Executive Summary - Cancer Survivability Project

Ray Lee Samuel Ogunfuye Karla Paulette Flores Silva Dilruba Sofia
Funmilola Mary Taiwo Enayon Sunday Taiwo

June 1, 2024

1 GitHub

May 2024 Cancer Survivability

2 Overview

Our project focuses on predicting the survival outcomes of cancer patients by analyzing a range of clinical and demographic factors. Employing diverse classification algorithms, we aim to identify the most effective model for accurate predictions.

3 Background and Goal

Breast cancer is when abnormal cellular growth happens in the breast tissue [1, 2]Breast cancer is the second most frequently diagnosed and is the leading cause of cancer death among women [1, 2]. In 2022, globally there were 2.3 million cases and 670000 death due to breast cancer [1]. While breast cancer generally occur among women, about $0.5 - 1\%$ cases include men [2]. Surgical resection, radiation therapy, and pharmaceutical treatment are the most common type treatment for breast cancer for both gender [2]. Since breast cancer has devastating consequences, it is important for healthcare providers to provide the best care to patients, insurance companies to make reimbursement decisions, government organizations to plan resource allocations, and patients and their families to make decisions in their care. One way to help these stakeholders in these procedure to make decision is to know the prognoses of patients. Our goal for this project is to find a good classification algorithm to predict patients' vital status based on their demographics and tumor characteristics.

4 Data

The breast cancer data by the TCGA-BRCA project is collected from the GDC data portal [3]. We take the open-access clinical information of the patients who have primary breast tumors. The JSON data file contains the patients' information such as demographics, tumor staging, and treatment type whether pharmaceutical or radiation therapy that patients received. It comprises 1006 observations with 42 variables.

5 Data Preprocessing

5.1 Cleaning

Data format: JSON file with datasets such exposure, demographics, and diagnoses type as dictionary entries.
Process:

- Extract exposure, demographics, and diagnoses files from the main data file since they series of separate dictionary entries and combine them.
- In the diagnoses file, treatment series is stored by two types; pharmaceutical treatment and radiation therapy. In other words, we have information about whether each patient received pharmaceutical treatment or radiation therapy.
- We drop the uninformative and redundant columns such as days to diagnosis, gender, site of resection or biopsy, last known disease status, morphology, synchronous malignancy, tissue or organ of origin, state, submitter ID, classification of tumor, ICD 10 code, tumor grade, progression or recurrence, demographic ID, updated date & time, age at diagnosis, days to birth, and year of death.
- We update less than zero values of days to last follow-up bu days to death and drop days to death column.
- Some patients who did not AJCC pathologic stage, we recover them based on AJCC pathologic Tumor size, Node, and Metastasis site (TNM) grading. For 15 patients there was no way to recover pathologic stage. Thus, we drop them from the data.
- We also add pathologic sub-stage information based on TNM grading, and further drop three samples that had no option to be sub-staged.
- We then replace pathologic stages by ordinal values in range $1 - 8$.
- We TNM sate columns and case ID, and the remaining columns are now AJCC pathologic stage, days to last follow up, primary diagnosis, prior malignancy, year of diagnosis, prior treatment, pharmaceutical treatment, radiation, ethnicity, race, vital status, age at index, and year of birth.
- The final dataset has 991 observations and 13 features.

5.2 Exploratory Data Analysis:

First, we plot the histograms of categorical features against each vital status, alive or deceased (see figure 1). It can be seen that in all categories number of deceased patients are significantly less as the the total number of deceased patient in this study is much less than alive patients.

Secondly, we plot box-plots for the features AJCC pathologic stage, days to last follow-up, year of diagnosis, pharmaceutical treatment, radiation therapy, age at index, and year of birth against vital status. We notice that there are relatively higher number deceased patients with later pathologic stage and older age (see figure 2).

