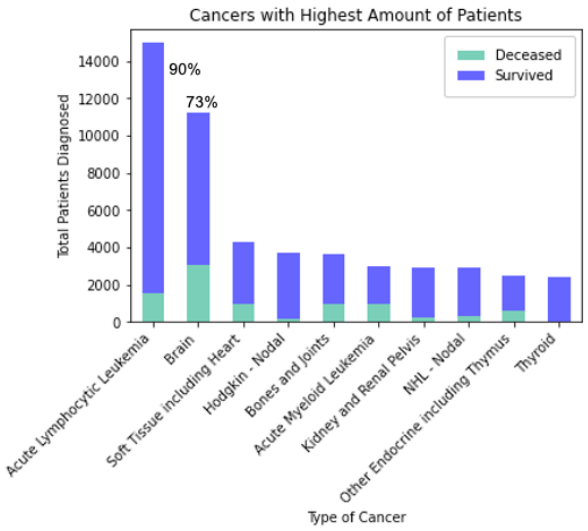
Survival Prediction of Pediatric Brain Cancer Patients
By Teresa Cameron

Introduction

Pediatric brain cancer is the 2nd most common type of brain cancer and the deadliest. Figure 1 shows the total number of pediatric cancer patients by type and the total amount by survival.

Bar Graph



Table

	Deceased	Survived
Acute Lymphocytic Leukemia	1556	13406
Brain	3051	8205
Soft Tissue including Heart	945	3331
Hodgkin - Nodal	131	3544
Bones and Joints	987	2613
Acute Myeloid Leukemia	942	2050
Kidney and Renal Pelvis	250	2671
NHL - Nodal	336	2573
Other Endocrine including Thymus	600	1852
Thyroid	14	2364

Figure 1

37% or 3,051 pediatric brain cancer patients have died from this cancer from 2000 – 2018. Advances in technology for treatment and detection have improved these outcomes over the years as shown in Figure 2.

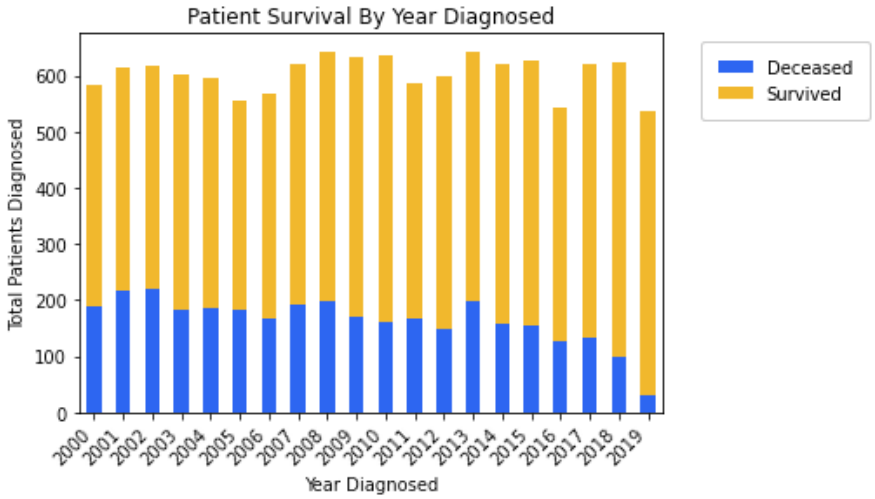


Figure 2

A study done in 2021 examined factors such as age at diagnosis, sex, race, primary site, tumor grade, histology, year of diagnosis, percentage of persons in county below poverty level, and geographic regions and if they influenced survival of pediatric cancer patients. The findings were published in a paper titled *Epidemiology and prognostic factors of pediatric brain tumor survival in the US: Evidence from four decades of population data.* The study found survival varies by tumor location, histology, age, race-ethnicity, and poverty. Children who are older when diagnosed have better survival outcomes and African American and Hispanic children are associated with higher mortality.

Patients have been documented with over 150 different brain tumors. The location and type of tumor influence the survival of the patients. Most of the patients have tumors that are located on the Cerebellum and develop from a type of glial cell called astrocytes. These are star shaped cell that support the nerve cells in the brain. Figures 3 and 4 show the breakdown of patients by primary tumor site and tumor type.

