Personalized colorectal cancer survivability prediction with machine learning methods

Samuel Li
Princeton University, Princeton, NJ
seli@princeton.edu

Abstract

In this work we investigate the importance of ethnicity in colorectal cancer survivability prediction using machine learning techniques and the SEER cancer incidence database. We compare model performances for 2-year survivability prediction and factor importance rankings between Hispanic, white, and mixed patient populations. Our models consistently performed better on single-ethnicity populations, and provided different factor importance rankings when trained on different populations. Additionally, after training our models to achieve higher AUC score than the best reported in literature, we improved G-mean score with imbalanced classification techniques. These results provide evidence in favor for increased consideration of patient ethnicity in cancer survivability prediction, and for more personalized medicine in general.

Keywords: Cancer survivability prediction, SEER, machine learning, personalized medicine, imbalanced classification

1. Introduction

Colorectal cancer, defined as cancer starting in the colon or the rectum, is among the most common types of cancer for both men and women. The probability of developing colon or rectum cancer is about 4.5% for men and 4.15% for women. In 2018, the National Cancer Institute estimates there will be 140,250 new diagnoses of colorectal cancer and over 50,630 deaths caused by these cancers in the United States[1]. This makes colorectal cancer the third leading cause of cancer-related deaths in the country. A patients survival time is largely dependent on the state of their cancer at the time of diagnosis: the 5-year survival rate for people with stage I colon cancer is approximately 92%, while the 5-year survival rate of stage IV colon cancers is approximately 11%[2].

Upon receiving their diagnosis, a patient will likely want to know their x-year survivability, defined as the probability that they will survive beyond x years. An estimate of their survivability based only on stage of cancer does not account for all the personal factors specific to that patient, and will be far less accurate than a prediction model that takes into account factors like patient age, race, primary site, etc. While it is difficult for a doctor to give consideration to many factors, machine learning algorithms can efficiently find patterns in large datasets of patient data. Given the vast amount of diagnostic data that is available and continuously being generated, there are many opportunities to use machine learning techniques to provide more personalized and accurate survivability predictions.

There is a sizable literature dedicated to the application of machine learning in cancer prognosis and survivability prediction. Delen et al. used data mining methods to predict breast cancer survivability in 2004[3]. Since then, with the increased adoption of machine learning methods, researchers have applied a wide variety of algorithms to survivability prediction of many types of cancer. Studies include use of deep learning[3, 4], ensemble methods[5], and imbalanced classification techniques[6]. These studies all predict survivability based on detailed personal and diagnostic information for each patient.

One approach to provide patients with more personalized healthcare is to take the patients race into account. It is well known that people of different racial groups have varying levels of susceptibility and responses to different diseases. Causes for discrepancies may include anything from genetic differences to environmental influences to cultural factors. Currently, there is significant disparity in colorectal incidence and mortality rates between different racial groups. From 2011 to 2015, the National Cancer Institute reported that the colorectal cancer incidence rate was 38.8 per 100,000 for the white population, while the Hispanic and black populations had incidence rates of 47.6 and 33.5, respectively. In the same years, the white, Hispanic, and black populations had recorded mortality rates of 14.1, 11.5, and 19.4, respectively[7].

The aim of our research is to develop models which effectively predict colorectal cancer survivability, and to use these models to compare factor importances between populations of different race. In this paper, we apply machine learning methods to predict 2-year survivability for Hispanic and white colorectal cancer patients, and compare factor importances in determining survivability. These two groups were chosen since they were the most frequently appearing ethnic groups in our dataset. While Hispanic describes an ethnicity, and therefore

the two groups are not automatically disjoint by official race and ethnicity records, most self-described Hispanic people consider their Hispanic background as part of their racial background[8]. In this study, we consider a patient white if their race encoding is white and their Hispanic origin encoding labels them as non-Hispanic.

We created a separate dataset for each race and trained classifiers to predict survivability on the separated data. The trained models provide importance rankings of the patient factors (age at diagnosis, tumor size, primary site, etc.), which allow factor comparison between Hispanic and white patients. We also apply imbalanced classification techniques to improve the performance of our models.

This paper is organized as follows: Section 2 provides background information about the dataset, data processing procedures, models used, and performance metrics. Section 3 details our methodology. Results are reported in Section 4, and we discuss our findings in Section 5 before concluding with Section 6.

