Survival Predication for Pediatric Brain Cancer Patients

By Teresa Cameron

Table of Contents

01

About the Patients

Data Introduction

Introduction

Brain cancer is the 2nd most common type of childhood cancer and the most deadliest.

SEER Cancer Data

Total Patients

331,449,281
Total Cancer Patients

Variables

263 Variables available for analysis

Data Collection

SEER has been funded by the NCI since 1973. It collects cancer data that covers around 48% of the U.S. population.

Patients Included in Analysis

Total Number of Patients

12,164

X-Variables Considered

Age
Race
Gender
Primary Site
Tumor Type
Grade
Laterality
Median Household Income
Rural Urban Code

Y-Variable

Patient Survival

101 Patients removed due to unknown cause of death

Years

2000 - 2019

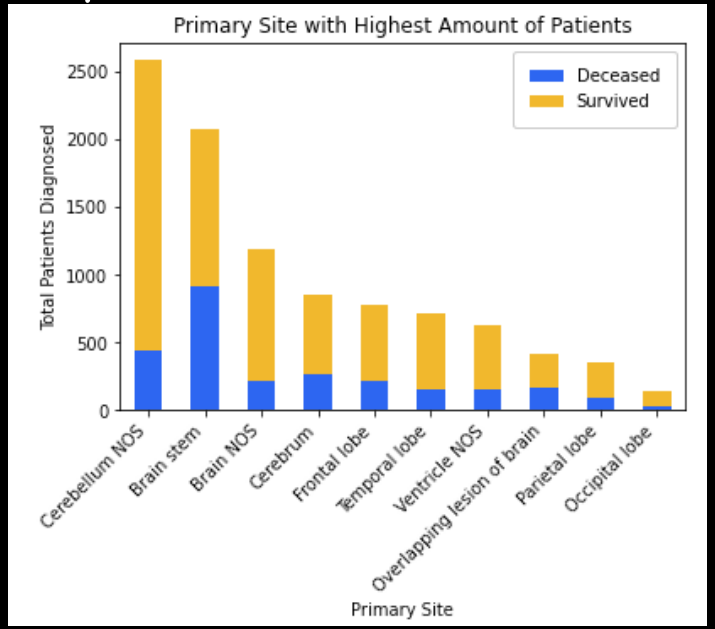
Year Diagnosed

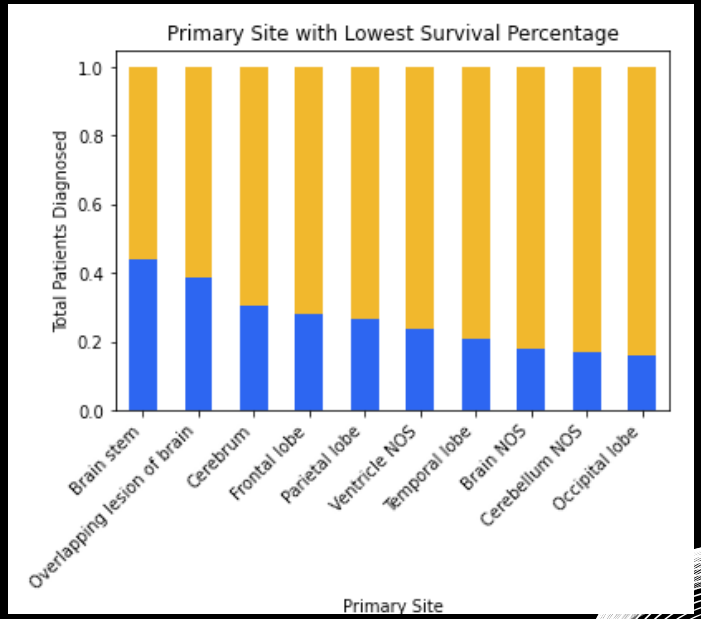
	Deceased	Survived
2000	32.59%	67.41%
2001	35.40%	64.60%
2002	35.60%	64.40%
2003	30.02%	69.98%
2004	31.09%	68.91%
2005	32.97%	67.03%
2006	29.45%	70.55%
2007	31.08%	68.92%
2008	31.05%	68.95%
2009	26.81%	73.19%
2010	25.12%	74.88%
2011	28.50%	71.50%
2012	24.75%	75.25%
2013	30.64%	69.36%
2014	25.60%	74.40%
2015	24.52%	75.48%
2016	23.57%	76.43%
2017	21.29%	78.71%
2018	15.92%	84.08%
2019	5.42%	94.58%

Reduced total Patients by removing those diagnosed before 2003 and after 2017.