Figure 3

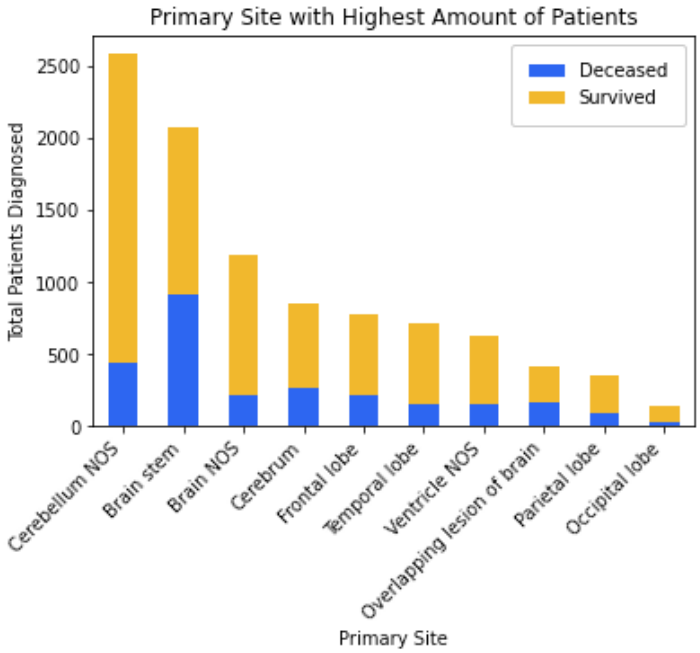
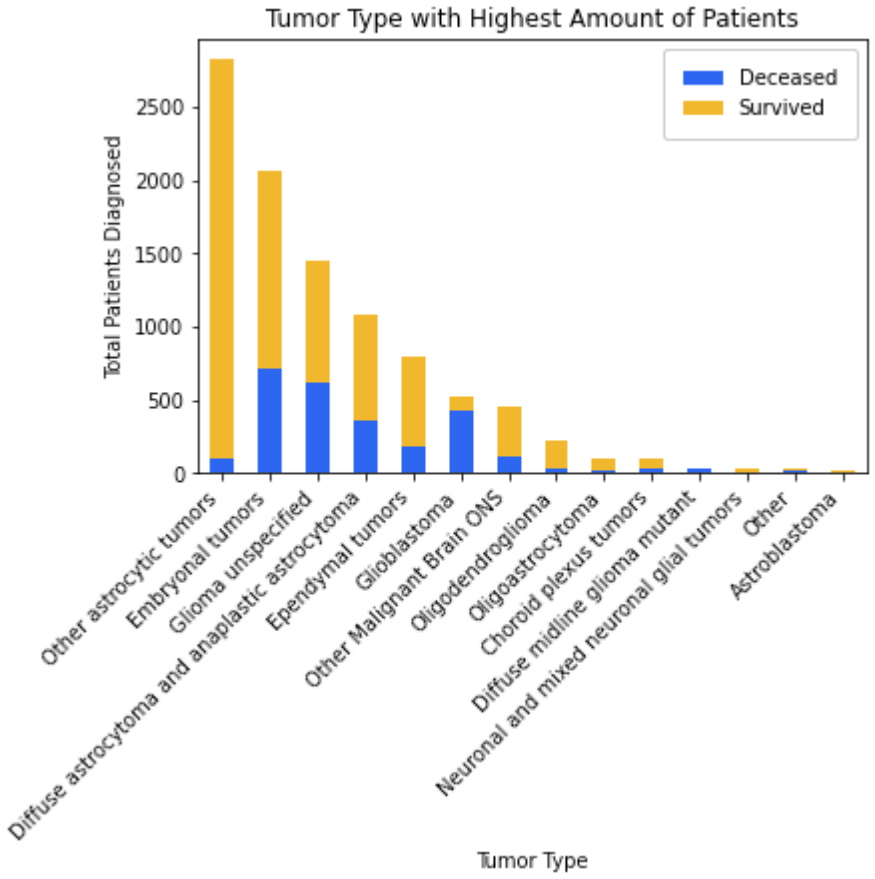


Figure 4



Patients with the lowest survival percentage had tumors located in the brain stem and a type of tumor called Glioblastoma. These are aggressive, fast-growing tumors that are primarily made up of abnormal astrocytic cells. They also include blood vessel cell types, which feed the tumor and allow the cells to reproduce rapidly and invade other regions of the brain. Figures 5 and 6 show the survival percentages of patients with these types of tumors.

