagnosis and treatment. The model was constructed based on data from 42,830 patients who were diagnosed between 1994-2003, from the SEER database. Conditional survivability prediction is calculated from a Cox proportional hazards model. The primary outcome variable was overall survivability conditional on having survived up to 5 years from diagnosis. Covariates included in the model were age, race, sex/gender and stage. The C-index for the model of this approach was 0.75 and the model is available as a web-based calculator. Valentini et al. [19] developed a tool to predict the probability that a rectal cancer patient will be alive or will have local recurrence or distant metastasis after delivery of long-course radiotherapy, with optional concomitant and/or adjuvant chemotherapy, over a 5-year period after surgery. Based on Cox regression, multivariate nomograms were developed through 2,795 individual patient data collected from five European randomized trials<sup>1</sup>, between 1992 to 2003. Selected by training data, the required information for the overall survivability calculator was gender, age at the date of randomization, clinical tumor stage, radiotherapy dose, surgery procedure, adjuvant chemotherapy (yes/no), pathological tumor and nodal stage. The concomitant chemotherapy (yes/no) is used to calculate the local recurrence. However, it must be inserted, even for overall survivability prediction, because it is a field required for the tool. The nomogram for overall survivability had a C-index of 0.70. Another SEER-based approach is the one developed by Bowles et al. [1], also made available in the form of an internet-based individualized conditional survivability calculator. This tool consists of four separate multivariate Cox regression models, taking into account: no radiotherapy, preoperative radiotherapy, postoperative radiotherapy and stage IV patients. These models were

---

<sup>1</sup>Trial name: European Organisation for Research and Treatment of Cancer, *Fdration Francophone de Cancrologie Digestive*, Working Group of Surgical Oncology/Working Group of Radiation Oncology/Working Group of Medical Oncology of the Germany Cancer Society, Polish and Italian.

created to determine adjusted survival estimates (at year 1 through 10) and used to calculate 5-year adjusted conditional survivability. They were constructed using registries of 22,610 patients with rectal adenocarcinoma, who were diagnosed from January 1988 to December 2002. Models developed for patients who underwent no radiotherapy, preoperative radiotherapy or postoperative radiotherapy, covariates were the same. They included age, sex/gender, race, tumor grade, surgery type and stage. In the model built for stage IV patients, i.e., for patients with distant metastasis, the surgery type was treated as a binary variable in the model (using any radiotherapy or primary tumor directed surgery as covariates). The measures of performance for this tool are not available.

The approach developed herein distances itself from already existing works by treating survival prediction as a classification problem and applying varied machine learning methods to obtain a qualified model of individualized survival prediction.

### 3 Development of the Prediction Model

The rectal cancer survival prediction model should have the capacity to accept a number of inputs for selected prediction features and, for each of the 5 years following treatment, produce an output stating whether the patient will survive that year or not, along with a confidence value for the prediction. Survivability prediction was approached as a binary classification problem, so that five classification models for each year were developed and were posteriorly combined, in a programmatic manner. The development of these prediction models involved the following phases by order of occurrence: preprocessing of SEER data, split dataset, balancing data, feature selection, modeling, and evaluation. The software chosen to develop the prediction model was RapidMiner, an open source data mining software. It has a workflow-based interface that offers an intuitive application programming interface (API).

#### 3.1 Preprocessing, Split Dataset, and Balancing Data

The colorectal cancer data from SEER were collected from 1973 to 2012. It contained 515,791 registries and 146 attributes, some of them only applicable to a limited period within the time of data collection. In the Preprocessing phase, it was defined that the period of interest would be from 2004 onwards, minimizing the occurrence of missing data due to the applicability of the attributes. Pediatric patients (age under to 18 years old) were removed. Patients who were alive at the end of the data collection whose survival time had not yet reached five years (the maximum period for which the model under development is supposed to predict survival), and those who passed away of causes other than colon or rectal cancer were sampled out from the training set as their inclusion was considered to be unsuited to the problem at hand. Binary classes (*survived* and *not survived*) were derived for the target labels 1-, 2-, 3-, 4- and 5-year survival. Finally, based on existing attributes and at the request of a physician who collaborated in this

work, new attributes, such as the number of regional lymph negative nodes, the ratio of positive nodes over the total examined nodes and also patient relapse, were calculated. After the Preprocessing phase, the attributes were reduced to 61, including the new attributes and the target labels and the data was reduced to 12,818 registries. During the Split Dataset phase, the data was separated into five sub-datasets by target label, taking into account the corresponding survivability year. Table 1 shows the class distribution in each sub-dataset.

