Utilizing Machine Learning Algorithms to Predict Survival Days in Glioblastoma Patients

Sabrina Imhof
Spinal Cord Injury Center
Balgrist University Hospital
sabrina.imhof@balgrist.ch

Selina Kohler
Health Sciences and Technologie BSc
ETH Zurich
selkohler@student.ethz.ch

Seraina Kämpf
Health Sciences and Technologie BSc
ETH Zurich
skaempf@student.ethz.ch

Melanie Grimm
Health Sciences and Technologie BSc
ETH Zurich
melgrimm@student.ethz.ch

Abstract

INTRODUCTION: Monitoring after glioblastoma removal relies on magnetic resonance imaging to detect any signs of recurrence. The aim of this group project was to determine the key MRI features that can accurately predict the number of days of survival following tumor removal surgery. METHODS: Five machine learning models were used to analyze the dataset from the University of Pennsylvania glioblastoma (UPenn-GBM) cohort, aiming to identify the most effective model for predicting the duration of survival following tumor removal. RESULTS: All models showed poor performance as indicated by low R^2, high root mean square error and mean absolute error. Final feature selection demonstrated a non-linear distribution. CONCLUSION: Due to the unsatisfactory performance of the models, interpretability of the final features is limited. For future analysis, it is recommended to employ regression algorithms suitable for non-linear data. Alternatively, transforming the label data and subsequently applying a classification algorithm can also be considered.

1 Introduction

As emphasized by Bakas et al. (2022) and Salvati et al. (2020), glioblastoma (GBM) is the most common, complex, and highly malignant primary tumor of the central nervous system (CNS). The overall survival of GBM patients is not more than one-third at 2 years and only around 10 percent at 5 years after diagnosis. There has been limited improvement in the overall survival rate of the patients, despite the advancements in these standard treatment options over the past two decades, according to Bakas et al. (2022). As stated by Stupp et al. (2009), the current standard of care still involves surgically removing the tumor followed by chemoradiation treatment. Postoperative surveillance after tumor excision heavily relies on magnetic resonance imaging (MRI) to detect any signs of recurrence. Unfortunately, the standardization of radiographic metrics for treatment response and for determining disease progression in GBM has proven to be very challenging. In the opinion of Akbari et al. (2020), particularly differentiating between MRI changes caused by treatment and genuine tumor progression holds significant implications for clinical decision-making. Although different studies tried to address this problem, the used data sets often showed numerous limitations, such as the number of included subjects, lack of consistent acquisition protocol, variable quality of data or accompanying clinical, demographic, and molecular information. In 2022, the study group of Bakas et al. (2022) introduced the "University of Pennsylvania Glioblastoma Imaging, Genomics and

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Radiomics" (UPenn-GBM) dataset. This is currently the largest publicly available comprehensive dataset with 630 patients diagnosed with de novo GBM and includes clinically acquired data, such as advanced multi-parametric magnetic resonance imaging scans, clinical and demographic data, and molecular status for mutation. The dataset further includes preprocessed scans according to a standardized protocol, extracted perfusion, and diffusion derivative volumetric scans, computationally derived and manually revised expert annotations of tumor sub-region boundaries, and quantitative imaging features. This large data set presents new opportunities for evaluating the relevance of MRI features in relation to survival time and, consequently, improving the ability to differentiate true tumor progression from MRI changes caused by treatment. Therefore, the aim of this group project was to identify the crucial MRI features within the UPenn-GBM dataset that can effectively predict the duration of survival following surgical removal of the tumor.

2 Methods

2.1 Patient cohort

The patients included in the cohort were between 18 and 89 years of age at the time point when the MRI scans were conducted. The ratio of the gender distribution is 60:40 male:female. The resection status of the 611 patients with available pre-operative baseline scans was partitioned in the three categorical entries Gross Total Resection (GTR, n = 362) table 4, Partial Resection (PRe, n = 211), and Not Available (NA, n = 38). For 452 patients the overall survival data, ranging from 3 to 2207 days was provided. Mutations in the IDH1 (isocitrate dehydrogenase-1) gene and methylated MGMT (O-6-methylguanine-DNA methyltransferase) promoter have been associated with several cancer types and therefore their mutation status was examined. Other features included, were the Karnofsky performance score (KPS) prior to treatment and the treatment response classification of pseudo-progression vs. true progression (`PsP_TP_score`), for which no data were provided.