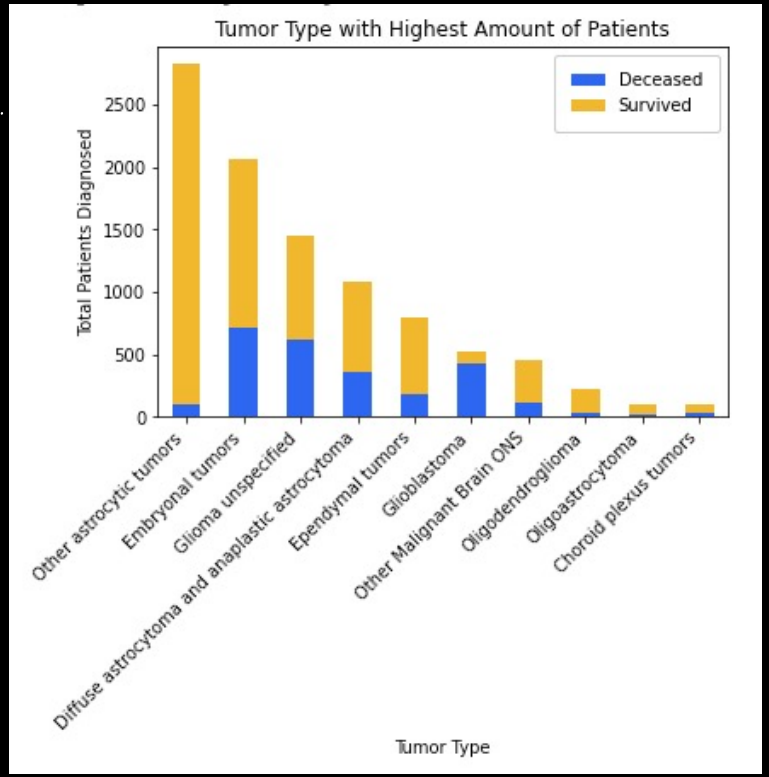
New total patients: 9,714

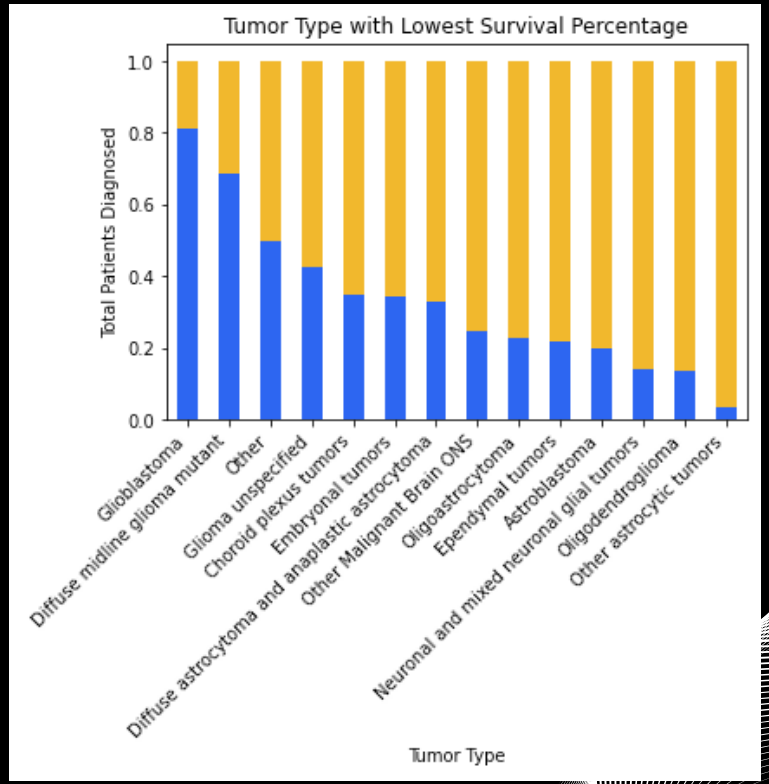
Primary Site





Tumor Type


Tumor Type with Highest Amount of Patients
Deceased
Survived
Total Patients Diagnosed
0
500
1000
1500
2000
2500
Other astrocytic tumors
Embryonal tumors
Glioma unspecified
Diffuse astrocytoma and anaplastic astrocytoma
Ependymal tumors
Glioblastoma
Other Malignant Brain ONS
Oligodendroglioma
Oligoastrocytoma
Choroid plexus tumors
Tumor Type