Observing Table 1 is seen that the classes are not equally represented. Studies [4, 11] show that the problem of using imbalanced datasets is important, from both the algorithmic and performance perspectives. An overview of classification algorithms for the resolution of this kind of problem [7] concluded that hybrid sampling techniques, i.e., combining over-sampling of the minority class with under-sampling of the majority class, can achieve better performances than just oversampling or undersampling. As such, in the Balancing Data phase, hybrid sampling, as described in [7], was applied in order to generate balanced sub-datasets with 12,818 records each.

Table 1: Class distribution for each target label in the sub-datasets.

	Target Labels				
	1 Year	2 Year	3 Year	4 Year	5 Year
Not Survived	4.03%	5.89%	7.17%	8.08%	8.70%
Survived	87.88%	82.27%	78.41%	75.68%	73.79%

### 3.2 Feature Selection

For the Feature Selection phase was used the Optimize Selection operator (implementing a deterministic and optimized selection process with decision trees and *forward selection*)[15] of RapidMiner. This phase was essential to discover the most influential features on the survival of rectal cancer patients. The process was applied to each sub-dataset for the target label and the common selected features to all the sub-datasets were used to construct the prediction models. Table 2 shows the selected features and their meaning. The 6 selected features were compared with a set of 18 features indicated by a specialist physician on colorectal cancer. In the subsequent modelling, it was assumed that the features had an equal weight, but further experimentation with biased models is required.

### 3.3 Modeling and Evaluation

During the Modeling phase, the classification strategy adopted consisted in the application of ensemble methods. In order to boost basic classifiers and improve their performance, the classification schemes used were meta-classifiers. All the classifiers combinations were explored, according to the algorithms and type of attributes allowed. The tested meta-classifiers were the same used in the previous

Table 2: Attributes selected in the Feature Selection process.

Attribute	Description
Age recode with < 1 year old	Age groupings based on age at diagnosis (single-year ages) of patients (< 1 year, 1-4 years, 5-9 years, ..., 85+ years)
CS Site-Specific Factor 1	The interpretation of the highest Carcinoembryonic Antigen (CEA) <sup>2</sup> test results
CS Site-Specific Factor 2	The clinical assessment of regional lymph nodes
Derived AJCC Stage Group	The grouping of the TNM information combined
Primary Site	Identification of the site in which the primary tumor originated
Regional Nodes Examined	The total number of regional lymph nodes that were removed and examined by the pathologist

work about colon cancer survivability prediction [17]: AdaBoost [6], Bagging [3], Bayesian Boosting [15], Stacking [5], and Voting [10].

Since survivability prediction is being handled as a classification problem, a group of basic classifiers were selected to be used in ensembles with the above-described meta-classifiers. The group includes some of the most widely used learners [15] available in RapidMiner, namely the k-NN (Lazy Modeling), the Naive Bayes (Bayesian Modeling), the Decision Tree (Tree Induction), and the Random Forest (Tree Induction). Fourteen classification schemes were constructed for the sets of 6 and 18 attributes and, for 1, 2, 3, 4, and 5 survival years. The combinations of meta-classifiers with basic classifiers were as follows. The Voting model used k-NN, Decision Tree and Random Forest as base learners. The Stacking model used k-NN, Decision Tree, and Random Forest classifiers as base learners, and a Naive Bayes classifier as a Stacking model learner. The AdaBoost, Bagging, and Bayesian Boosting models were combining with each basic classifier. In the evaluation process, 10-fold cross-validation [16] was used to assess the prediction performance of the generated prediction models and avoid overfitting. All classification schemes was evaluated and compared using the prediction accuracy and the area under the ROC curve (AUC). The accuracy is the percentage of correct responses among the examined cases [2]. The AUC can be interpreted as the percentage of randomly drawn data pairs of individuals that have been accurately classified in the two populations, and it is commonly used as a measure of quality for classification models [2].

## 4 Experimental Results and Discussion

A vast quantity of results was analyzed. From the results obtained, the top three performing algorithms, for each evaluating method described in Section 3, are present in Tables 3 and 4, for each of the 5 years. For an easier interpretation, the average performances were calculated, allowing a better comparison between

algorithms and the selection of the best model. This is shown in Figure 1a and Figure 1b for the top three performing algorithms.