56% Survived

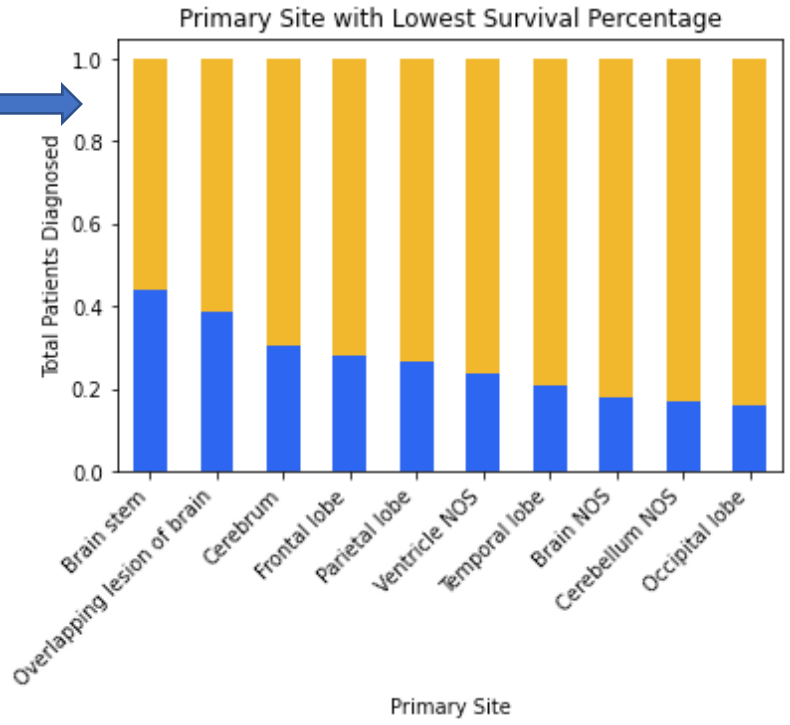


Figure 5

19% Survived

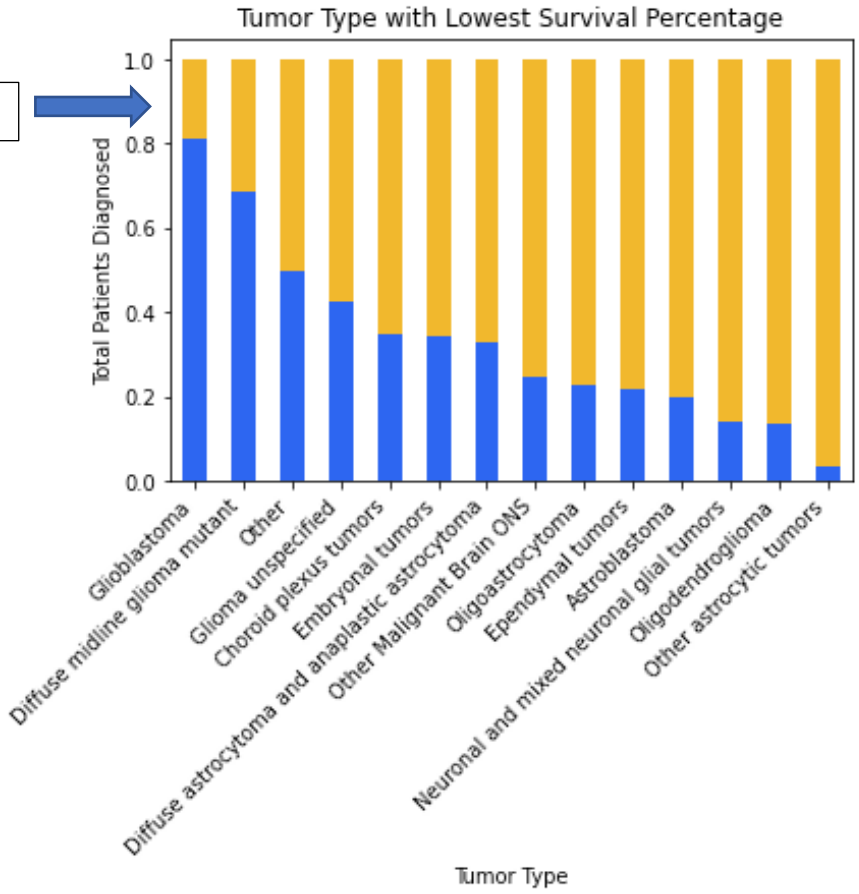


Figure 6

As mentioned in the study, race and age were indicators of survival. Figure 7 shows the survival percentage of patients by race and origin code.