Tumor Type with Lowest Survival Percentage
Total Patients Diagnosed
0.0
0.2
0.4
0.6
0.8
1.0
Glioblastoma
Diffuse midline glioma mutant
Other
Glioma unspecified
Choroid plexus tumors
Embryonal tumors
Diffuse astrocytoma and anaplastic astrocytoma
Other Malignant Brain ONS
Oligoastrocytoma
Ependymal tumors
Astroblastoma
Neuronal and mixed neuronal glial tumors
Oligodendroglioma
Other astrocytic tumors
Tumor Type

Race Origin Recode

	Deceased	Survived
Non-Hispanic Asian or Pacific Islander	33.05%	66.95%
Non-Hispanic Black	31.93%	68.07%
Hispanic (All Races)	30.87%	69.13%
Non-Hispanic American Indian/Alaska Native	24.32%	75.68%
Non-Hispanic White	23.91%	76.09%

Removed 102 Non-Hispanic Unknown Race patients as their survival rate was 96.08%.

Grade

Unknown	66.052851
Undifferentiated; anaplastic; Grade IV	14.180191
Blank(s)	6.377445
Moderately differentiated; Grade II	5.815647
Well differentiated; Grade I	5.742821
Poorly differentiated; Grade III	1.831045

Removed Grade variable because majority of patients had missing information.

Age

Created bins for patients age ranging based on quartiles.

	Deceased	Survived
1	30.63%	69.37%
2	30.07%	69.93%
3	24.02%	75.98%
4	23.29%	76.71%

0-4

5-8

9-13

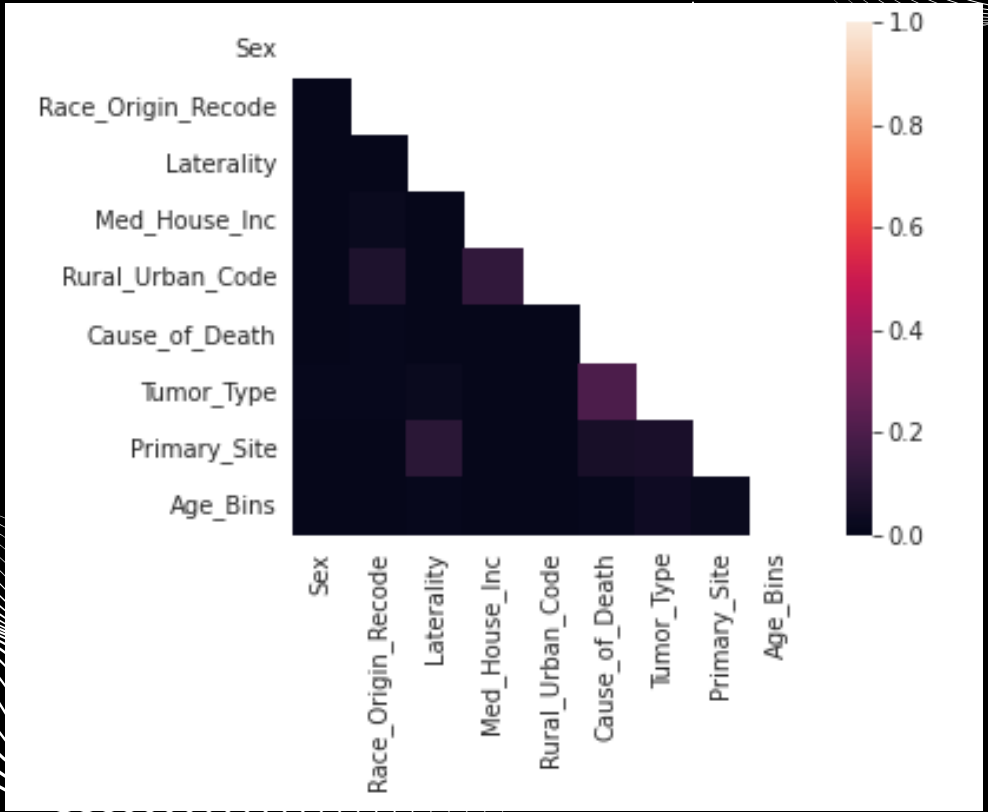
14 - 19

Total Patients for Machine Learning: 9,611
Total X-Variables: 8

02

Machine Learning

Cramer's V Correlation Matrix



Split training and testing data based on Race

4	0.553822
0	0.257930
3	0.106604
2	0.073843
1	0.007800

Non-Hispanic White
Hispanic (All Races)
Non-Hispanic Black
Non-Hispanic Asian or Pacific Islander
Non-Hispanic American Indian / Alaska Native