Table 1: IDH1 gene

IDH1:	
Wildtype	499
NOS/NEC	96
Mutated	16

Table 2: MGMT gene

MGMT:	
Not Available	322
Unmethylated	151
Methylated	111
Indeterminate	27

Table 3: Karnofsky Performance Status (KPS)

KPS scores:	
Not Available	536
30	1
40	2
60	6
70	6
80	18
90	37
100	5

Table 4: Gross Total Resection (GTR) >90%

GTR_over90percent:	
Yes	362
No	211
Not Available	38

2.2 Data cleaning, preprocessing, and splitting

First, some basic steps of data cleaning were conducted. The missing information of the treatment response (`PsP_TP_score`) in the clinical dataset was removed and only patients with available information on the number of survival days were considered. Additionally, the dataset was restricted to only complete cases. Furthermore, the categorical features of the GTR score and the gender were numerically transformed. In the end, the remaining features of the clinical dataset were added to the radiomic dataset, which contained the 4753 extracted features of the MRI scans. The data was visualized with seaborn using a customized color palette with black (#000000), malibu blue

(#6db6ff), and silver chalice (#a3a3a3). Before splitting the dataset into training and test sets the feature `"Survival_from_surgery_days"` was set as our label. Then the dataset could be split with a ratio of 80:20 and normalized with a standard scaler. The random seed was set to 2023. A large number of features incentivized to first conduct a feature correlation prior to the feature selection. Additionally, the data were tested for normality with the Spearman method. The features, which showed a correlation higher than +/-0.7 were dropped. Recursive feature elimination (RFE) was used for the univariate feature selection. The best 236 features, which had an importance of over 0.001, were selected for further applications.

2.3 Machine learning models and final feature selection

There is no clear classification detectable in our label and therefore, four different types of supervised machine learning models specific for regression tasks were applied. The models, linear regression, Lasso and Ridge regression, as well as the random forest, were chosen. The final model performance evaluation gave the incentive to additionally program a non-linear model on the given dataset. The Support Vector Machine algorithm (SVM), which can solve linear, as well as non-linear problems, was applied and its performance was compared to the linear models. A cross-validation, iterating over three folds for all models, was performed. The coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE) were calculated to evaluate the model performance for each fold. For the final feature selection of the top ten features and evaluation of the test data, the ridge regression model was chosen. The number of top features was chosen arbitrarily. Lastly, the linearity of the top ten features was subject to verification and the final model was evaluated on the test dataset.

3 Results

3.1 Data visualization

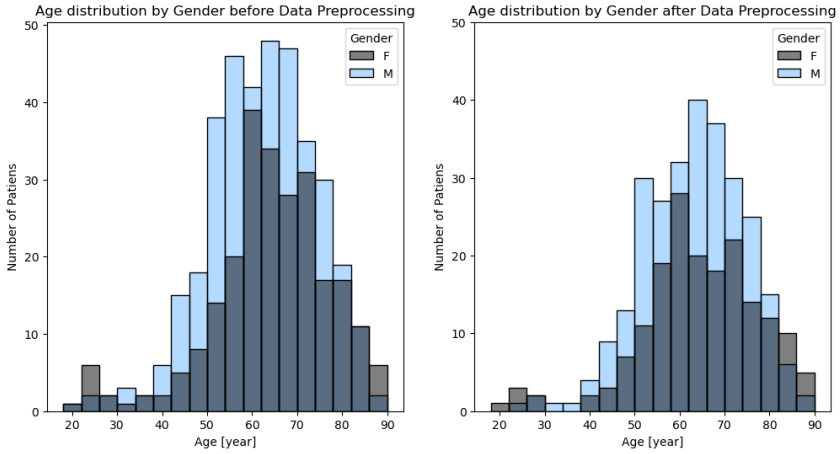


Figure 1: Age distribution split by gender

3.2 Correlations between variables

To test for the normality of the clinical dataset, the distribution of the correlation coefficient was plotted. The bell-shaped curve showcases a normal distribution, centered at 0.00.