2. Background and Methodology

Machine Learning Pipeline

Our machine learning pipeline consist of the distinct stages of preprocessing, model training, and evaluation. Preprocessing is everything that is done to convert raw data into a format that can be used as input to our models. The models that we select are trained on the cleaned data, and then they are evaluated by a variety of metrics. The following sections provide further details on each stage.

Data Source

The diagnostic data used in our research was obtained from the Surveillance, Epidemiology, and End Results (SEER) program database[9]. Sponsored by the National Cancer Institute, SEER is the principal repository for cancer incidence and survival data in the United States. The SEER database contains over 10 million diagnostic records from 1973-2015. These records are collected from cancer registries across the country, and include data from over ten types of cancer. We aggregate patient data from all geographical registries, and filtered our data to only include white and Hispanic patients. The mixed dataset is simply formed by combining the Hispanic and white datasets. Table 1 shows the survivability distribution by ethnicity.

Ethnicity	Not Survived	Survived	Total
White	53343 (19.4%)	222204 (80.6%)	275547
Hispanic	6762 (18%)	30813 (82%)	37575

Table 1: Dataset label distribution by ethnicity

Data Preprocessing

In this study, we used SEER data from the years 2004-2015, for a couple of reasons. There was a major variable recode in the data beginning in 2004 which may have compromised data consistency if we had used patient data both prior and after 2004. Additionally, using relatively recent data is ideal because medical technology is continuously developing and improving, resulting in more timely, accurate diagnoses and better treatments. We used colorectal cancer patient data from all available registries.

Raw SEER data is encoded in pure ASCII files. These files were converted via Python scripts to a tabular format, from which we created two distinct datasets: one containing only Hispanic patients, the other containing only white patients. This separation is done prior to normalization and imputation so that those modifications are specific to each race.

In this paper, we will refer to instances and predictions where a patient survives beyond two years as positive, and instances and predictions where a patient dies of colorectal cancer before two years as negative. Accordingly, an instance is labeled negative if the survival time in months field is less than 24, and the cause of death is colon or rectum cancer. All instances not satisfying those two conditions are labeled positive.

Of the 133 features provided by SEER, we selected the 20 features detailed in Table 2 based on feature importances reported in the literature[3, 5]. We selected 16 categorical features and 4 continuous features. Each feature was carefully examined for outliers, missing values, and other inconsistencies. For the age diagnosed, number of malignant tumors, and months survived features, we deleted entire instances if the feature was missing. For the tumor size and number of positive nodes features, where there were a large proportion of missing or unknown values, we employed the MICE imputation technique, described shortly. It is worth noting that many other related studies cited in this paper have dropped large proportions of instances with missing data. However, we wanted to minimize instance eliminations because the SEER dataset did not have abundantly many Hispanic patient instances to begin with.

Categorical Variable	Unique values
Marital status	7
Sex	2
Primary site	13
Histology	139
Behavior	2
Grade	5
Diagnostic conf.	8
Extension	65
Lymph Nodes	18
Metastasis	25
Tumor size eval.	7
Node eval.	7
Metastasis eval.	7
Surgery site	34
Reason no surgery	8
Summary stage	5
Registry	18

Continuous Variable	Mean	S.D.
Age diagnosed	68.3	14
Positive nodes	1.57	4.26
Number of tumors	1.4	0.717
Tumor size	43.1	37.3

Table 2: Features used in models

In order to reconcile categorical and continuous features, we used a one-hot encoding scheme to binarize the categorical features. For each categorical feature, one-hot encoding creates a new binary indicator variable for every possible value of the original feature.

MICE Imputation

MICE is among the most common strategies for handling missing values in electronic health data[10]. MICE initially fills all missing values with the mean or median of the corresponding non-missing features, and then iteratively cycles through each feature requiring imputations and performs regression by treating all other features as independent variables. Algorithm 1 provides pseudocode for the MICE algorithm.

Algorithm 1 MICE Imputation

```
function MICE(Data, numcycles)
    Perform mean/median imputation on all missing values in dataset
    for i in numcycles do
        for feature f_m in set of features with any missing values do
            Regression(Data - column_{f_m}, column_{f_m})
            Replace originally missing features in column_{f_m} with regression result
        end for
    end for
end function
```

Models

We selected the following classifiers to predict survivability. Models were selected based on preliminary test performance as well as models previously reported in literature.