	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%

Figure 7

There is a 9.14% difference between Non-Hispanic White patients and Non-Hispanic Asian or Pacific Islander patients. Figure 8 shows the survival percentages by age bins.

	Deceased	Survived	
1	30.63%	69.37%	0-4
2	30.07%	69.93%	5-8
3	24.02%	75.98%	9-13
4	23.29%	76.71%	14 - 19

Figure 8

The older a patient is at time of diagnosis, there is a better chance of survival.

Description of the Problem

By identifying which combination of variables lead to a higher probability of death, doctors can treat patients with these factors more aggressively. Additionally, research and analytics should be focused on these factors to improve survival.

A method to determine these variables is machine learning. In 2021, a paper was published titled *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. This study compared the most widely used method for survival analysis, Cox Proportional Hazard regression (CPH), to the machine learning algorithms Random Survival Forests (RSF) and Classification and Regression Trees (CART). The patients analyzed were diagnosed with glioblastoma brain cancer. The analysis found RSF achieved the best performance and highest accuracy followed by CPH and lastly CART for both short- and long-term predictions. This study reveals machine learning algorithms may be a powerful tool for survival analytics of cancer patients.

This project will analyze the variables referenced in the pediatric brain tumor survival paper with machine learning algorithms to predict patient survival.

Data

Both studies used data from the Surveillance, Epidemiology, and End Results (SEER) program. This program has been funded by the National Cancer Institute (NCI) since 1973. It is comprised of 18 core registries where data regarding patients with cancer are recorded. There are a total of 331,449,281 patients with 263 variables available for analysis.

This project focuses on patients diagnosed with brain cancer who range in age from 0 to 19. The variables selected for analysis include year of diagnosis, age, race, gender, primary site, tumor type, grade, laterality, median household income, cause of death, and rural urban code. The year of diagnosis spanned from 2000 to 2019.

The data was obtained using SEER*Stat program. This program allows users to have access to the database and filter the data by variables. The query produced 12,164 patients.

Methods for Data Analytics

Of the 12,164 patients, 101 were first removed because the cause of death was unknown. Patients who were diagnosed before 2003 and after 2017 were also removed because the earlier overall survival rate was much lower. Patients who were diagnosed in 2018 and 2019 had a much higher survival rate. This is probably because these patients were diagnosed less than 5 years ago.

An additional 102 were removed because their race and origin recode was listed as Non-Hispanic Unknown Race. The survival rate for these patients was 96.08%, which is significantly higher than the other categories.

The Grade variable was removed from analysis because over 66% of the patients had unknown or blanks in this category as shown in Figure 9.

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

Figure 9

Age was binned according to quartiles. There were more younger patients in the analysis as shown by the histogram in Figure 10.

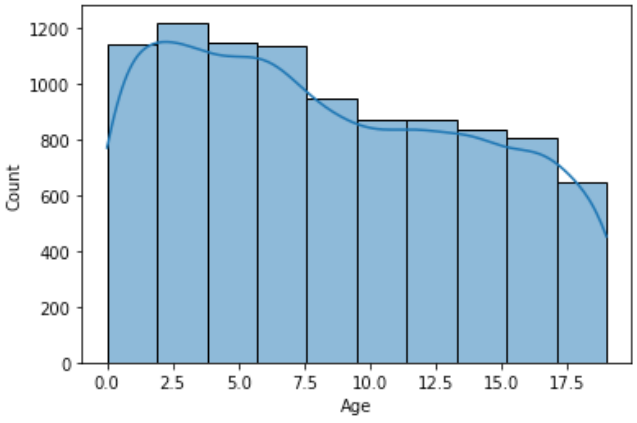


Figure 10

Figure 8 on page 5 above shows the range of ages within the bins. The bin with the highest survival ranges from age 14-19 at 77% while the lowest is 0-4 years at 69%.

The final totals for analysis were 9,611 patients with 9 variables.

Next a Cramer's V Correlation Matrix was done to see if there were any correlations between the X-variables for feature selection. Figure 11 shows a heat map revealing there were no correlations between the variables, so none were removed.