3.3 Supervised machine learning models

The performances of the 5 models applied are summarized in Table 5 and visualized in the adjoining figures 4a & 4b. There are no large differences between the performance of the different models. All R^2 values fluctuate around zero with the Linear Regression and Support Vector Machine model

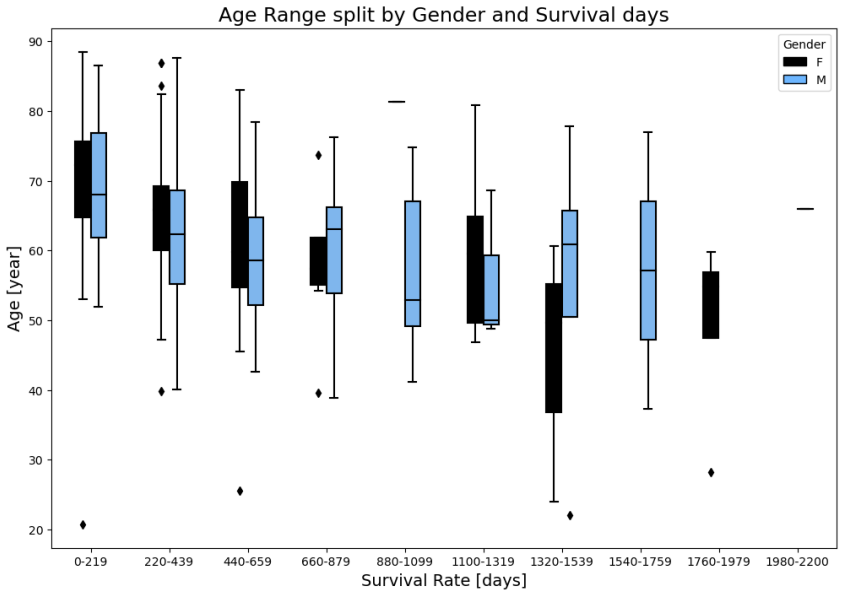


Figure 2: Survival days based on age and gender

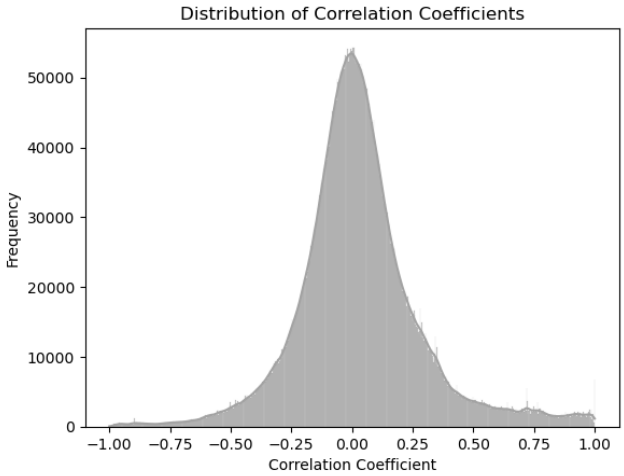
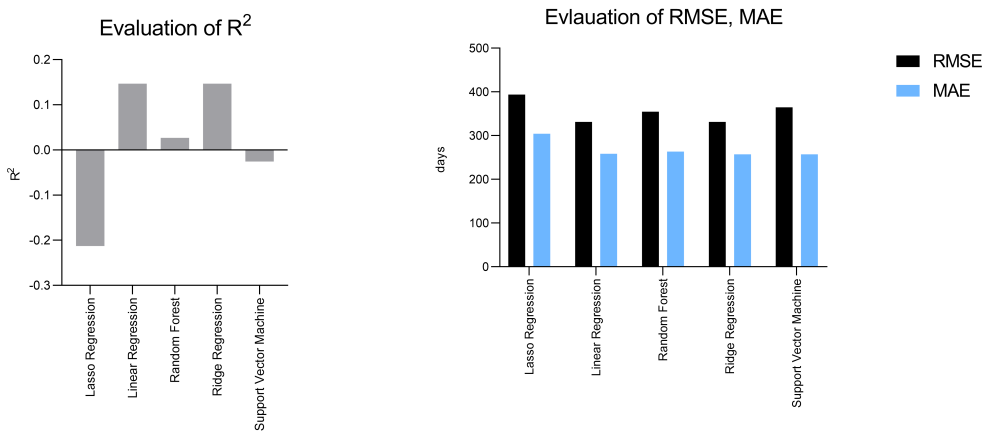