- *Logistic Regression* Logistic Regression is a classifier that evaluates the weighted sum of the input components, and applies the sigmoid function to the weighted sum. The output from the sigmoid function can be interpreted as the probability of the positive class in the binary classification scenario. The weights are learned during training.

- *Random Forest* A Random Forest is an ensemble of decision trees, where each tree splits nodes based on random features (instead of best feature)[11]. The final output class is the class which received the majority vote over the individual trees. Randomly determining which features to split on reduces overfitting because there is no guarantee that a certain feature will be included in each model.

- *AdaBoost* AdaBoost[12] is an ensemble of sequentially trained classifiers. Each instance in the training set is initialized with equal weight. Misclassified instances are given increased weight while training the next classifier. Prediction consists of a weighted prediction on all the classifiers, where each is weighted by its accuracy during training. In this study, the decision stump is used as the base classifier.

- *Neural Network* A neural network is composed of layers of nodes with interconnected weights. Each node receives a weighted sum of the outputs from all the nodes in the previous layer. An activation function is applied to the weighted sum, and the result is the output of the node. The weights in each layer are determined by the backpropagation training method, which updates each weight based on its influence on the loss function. Neural networks are at the forefront of recent deep learning progress[13].

- *Imbalanced Classification* We have additionally employed the imbalanced classification techniques of weighted classes and undersampling to improve the performance of the above models. Weighted classes penalizes the model more harshly for mislabeling an instance of the minority class. Undersampling only uses a fraction of the majority class instances, to make the proportion between class instances more balanced[14].

Evaluation Metrics

As our class labels are imbalanced, the conventional metric of accuracy is an inadequate indicator of model performance: A model which always predicts that the patient survives would achieve over 80% accuracy. The most common evaluation metric in the literature is Area Under Curve (AUC) of the ROC curve. The AUC score is intuitively interpreted as the probability that, given a positive and a negative instance, the positive instance is ranked more likely to be positive than the negative instance. The AUC score can be calculated given a set of predicted labels and the corresponding true labels. We employ AUC as the primary evaluation metric so that our results can be more comparable to results reported in the literature.

Additionally, we use the G-mean as an indicator for our performance on an imbalanced classification problem. To define the G-mean, we first define the following four metrics:

TP (true positive): the number of positive instances labeled positive

TN (true negative): the number of negative instances labeled negative

FP (false positive): the number of negative instances labeled positive

FN (false negative): the number of negative instances labeled negative

The confusion matrix shown in Table 3 summarizes the definitions.

		True Label	
		Not Survived	Survived
Predicted	Not Survived	TN	FN
	Survived	FP	TP

Table 3: The confusion matrix groups instances by the combination of their true and predicted labels.

The G-mean is defined as follows:

$$\text{G-mean} = \sqrt{\text{sensitivity} \cdot \text{specificity}}$$

where sensitivity and specificity are defined as:

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad \text{specificity} = \frac{TN}{TN + FP}$$

The G-mean is a viable metric for imbalanced classification because it weights the rate of correctly predicting the majority and minority classes equally.

3. Results

We trained our models with Scikit-learn, an open-source Python machine learning library. We used Scikit-learn version 0.19.1 and Python 3.6.4. All experiments were executed on a Linux machine with an Intel i7-3700 3.4 GHz processor and 16GB RAM running Ubuntu 18.04 LTS.

We used 5-fold cross-validation to determine the best hyperparameters for each model, and recorded results on the test sets, each composed of a randomly-selected 10 percent of the data. All code used is in our GitHub repository[16]. Table 4 details the AUC scores for each model and dataset. The AUC scores we achieved with our neural networks are above the best reported in literature of 0.8675 by[4], which was also produced by neural networks. Our AUC scores are calculated on each model's prediction of the probability that the true label is positive. This will provide a higher AUC score because wrong predictions are not penalized as heavily if their probability prediction was closer to 0.5.