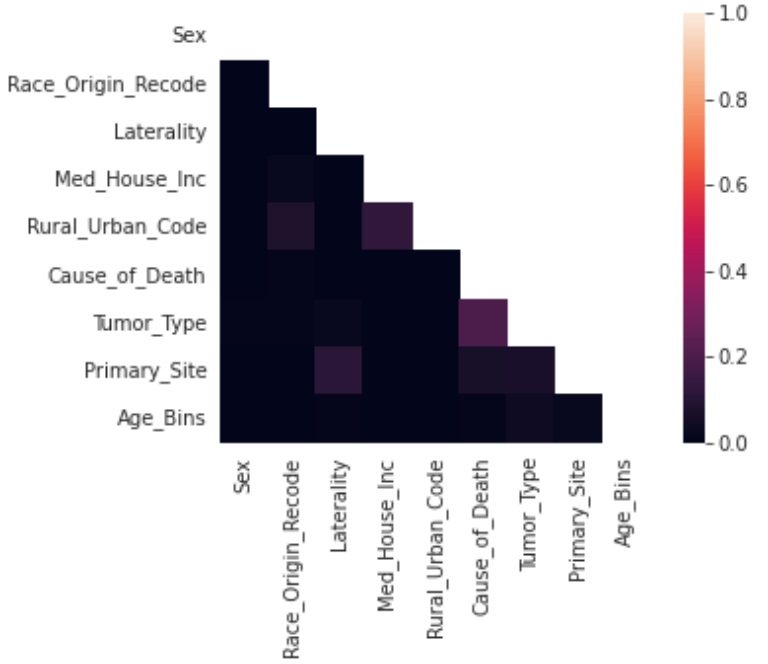


Figure 11

The training and testing data was stratified to reflect the distribution of race and origin categories in the original data. This ensures the model is trained on a population representative of the original so it will not give inaccurate predictions for patients belonging to less populated groups.

Finally, the y-variable was unsampled as there were significantly more patients who survived than were deceased.

After the get dummies function was applied to the testing and training sets, the total number of training rows was 11,216 and the testing set had 1,923 with 59 columns for both.

Machine Learning

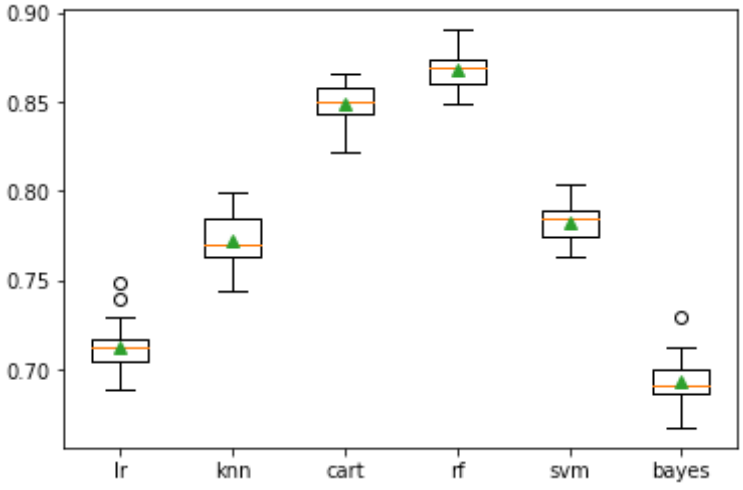
The classifier models used for this analysis are Logistic Regression, K-nearest neighbor, Decision Tree, Random Forrest, Support Vector Classifier, and Gaussian Naïve Bayes. The models were evaluated using repeated stratified K fold with a total of 10 splits and the accuracy was found using the training data.

Finally, a stacking classifier was used to generate predictions with the testing data. The stacking classifier uses the output from several classifiers to make predictions with a final estimator. The classifiers listed above were used and the final meta learner was a logistic regression classifier. A confusion matrix was created so the results could be analyzed.

Evaluation of Results

The models were first evaluated using the accuracy of prediction for the training set. Figure 12 shows a boxplot of the distribution of the accuracy scores for each model.

Figure 12




```
lr 0.713 (0.013)
knn 0.772 (0.014)
cart 0.849 (0.011)
rf 0.868 (0.010)
svm 0.783 (0.011)
bayes 0.693 (0.013)
```

The classifier with the highest accuracy was the random forrest followed by the decision tree. The classifier with the lowest accuracy was the Gaussian Naïve Bayes followed by the logistic regression.

The confusion matrix for predictions is below as Figure 13.