Balance Y Variable in Training Set

Survived: 5,608
Deceased: 2,080

Final Training Set:
11,216 rows
59 columns

Final Testing Set:
1,912 rows
59 columns

Models

Classification Models Selected

Logistic Regression
KNN
Decision Tree
Random Forest
SVC
Gaussian NB

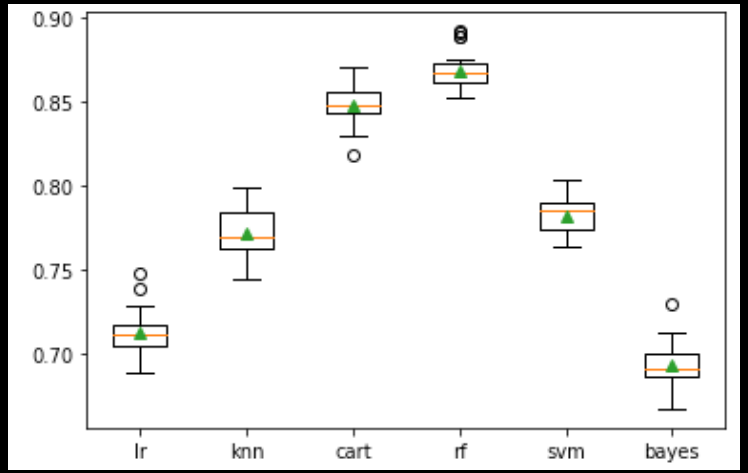
Stratified K Fold with 10 splits

03

Model Evaluation

Evaluation of Standalone Models on Training Data

Findings



Random Forrest had the highest accuracy on the training data.

Accuracy

```
lr 0.713 (0.013)
knn 0.772 (0.014)
cart 0.849 (0.011)
rf 0.869 (0.010)
svm 0.783 (0.011)
bayes 0.693 (0.013)
```

Stacking Classifier for Prediction

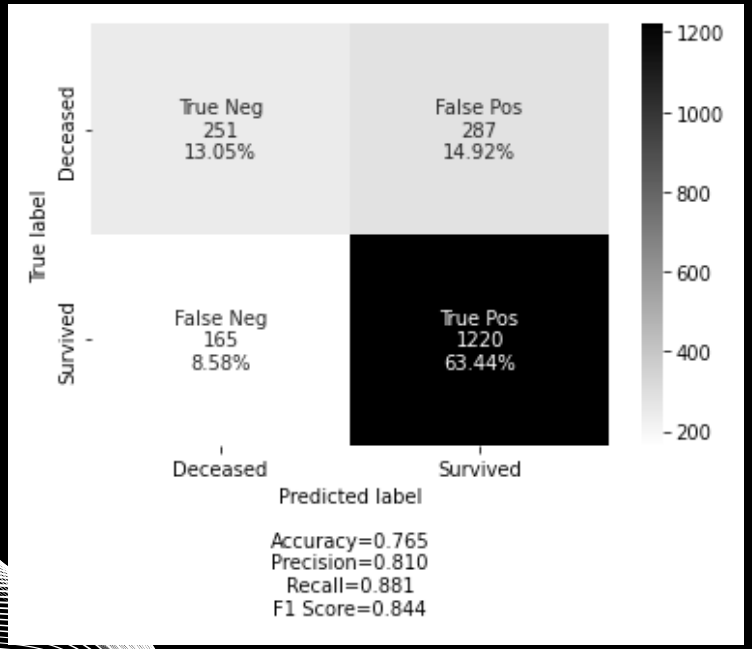
The stacking classifier function allows the strength of many classifiers to make a prediction. The output of each estimator is used as input of a final estimator.

04

Results

Evaluate predictions

Results



Of the 1,923 patients, 1,471 were correctly classified.

287 of the incorrectly classified were labeled as surviving and 165 as deceased.

The model was better at identifying patients who survived than not labeling deceased patients as survived.