Figure 3: Testing for normality: Distribution of the spearman correlation coefficient

showcasing negative R^2 values. Taking also the RMSE and MAE into account, the Ridge Regression model performed the best.

Table 5: Mean of model performance in the three-fold cross-validation

clf	r2	rmse	mae
LassoRegression	-0.212516	394.004500	304.452306
LinearRegression	0.146758	331.623527	258.423394
RandomForest	0.026725	354.890583	263.549116
RidgeRegression	0.146804	331.606612	257.490599
SupportVectorMachine	-0.025714	364.656692	257.304354



(a) R^2

(b) RMSE, MAE

Figure 4: Performance evaluation of the supervised machine learning models

The importance of the top ten features is visualized in Fig. 5. The mean absolute deviation of the measured intensity of the enhancing tumor region derived by Diffusion Tensor Imaging (`DTI_TR_ET_Intensity_MeanAbsoluteDevation`) was identified as the top feature with an absolute importance of 0.0135. No linear correlation between survival days and the top ten features can be identified (Fig. 6).

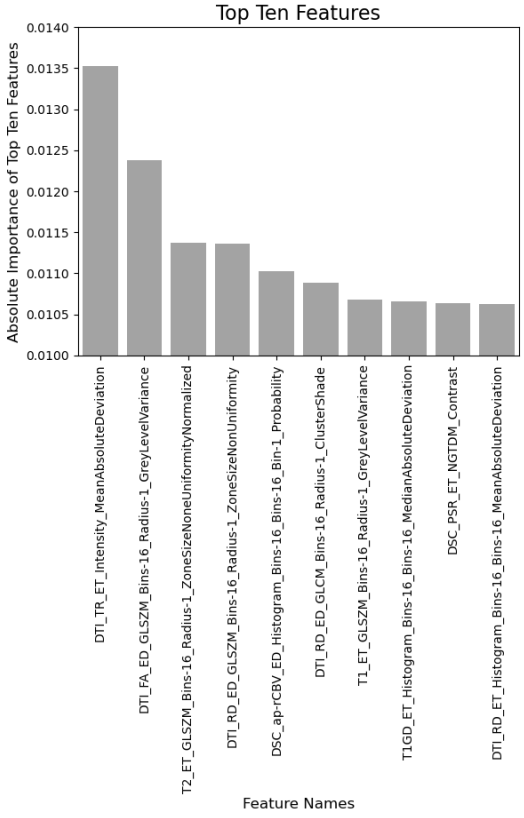


Figure 5: Final feature selection: The top ten features

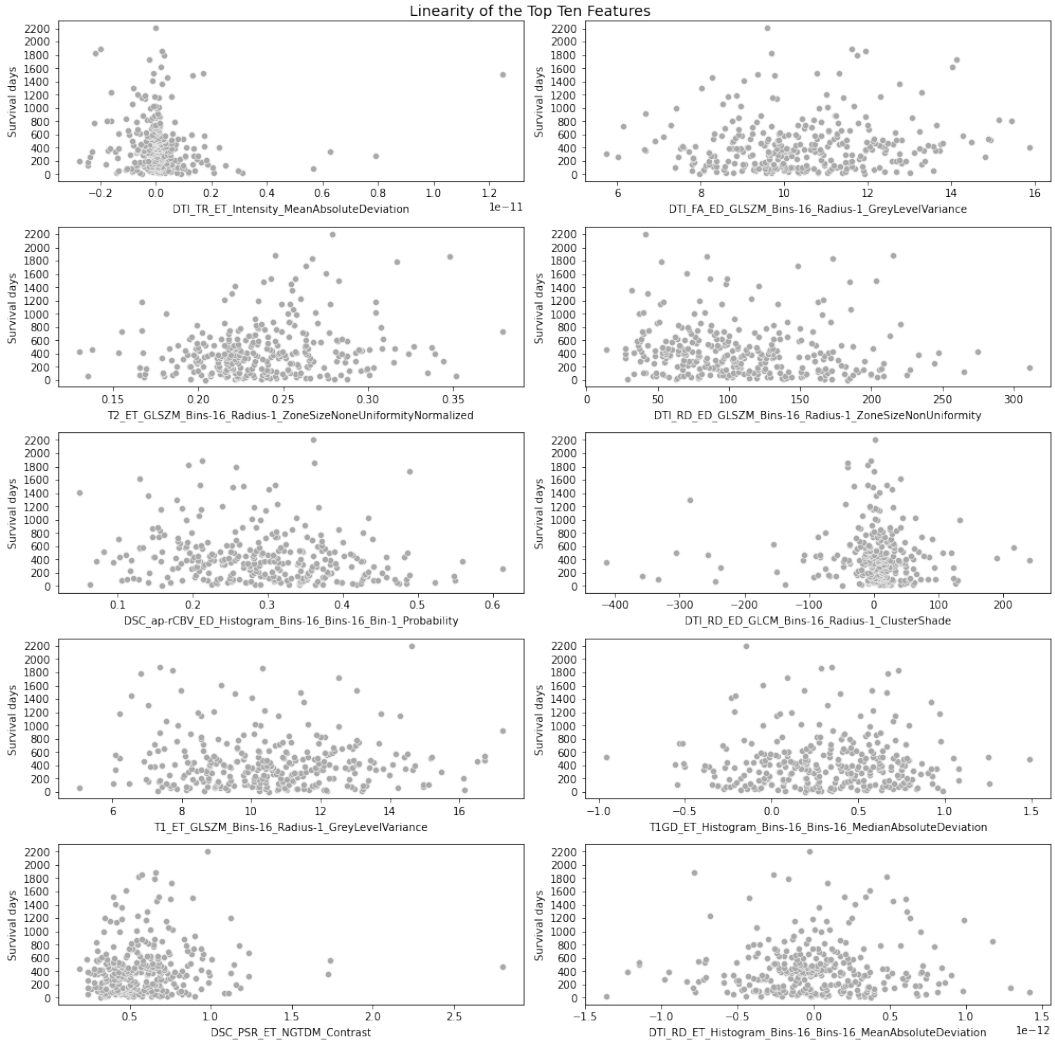


Figure 6: Final feature selection: Testing for the linearity of the top ten features

The performance of the final model on the test data set is summarized in Table 6.

Table 6: Final evaluation model

clf	r2	rmse	mae
Ridge Regression	-1.0747	648.1000	478.9046

4 Discussion

The present project tries to identify the key MRI features that can accurately predict the number of survival days from the UPenn-GBM dataset following tumor removal by applying different machine-learning models. The models used were Lasso Regression (LaR), Linear Regression (LiR), Random Forest (RF), Ridge Regression (RR), and Support Vector Machine (SVM). The final feature selection process was performed on the test set using the Ridge Regression (RR) model, which demonstrated the highest performance in the pre-analysis stage. However, the overall performance of all five models was unsatisfactory, indicated by a low R^2 value, suggesting a weak relationship between the response and predictor variables. Furthermore, the high values for RMSE and MAE indicate that the

predictions generated by the model are not accurate. In conclusion, it can be stated that the model lacks utility in accurately predicting patient survival.