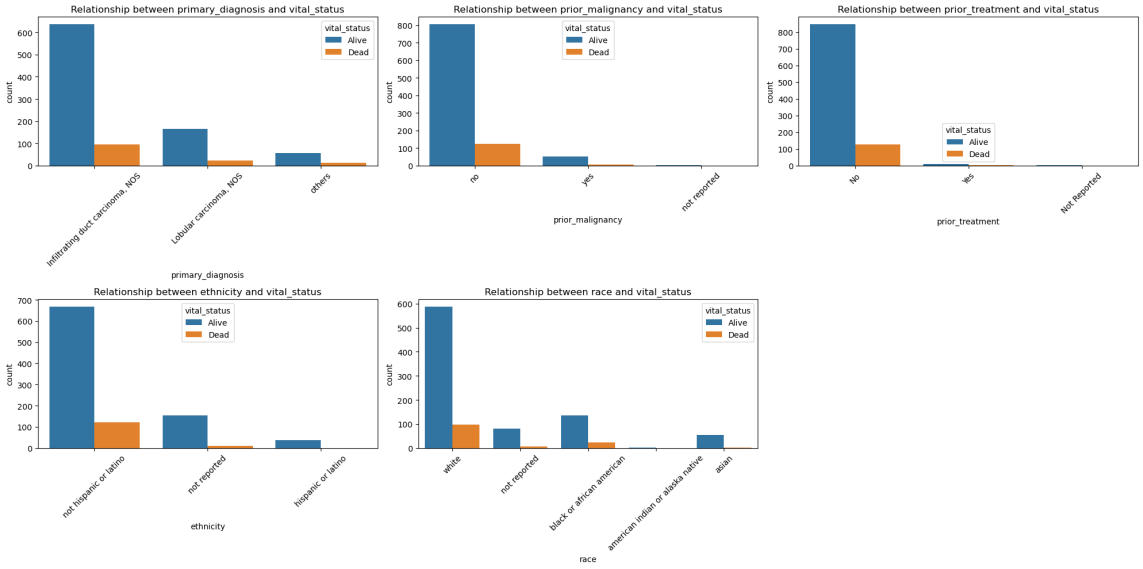


Figure 1: Histograms of categorical features primary diagnosis, prior malignancy, prior treatment, ethnicity, and race plotted against alive or deceased vital status.

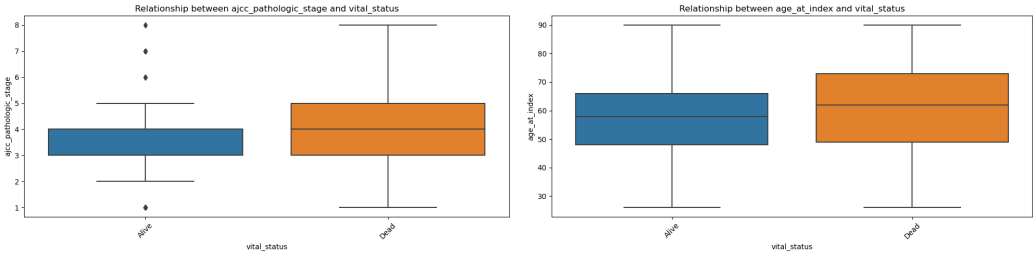


Figure 2: Box-plots of AJCC pathologic stage and age at index of patients against their vital status.

6 Methods and Results

6.1 Classifying patient outcome

Our first goal was to develop classifiers to predict patient outcome as "alive" or "deceased." We implement Random Forest, AdaBoost, Support Vector Classifier, Logistic Regression, K-nearest neighbors, and Multilayer Perceptron classification algorithms, and compare the performance of these algorithms against the Decision Tree algorithm.

Our data ($n = 991$) is split into a training set ($n = 772$) and test set ($n = 219$). 3-fold cross-validation is performed using the training data. We compute model accuracy, precision, sensitivity, selectivity, false positive rate, and false negative rate with respect to each validation set. Here, a 'positive' result corresponds to an outcome of 'deceased'. We summarize average model performance for each metric below (Figs. 3 and 4).

	accuracy	precision	sensitivity	selectivity	false positive	false negative
rf	0.892677	0.652233	0.419067	0.965067	0.034933	0.580933
adaboost	0.892677	0.665167	0.390500	0.969433	0.030567	0.609500
svc	0.891414	0.911100	0.209500	0.995633	0.004367	0.790500
logreg	0.901515	0.832800	0.323800	0.989800	0.010200	0.676200
knn	0.892677	0.925000	0.209500	0.997067	0.002933	0.790500
mlp	0.887626	0.607900	0.438067	0.956333	0.043667	0.561933

Figure 3: Average validation metrics

	accuracy	precision	sensitivity	selectivity	false positive	false negative
Decision Tree	0.848485	0.436533	0.533333	0.896667	0.103333	0.466667

Figure 4: Average validation performance of Decision Tree model

Since it is important to identify patients who are at increased risk of mortality, model sensitivity is a particularly significant metric. The most sensitive models were generated by the Decision Tree, MLP, and Random Forest algorithms. However, none of our models were very sensitive to identifying patients whose eventual outcome was 'deceased'. This suggests that additional features are needed to identify patients who are most at-risk.

Taking account all performance metrics, we decided to develop the Random Forest model, which was relatively sensitive with respect to the other models, while also being significantly more accurate, precise, and selective than the base model. After tuning model parameters via grid search, we evaluate the Random Forest Model on test data, and obtain the following result (Figs. 5):

	accuracy	precision	sensitivity	selectivity	false positive	false negative
Tuned Random Forest	0.924623	0.8235	0.5385	0.9827	0.0173	0.4615