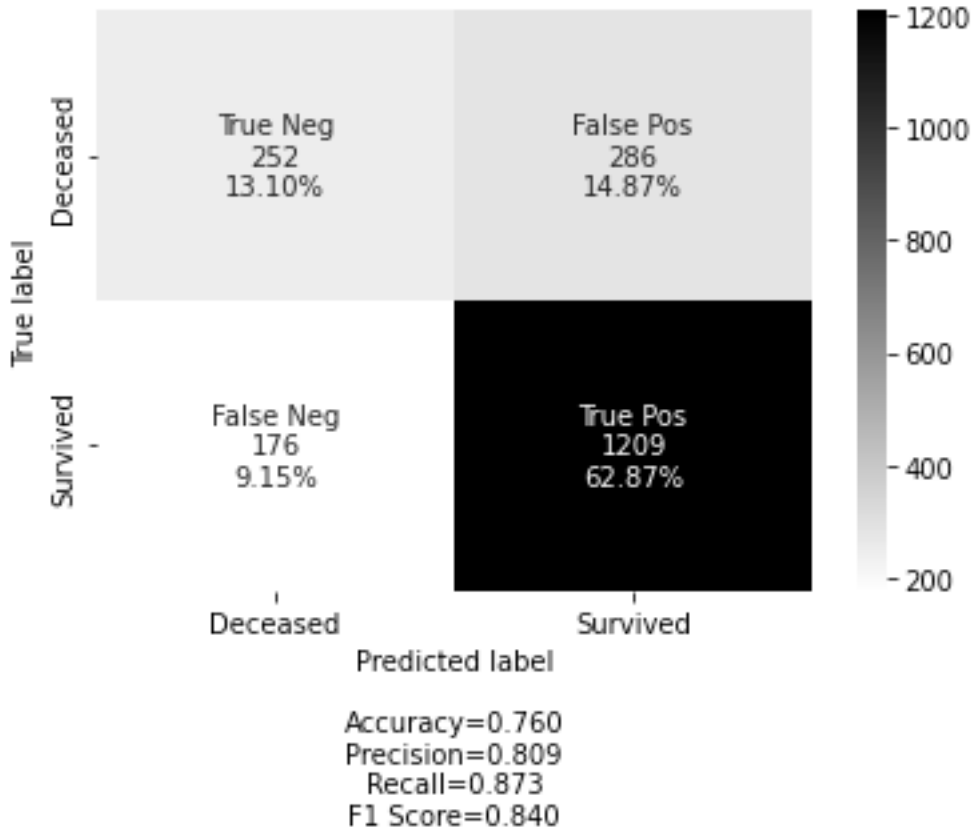


Figure 13

The accuracy was 76% and the recall was higher than the precision indicating the model was better at identifying patients who survived than not labeling deceased patients as survived.

The prediction data was further analyzed by examining the distribution of the variables from patients incorreclty classified. Most of the distributions resembled those of the population. For example, 52% of the incorrectly classified patients had a race code of Non-Hispanic White and

this population represents 55% of the data. Two variables of interest would be tumor type and primary site. The distribtion for these variables is shown below as Figure 14.

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

Figure 14

For tumor type, 40% of the incorrect predictions were labeled as Embryonal tumors and these tumors are the 2nd most common type and the 6th deadliest. Glioma unspecified made up 15% of the incorrect classifications and this is the 3rd most common type of cancer and 4th deadliest. For primary site, the brain stem is the deadliest site for cancer and the 2nd most common and this site makes up 27% of the incorrect classifications. The cerebellum is the site for 26% of the incorrectly classified data but this is the most frequent site in patients and the 3rd deadliest.

To increase model performance, it would be beneficial to either apply weights or improve the distribution of this variable in the training model.

Results

The model performed well in predicting patient survival with an accuracy of 76.5%. The recall was higher than the precision indicating the model was better at identifying patients who survived than not labeling deceased patients as survived. It would be more beneficial to have a model that was more sensitive to detecting deceased patients.

Analysis of the incorrectly classified predictions indicated the variables for primary site and tumor type should be re-evaluated as they contributed to incorrect predictions. This could be done by stratifying the test set so these variables are representative of the distribution of the original population or weights could be applied.

Discussion and Conclusion

Machine learning algorithms have been shown to be better predictors of brain cancer patients than traditional methods in previous studies. This analysis revealed the importance of evaluating the results to better improve the models so they can make predictions that are more helpful in saving patient's lives. The next steps for this analysis would be to predict the survival months of patients who were classified as deceased.

The importance of survival prediction is so doctors can aggressively treat patients with tumors that have a higher risk of death. Additional research and analytics could be reallocated to these indicators as well. Further, patients who match the demographics that indicate a higher risk of death could be given resources to try and ameliorate the obstacles during treatment they may face.

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