In current literature, most studies published on GBM and machine learning primarily focus on classification tasks rather than regression analysis, as shown in the study of Cao et al. (2020). This preference stems from the clinical importance of determining whether an MRI image displays tumor tissue or not, rather than precisely predicting the survival time once GBM is confirmed through imaging. Consequently, there is limited literature available for directly comparing and validating the top features found in this project with other studies. However, in the study conducted by Massett et al. (2023), they assess and validate a regression model that utilizes MRI to identify specific contributions of distinct neuroanatomical structures to brain aging in a cohort of healthy subjects. Their model performance is fairly good achieving a value of R^2 = 0.81 on training and R^2 = 0.79 on test data. At the same time, their predictions are quite accurate with MSE = 70.94 and MAE = 6.57 for training data and MSE = 81.96 and MAE = 7 for test data.

In the study conducted by Senders et al. (2020) the prediction of survival in patients diagnosed with GBM was performed using both classical statistical methods and machine learning techniques. The authors employed SVM and RF models, among others, to assess the survival duration of GBM patients. While the study did not report specific metrics such as R^2, RMSE, MAE, the models demonstrated favorable performance, with a concordance index ranging from 0.68 to 0.70 across all models. It is worth noting that the dataset utilized by Senders et al. consisted of over 20,000 patients and included only a limited number of features. Therefore, they did not encounter the challenge of dealing with feature multidimensionality, as observed in the current group project. Furthermore, their dataset also had a limited number of radiographic features, making it more challenging to directly compare their results with the findings of the present study.

Due to the bad performance of the models utilized in this project, it raises the question of whether regression models are suitable for analyzing the given dataset. Spiess and Neumeyer (2010) propose, that although often used, R^2 is not an adequate measure for nonlinear data in pharmacological and biochemical research. Because of the large number of features in the dataset, it was not feasible to examine the distribution of all the features during the preprocessing phase. However, upon analyzing the distribution of the top 10 features after the final analysis (Fig. 6), it became evident that our data does not follow a normal distribution. So instead of linear models the authors recommend employing bias-corrected measures such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which have demonstrated superior performance Penny (2012). Moreover, implementing polynomial regression analysis could potentially improve the prediction of non-normally distributed data.

Apart from the mathematical evaluation of the model performance, it is also necessary to look at the top 10 features found with the RR method to evaluate clinical interpretability. As none of the members of this project group are normally working with MRI data and because of the absence of literature to compare the features found, a discussion with an expert working in the MRI field was held instead. In summary, the features revealed by the RR method seemed to be reasonable, as they assess different structures as myelin or contrast agent-associated MRI features as well as perfusion and diffusion parameters are in our top features. Nevertheless, we were advised to further improve the model performance before using or recommending the features found.

4.1 Limitations

As a label, we chose the survival days after surgery of GBM. As some patients in the dataset were still alive and at the same time we didn't know how many days ago the operation was, we had to drop these patients which was roughly half of the subjects in the dataset. This led to an even smaller number of patients compared to numerous features in the dataset and with that to a smaller number of subjects in the train/test split. For further analysis, it could be beneficial to use another label where fewer dropouts have to be stated when it becomes evident that the label chosen has too many dropouts.

Additionally, in our analysis, non-numerical features, except for gender and the gross total resection rate over 90 percent (`GTR_over90percent`), were excluded. These features were dropped due to either having numerous missing values or being almost uniformly constant (e.g. IDH1 gene), thereby not providing any additional or distinguishing information for the analysis. However, it is worth

considering the inclusion of other clinical features, as indicated by a separate study conducted by Verduin et al. (2021). Therefore, incorporating clinical features such as tumor type or the score for activities of daily living (KPS) may potentially enhance the performance of the models.