Model	**Hispanic**	**White**	**Mixed**
Logistic Regression	.859	.872	.87
Random Forest	.855	.865	.849
AdaBoost	.859	.871	.859
Neural Network	.873	.875	.856

Table 4: AUC scores for trained models on the test set. The highest reported AUC for 2-year colorectal cancer survivability had been 0.8675[4]. Our neural network architecture had 3 fully-connected hidden layers of 400 neurons, ReLU activation, 0.1 dropout between each layer besides input layer, with sigmoid output activation.

Table 5 ranks the top seven factors for each model and dataset. Factor models were obtained from the coefficient and parameter values of the trained models, in the context of each model. In the logistic regression model, a larger magnitude coefficient for a feature indicates that the feature has relatively greater impact on the label prediction. For decision tree based models like random forest and AdaBoost, we use Gini impurity and entropy scores for each feature, which reveal the feature that provide the most information when a decision tree node is split on that feature. Since our models are trained on one-hot encoded features, we sum the feature importance values for all the one-hot encoded columns corresponding to a single original feature (histology, surgery site, etc.) to obtain the score used to rank

feature importances. Scikit-learn allows access to model parameters, and also includes class functions which calculate factor rankings.

Logistic Regression		**Random Forest**		**AdaBoost**	
Hispanic	**White**	**Hispanic**	**White**	**Hispanic**	**White**
Histology	Histology	Metastasis	Metastasis	Extension	Extension
Extension	Lymph node inv.	Stage	Stage	Histology	Age
Metastasis	Extension	Age	Age	Age	Histology
Surgery site	Surgery site	No surg. reason	No surg. reason	Tumor size	Positive nodes
Diagnostic conf.	Metastasis	Positive nodes	Positive nodes	Positive nodes	Metastasis
Lymph node inv.	Diagnostic conf.	Surgery site	Surgery site	Metastasis	Tumor size
No surg. reason	Primary site	Tumor size	Tumor size	Surgery site	Surgery site

Table 5: Factor rankings for each model and ethnic population

Table 6 shows the improvements made by using the weighted classes and undersampling imbalanced classification methods. These methods were applied to the datasets prior to model training. Our results are comparable to the G-mean score of 0.792 reported by Al-Bahrani and Agrawal[4].

Model	**Hispanic**	**White**
Logistic Regression	.628	.683
Weighted Logistic Regression	.783	.8
Undersampled Logistic Regression	.79	.8
Random Forest	.623	.631
Weighted Random Forest	.782	.775
Undersampled Random Forest	.787	.796

Table 6: G-mean scores for selected models and ethnic populations. We undersampled until we had an equal number of both labels, and we weighted the minority class five times as the majority class

4. Discussion

Our results provide insight into the importance of patient ethnicity in cancer survivability prediction, as well as the cancer survivability prediction problem itself. From the AUC score table 4, we observed that models trained on mixed-ethnicity patient data performed consistently worse than those trained on single-ethnicity data. Logistic regression was the

only model in which the mixed dataset yielded better results either the Hispanic or white datasets. The white dataset had significantly better results than the mixed dataset for all models, while the Hispanic dataset had slightly better results. This is very likely due to the Hispanic dataset being far smaller than the white dataset, rather than the Hispanic dataset being intrinsically harder to train on.

Our factor rankings results show fairly consistent variable rankings between the Hispanic and white datasets for the same model, with minor differences. For example, lymph node involvement is the second most important factor for white patient instances in the logistic regression model, but the sixth most important factor for Hispanic patient instances. It is expected that there are not dramatic differences in factor rankings between the two groups. However, even minor differences can be of use in the analysis of a patients diagnostic data; for example, by giving an observation of a feature more consideration for a patient with a certain race. Additionally, we observe in table 5 that different models have widely varying factor rankings. This confirms that many factors are not independent, and that there are many viable ways to make survivability predictions with competitive results. We note that neural networks are unable to provide factor rankings due to the non-transparent nature of the model.

From the imbalanced classification results, we observe that the logistic regression and random forest models both perform better with respect to the G-mean metric with imbalanced classification techniques. Both weighted classes and undersampling improved the G-mean significantly. Applying these techniques to the AdaBoost and neural network model achieves similar results(appendix Table 7). The same table shows that G-mean scores greatly improved, while AUC scores slightly increased after applying undersampling to AdaBoost, and slightly decreased after applying undersampling to a neural network. Applying imbalanced classification methods demonstrates the trade-off between sensitivity and specificity, and the inherent hardness of this classification problem. In the original models sensitivity and specificity values were around 0.9 and 0.4, respectively. After applying either weighted classes or undersampling, sensitivity and specificity values were both close to 0.8.