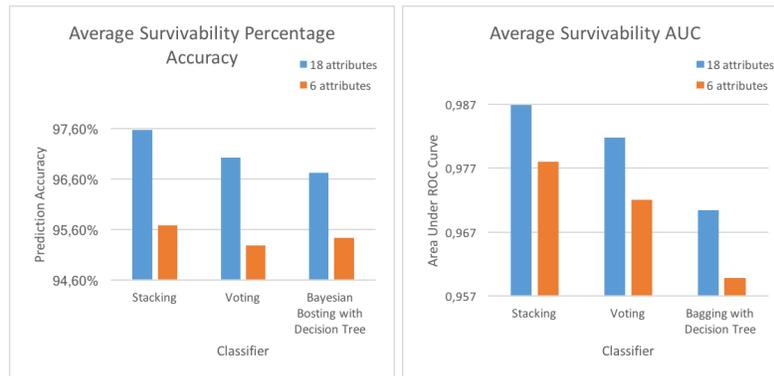
Table 3: Survivability Percentage Accuracy.

Ensemble Model	1 Year		2 Year		Accuracy 3 Year		4 Year		5 Year	
	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes	18 attributes	6 attributes
Stacking	98.32%	96.45%	98.00%	96.15%	97.72%	95.79%	96.97%	95.05%	96.88%	95.01%
Voting	97.37%	95.97%	97.16%	95.91%	97.20%	95.32%	96.79%	94.63%	96.62%	94.64%
Bayesian Boosting with Decision Tree	97.16%	96.26%	97.08%	96.06%	96.75%	95.19%	96.22%	95.01%	96.42%	94.66%

Table 4: Survivability AUC.

Ensemble Model	1 Year		2 Year		AUC 3 Year		4 Year		5 Year	
	18 attributes	6 attributes								
Stacking	0.988	0.976	0.989	0.981	0.987	0.979	0.984	0.977	0.986	0.977
Voting	0.985	0.975	0.983	0.976	0.983	0.971	0.979	0.969	0.979	0.969
Bagging with Decision Tree	0.985	0.981	0.974	0.964	0.964	0.955	0.967	0.948	0.962	0.951

Among the 18-attribute models, the stacking algorithm stood out. Among the 6-attribute models, the same has happened. The 18-attribute models had an average accuracy of 97.58%, with values for years 1 to 5 of 98.32%, 98.00%, 97.72%, 96.97% and 96.88%. The average AUC was 0.987, and the remaining values were 0.988, 0.989, 0.987, 0.984 and 0.986 for years 1 to 5. With an average of 95.69% for accuracy and 0.978 for AUC, the 6-attribute stacking models had prediction accuracies for years 1 to 5 of 96.45%, 96.15%, 95.79%, 95.05% and 95.01% (as seen in Table 3), and AUCs of 0.976, 0.981, 0.979, 0.977 and 0.977 (as seen in Table 4). Comparing the results of the 6-attribute stacking models with those of the 18-attribute models, the performances values are close, being slightly better for the 18-attribute models. The gap between accuracy measures are 1.89% and 0.009 for AUC. It is possible to say that the differences are not significant, taking account the contrast of feature numbers. The results show that it is possible to build a model with less than half of the features indicated by the expert physician. Regarding the attributes obtained in the feature selection process, with the exception of the site-specific factors, they were all connected with the features indicated by the specialist physician. The regression based approaches mentioned in Section 2 utilized the C-index to evaluate the models. This measure and AUC are considered numerically identical [8]. Since both correspond to the probability of giving a correct response in a binary prediction problem. As such, the present work represents an improvement and was able to achieve better results. In addition, comparing this approach with the previous work developed for colon cancer[17], results were similar. The third performing algorithm was not the same. However, the best performance scheme also was Stacking. In the colon cancer prediction model, the performance values were slightly better, but not more than 1.13% for accuracy and 0.011 for AUC. The most surprising result of the both approaches are the selected features, which were the same.



(a) Average survivability percentage accuracy. (b) Average survivability AUC.

Fig. 1: Comparison of the 18-attribute models with the 6-attribute models.