In order to enhance the performance of the model using the selected survival days as the target label, one potential approach would be to switch from regression to a classification task. This would involve transforming the continuous data of the label into distinct bins. The project group initially contemplated pursuing a classification task at the project's outset. However, the idea was ultimately dismissed due to concerns about determining appropriate bins based on survival thresholds or utilizing mathematical thresholds like quartiles. The project group wanted to avoid any loss of information that could result from artificially assigning the bins.

5 Conclusion and Outlook

GBMs are highly malignant and common primary tumors of the CNS. One of the main treatment approaches involves surgical tumor removal followed by chemoradiation, along with close postoperative monitoring using MRI. However, despite its frequent use, standardizing radiographic metrics for treatment response assessment and disease progression determination in glioblastoma remains challenging. The objective of our group project was to identify the key MRI features that could predict the survival duration after surgical tumor removal, utilizing the UPenn-GBM dataset. Unfortunately, none of the five different machine learning models employed performed satisfactorily. Therefore, it is important to interpret our final feature selection with caution. To enhance our analysis further, we recommend implementing a nonlinear model such as polynomial regression to account for the non-linear distribution of the data. Additionally, it could be worth considering transforming the label data into classes or a dichotomous outcome, allowing for the use of a classification paradigm instead of regression. This approach would align more closely with the current research methods employed in GBM studies.

References

Akbari, H., Rathore, S., Bakas, S., Nasrallah, M., Shukla, G., Mamourian, E., and et al. (2020). Histopathology-validated machine learning radiographic biomarker for non- invaisve discrimination between true progression and pseudo-progression in glioblastoma. *Cancer*, 126(11):2625–2636.

Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., and et al. (2022). The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Sci Data*, 9(1):1–12.

Cao, H., Erson-Omay, E., Li, X., Günel, M., Moliterno, J., and Fulbright, R. (2020). A quantitative model based on clinically relevant MRI features differentiates lower grade gliomas and glioblastoma. *European Radiology*, 30(6):3073–3082.

Massett, R., Maher, A., Imms, P., Amgalan, A., Chaudhari, N., Chowdhury, N., and et al. (2023). Regional neuoranatomic effects on brain age inferred using magnetic resoncance imaging and ridge regression. *J Gerontol A Biol Sci Med Sci*, 78(6):872–881.

Penny, W. (2012). Comparing Dynamic Causal Models using AIC, BIC and Free Energy. *Neuroimage*, 59(1):319–330.

Salvati, M., Bruzzaniti, P., Relucenti, M., Nizzola, M., Familiari, P., Giugliano, M., and et al. (2020). Retrospective and randomized analysis of influence and correlation of clinical and molecular prognostic factors in a mono-operative series of 122 patients with glioblastoma treated with STR or GTR. *Brain Sci*, 10(91):1–14.

Senders, J., Staples, P., Mehrtash, A., D.J., C., Taphoorn, M., D.A., R., and et al. (2020). An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning. *Neuorsurgery*, 86(2):E184–E192.

Spiess, A. and Neumeyer, N. (2010). An evluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacology*, 10(6):1–11.

Stupp, R., Hegi, M., Mason, W., van den Bent, M., Taphoorn, M., Janzer, R., and et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol*, 10(5):459–466.

Verduin, M., Primakov, S., Compter, I., Woodruff, H., van Kuijk, S., Ramaekers, B., and et al. (2021). Prognostic and Predictive Value of Integrated Qualitative and Quantitative Magnetic Resonance Imaging Analysis in Glioblastoma. *Cancers (Basel)*, 13(4):2–19.

List of Tables

List of Figures

Appendix

Python version: 3.10.10
Matplotlib version: 3.7.1
Pandas version: 1.5.3
NumPy version: 1.23.5
Seaborn version: 0.12.2
Scipy version: 1.10.0
Sklearn version: 1.2.2