The trade-off implies that instances with similar feature values have different labels. In some instances this is due to both actual survival times being close in value but on opposite sides of the 24-month cutoff. In other instances, survival times can actually differ widely, suggesting that we do not have all the relevant information. Indeed, this is the case: we

only have the colorectal cancer diagnostic information and basic personal information. We hypothesize that more extensive data which includes factors like personal health history and lifestyle information would improve performance beyond the best currently reported in literature. However, there is currently no dataset with such degree of patient information.

5. Conclusion

In this study we applied machine learning methods to predict 2-year colorectal cancer survivability. Specifically, we trained logistic regression, random forest, AdaBoost, and neural network models on SEER data we carefully preprocessed. Our models achieved AUC score of 0.87, improving upon the best-reported score in the literature [REF]. We used our models to investigate the impact of ethnicity on model performance, as well as diagnostic factor rankings. Our models achieved superior performance on single-ethnicity patient data compared to the mixed data, and we also observed minor differences in factor importance. Finally, we used the imbalanced classification methods of weighted classes and undersampling to improve the G-mean score to approximately 0.8.

We infer from our results that cancer survivability prediction is a difficult classification problem, and that the SEER dataset may not contain enough information to significantly improve the performance of our current models. Future work includes extending our focus on ethnicity beyond the Hispanic and white patient groups, as well as to other types of cancer.

References

[1] National Cancer Institute, Cancer stat facts: Colorectal cancer, `https://seer.cancer.gov/statfacts/html/colorect.html`, accessed: 2018-07-11.

[2] American Cancer Society, Survival Rates for Colorectal Cancer, by Stage, `https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html`, accessed: 2018-07-11.

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

[4] R. Al-Bahrani, A. Agrawal, A. Choudhary, Survivability prediction of colon cancer patients using neural networks, Health informatics journal (2017) 1460458217720395.

[5] R. Al-Bahrani, A. Agrawal, A. Choudhary, Colon cancer survival prediction using ensemble data mining on seer data, in: Big Data, 2013 IEEE International Conference on, IEEE, 2013, pp. 9–16.

[6] Y.-Q. Liu, C. Wang, L. Zhang, Decision tree based predictive models for breast cancer survivability on imbalanced data, in: Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on, IEEE, 2009, pp. 1–4.

[7] SEER Cancer Statistics Review 1975-2015, Age-adjusted seer incidence rates and trends for the top 15 cancer sitesa by race/ethnicity, https://seer.cancer.gov/csr/1975_2015/results_merged/topic_race_ethnicity.pdf, accessed: 2018-07-11.

[8] A. Gonzalez-Barrera, M. H. Lopez, Is being hispanic a matter of race, ethnicity or both?, http://www.pewresearch.org/fact-tank/2015/06/15/is-being-hispanic-a-matter-of-race-ethnicity-or-both/, accessed: 2018-07-08.

[9] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2015), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission.

[10] M. J. Azur, E. A. Stuart, C. Frangakis, P. J. Leaf, Multiple imputation by chained equations: what is it and how does it work?, International journal of methods in psychiatric research 20 (1) (2011) 40–49.

[11] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[12] Y. Freund, R. Schapire, A short introduction to boosting, Journal-Japanese Society For Artificial Intelligence 14 (771-780) (1999) 1612.

[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436.

[14] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge & Data Engineering (9) (2008) 1263–1284.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[16] S. Li, Personalized colorectal cancer survivability prediction, https://github.com/samuelli97/cancer_survivability (2018).

	Hispanic		White	
Model	**G-mean**	**AUC**	**G-mean**	**AUC**
AdaBoost	0.627	0.859	0.671	0.859
Undersampled AdaBoost	0.787	0.865	0.798	0.871
Neural Network	0.6	0.873	0.7	0.875
Undersampled Neural Network	0.788	0.868	0.796	0.873

Table 7: G-mean scores of AdaBoost and neural networks, as well as G-mean and AUC comparisons for imbalanced classification.