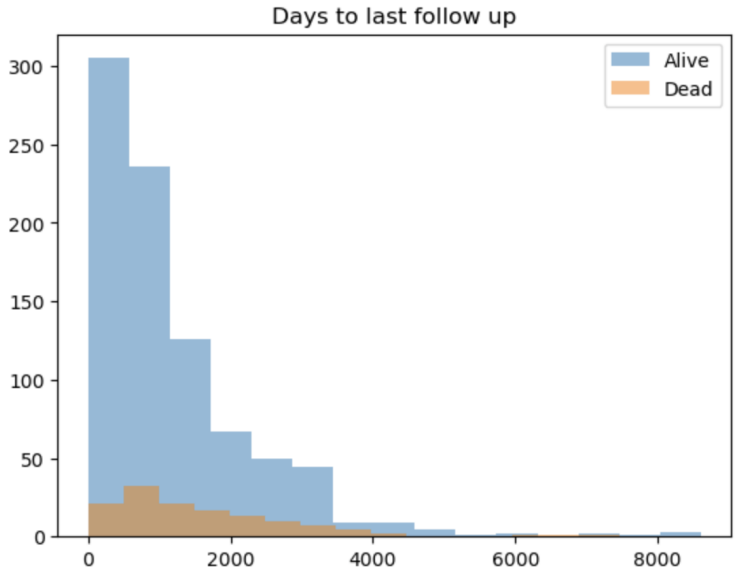
Figure 5: Average validation performance of tuned Random Forest model

6.2 More experiments..

Having build different models to predict survival state, we come to realization about the need of a time frame context for this prediction. Then, we build customized models for given time lapses.

To achieve this, we followed the workflow as mentioned:

1. The feature *'days to last follow up'* describes the time span the patients were in contact with medical care. The distribution of it helps us decide the timeframes to be considered



- Mean of alive: 1196.89
- Mean of dead: 1622.59

2. We designed models that predicts survival state within the first n years, where $n = 1, 2, 3, 4, 5..$ for our experimentation. Models based on Support Vector Machine and Random Forests.

3. We filter out the patients whose time span of follow ups is less than n years since it is uncertain from the data if the patient survived after this time. Therefore, models were trained in the rest of the data.

	ajcc_pathologic_stage	days_to_last_follow_up	primary_diagnosis	prior_malignancy	year_of_diagnosis
0	1	337.0	Infilt...	no	2010.0
1	3	5.0	Infilt...	no	2010.0
2	1	759.0	Lobula...	yes	2012.0
3	5	954.0	Lobula...	yes	2010.0
4	8	304.0	Infilt...	no	2009.0

Figure 6: Patients who were into follow ups with clinic less than 1 year are marked in red.

4. From the subset of data that is valid for the model, the feature *'days to last follow up'* for a dead patient was larger than the time frame into consideration, survival status is modified:

$$\textit{'Dead'} \rightarrow \textit{'Alive'}$$

since he/she survived during the time lapse of interest.

5. Best model for survival predictions in the first n years after diagnosis is: Support vector machine for $n = 1, 2, 3, 4, 5$ according to all metrics accuracy, precision and recall.

   SVC for prediction of survival within the first n years:

1 year	$accuracy : 0.97$	$precision : 0.00$	$recall : 0.00$
2 years	$accuracy : 0.98$	$precision : 1.00$	$recall : 0.80$
3 years	$accuracy : 0.97$	$precision : 1.00$	$recall : 0.80$
4 years	$accuracy : 0.98$	$precision : 1.00$	$recall : 0.85$
5 years	$accuracy : 0.96$	$precision : 1.00$	$recall : 0.91$

7 Discussion

7.1 Stakeholders

Our prediction model on vital status can benefit different set of stakeholders including patients:

- Healthcare providers would be able to decide best care for each patient.
- Insurance companies can decide pricing and reimbursement options.
- Government and patient support organization would be able plan resource allocation.
- Patients and their families make decision on treatment options.

7.2 Limitations

- We did not consider all types of data available TCGA-BRCA project. With inclusion of gene expression, immune cell abundance, and slide images, we might be able to better our model.
- We also did not consider other types or advanced data engineering and models that could potentially make the prediction precise.

7.3 Future Directions

- Include gene expression and comprehensive cellular abundance data, and slide images in the model.
- Explore more advanced algorithm that could potentially give better accuracy and precision.
- Check whether this model is transferable to other types of cancer and make appropriate extension.
- Make a web-based interface which will take patients information as input and output the predicted vital status or survival months.

Acknowledgements

We would like to thank our project mentor Soumen Deb and our instructors at the Data Science Boot Camp at The Erdós Institute.

References

[1] World Health Organization (2024). Breast Cancer. Retreived from https://www.who.int/news-room/fact-sheets/detail/breast-cancer (Accessed on May 31, 2024).

[2] National Cancer Institute (2024). Breast Cancer - Patient Version. Retrieved from https://www.cancer.gov/types/breast (Accessed on: June 1, 2024).

[3] Thennavan A, Beca F, Xia Y, Recio SG, Allison K, Collins LC, Tse GM, Chen YY, Schnitt SJ, Hoadley KA, Beck A, Perou CM. Molecular analysis of TCGA breast cancer histologic types. Cell Genom. 2021 Dec 8;1(3):100067. doi: 10.1016/j.xgen.2021.100067. PMID: 35465400; PMCID: PMC9028992.