## 5 Conclusions and Future Work

This work involved the application of different meta-classification schemes to construct survivability prediction models for rectal cancer patients. The best performing scheme presented uses a Stacking classification scheme, combining k-NN, Decision Tree, and Random Forest classifiers as base learners and a Naive Bayes classifier as a stacking model learner. The relevant number of features for rectal cancer survivability prediction was found to be 6, the same selected for colon cancer. The set includes: age, CS site-specific factor 1, CS site-specific factor 2, derived AJCC stage group, primary site, and regional nodes examined. Overall, the developed model was able to present a good performance with fewer features than the existing approaches. As future work we intend to construct a mobile application to make both models (colon and rectal cancer prediction models) available to the health care community and to integrate it in settings of ambient assisted living and group decision making [14, 12]. In order to have the tool always updated and adapted to new patients, an on-line learning scheme is being prepared. This functionality will allow to dynamically feed new cases to the prediction system and make it change in order to provide better survival predictions. Future work also includes the development of conditional survivability models, enabling the user to get a prediction knowing that the patient has already survived a number of years after diagnosis and treatment. Additionally, we intend to conduct experiments to assess how well the tool fulfils the needs of health care professionals and identify aspects to improve.

## Acknowledgements

This work has been supported by COMPETE: POCI-01-0145-FEDER-0070 43 and FCT Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/ 00319/2013. The work of Tiago Oliveira is supported by a FCT grant with the reference SFRH/BD/85291/ 2012.

## References

1. Bowles, T.L., Hu, C.Y., You, N.Y., Skibber, J.M., Rodriguez-Bigas, M.A., Chang, G.J.: An individualized conditional survival calculator for patients with rectal cancer. *Diseases of the colon and rectum* 56(5), 551–9 (2013)
2. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7), 1145–1159 (1997)
3. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
4. Chawla, N.V.: Data Mining for Imbalanced Datasets: An Overview. In: *Data Mining and Knowledge Discovery Handbook*, pp. 853–867 (2005)
5. Džeroski, S., Ženko, B.: Is combining classifiers with stacking better than selecting the best one? *Machine Learning* 54(3), 255–273 (2004)
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55(1), 119–139 (1997)
7. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. *Int. J. Emerg. Technol. Adv. Eng* 2(4), 42–47 (2012)
8. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143(1), 29–36 (1982)
9. U. S. National Institutes of Health, N.C.I.: Seer training modules, colorectal cancer
10. Kittler, J.: Combining classifiers: A theoretical framework. *Pattern Analysis and Applications* 1(1), 18–27 (1998)
11. Leon, M.R.C.D., Jalao, E.R.L.: Prediction Model Framework for Imbalanced Datasets (c), 33–41 (2014)
12. Lima, L., Novais, P., Neves, J., Bulas, C.J., Costa, R.: Group Decision Making and Quality-of-Information in e-Health Systems. *Logic Journal of the IGPL* 19(2), 315–332 (2011)
13. National Cancer Institute: Surveillance, epidemiology and end results program (2015), <http://seer.cancer.gov/data/>, last visited on 10/01/2015
14. Novais, P., Costa, R., Carneiro, D., Neves, J.: Inter-organization cooperation for ambient assisted living. *J. Ambient Intell. Smart Environ.* 2(2), 179–195 (Apr 2010)
15. RapidMiner: Rapidminer documentation: Operator reference guide (2016), <http://docs.rapidminer.com/studio/operators/>, last visited on 03/01/2016
16. Refaeilzadeh, P., Tang, L., Liu, H.: Cross-validation. In: LIU, L., ÖZSU, M. (eds.) *Encyclopedia of Database Systems*, pp. 532–538. Springer US (2009)
17. Silva, A., Oliveira, T., Novais, P., Neves, J., Leão, P.: Developing an Individualized Survival Prediction Model for Colon Cancer, pp. 87–95. Springer International Publishing, Cham (2016), [http://dx.doi.org/10.1007/978-3-319-40114-0\\_10](http://dx.doi.org/10.1007/978-3-319-40114-0_10)
18. Vachani, C., Prechtel-Dunphy, E.: All about rectal cancer (2015), <http://www.oncolink.org/types/article.cfm?aid=108&id=9457&c=703>, last visited on 27/12/2015
19. Valentini, V., van Stiphout, R.G., Lammering, G., et al.: Nomograms for Predicting Local Recurrence, Distant Metastases, and Overall Survival for Patients With Locally Advanced Rectal Cancer on the Basis of European Randomized Clinical Trials. *Journal of Clinical Oncology* 29(23), 3163–3172 (2011)
20. Wang, S.J., Wissel, A.R., Luh, J.Y., et al.: An interactive tool for individualized estimation of conditional survival in rectal cancer. *Annals of surgical oncology* 18(6), 1547–52 (2011)
21. Yamauchi, M., Lochhead, P., Morikawa, et al.: Colorectal cancer: a tale of two sides or a continuum? *Gut* 61(6), 794–797 (2012)