Distribution of the Variables from Patients Incorrectly Classified

Gender

Male	0.551948
Female	0.448052

Race

Non-Hispanic White	0.517316
Hispanic (All Races)	0.294372
Non-Hispanic Black	0.114719
Non-Hispanic Asian or Pacific Islander	0.071429
Non-Hispanic American Indian/Alaska Native	0.002165

Rural Urban Code

Counties in metropolitan areas ge 1 million pop	0.608225
Counties in metropolitan areas of 250,000 to 1 million pop	0.192641
Counties in metropolitan areas of lt 250 thousand pop	0.086580
Nonmetropolitan counties adjacent to a metropolitan area	0.062771
Nonmetropolitan counties not adjacent to a metropolitan area	0.049784

Laterality

Not a paired site	0.727273
Left - origin of primary	0.121212
Right - origin of primary	0.121212
Paired site, but no information concerning laterality	0.010823
Paired site: midline tumor	0.008658
Bilateral, single primary	0.008658
Only one side - side unspecified	0.002165

Age Bins

Bins_1	0.400433
Bins_2	0.235931
Bins_4	0.183983
Bins_3	0.179654

Median Household Income

$75,000+	0.285714
$65,000 - $69,999	0.199134
$60,000 - $64,999	0.164502
$70,000 - $74,999	0.086580
$50,000 - $54,999	0.075758
$55,000 - $59,999	0.075758
$45,000 - $49,999	0.058442
$40,000 - $44,999	0.030303
$35,000 - $39,999	0.015152
< $35,000	0.008658

Tumor Type

Embryonal tumors	0.398268
Glioma unspecified	0.153680
Diffuse astrocytoma and anaplastic astrocytoma	0.127706
Ependymal tumors	0.101732
Other astrocytic tumors	0.058442
Other Malignant Brain ONS	0.045455
Glioblastoma	0.043290
Choroid plexus tumors	0.025974
Oligodendroglioma	0.015152
Diffuse midline glioma mutant	0.010823
Oligoastrocytoma	0.008658
Other	0.006494
Neuronal and mixed neuronal glial tumors	0.002165
Astroblastoma	0.002165

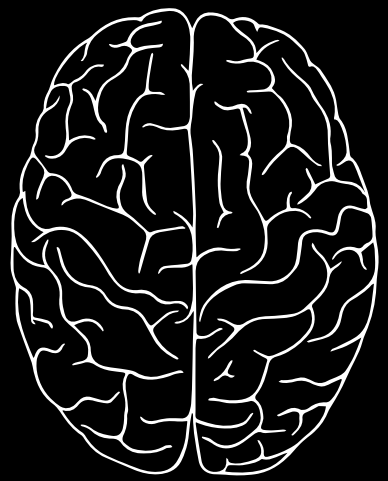
Primary Site

Brain stem	0.272727
Cerebellum NOS	0.264069
Brain NOS	0.114719
Cerebrum	0.080087
Ventricle NOS	0.077922
Overlapping lesion of brain	0.054113
Frontal lobe	0.051948
Parietal lobe	0.043290
Temporal lobe	0.030303
Occipital lobe	0.010823

05

Conclusion

Findings and Discussion



Discussion Summary

The predicted results could be improved by analyzing which features contributed to incorrect classification by the model. It would be best to examine why the model incorrectly identified patients as survived when they did not.

Further analysis of the data could be used to predict the survival months of the patients who were classified as deceased.
Additionally, it would be beneficial to see which treatments had the best success with the different patient variables.

Conclusion

The model performed well with identifying patients who were predicted to survive brain cancer. It would be beneficial to have a model that performed better to identify patients who are less likely to survive so that doctors could aggressively treat the patients and additional research and analysis could be allocated to those types of cancers.

References

Hossain MJ, Xiao W, Tayeb M, Khan S. Epidemiology and prognostic factors of pediatric brain tumor survival in the US: Evidence from four decades of population data. Cancer Epidemiol. 2021 Jun;72:101942. doi: 10.1016/j.canep.2021.101942. Epub 2021 May 1. PMID: 33946020; PMCID: PMC8142618.

Is it time to use machine learning survival algorithms for survival and risk factors prediction instead of Cox proportional hazard regression? A comparative population-based study. Sara Morsy, Truong Hong Hieu, Abdelrahman M Makram, Osama Gamal Hassan, Nguyen Tran Minh Duc, Ahmad Helmy Zayan, Le-Dong Nhat-Nam, Nguyen Tien Huy
medRxiv 2021.11.20.21266627; doi: https://doi.org/10.1101/2021.11.20.21266627

National Cancer Institute. (2022) Cancer Stat Facts: Childhood Brain and Other Nervous System Cancer (Ages 0-19). National Institutes of Health. https://seer.cancer.gov/statfacts/html/childbrain.html

Thank you!