Machine Learning-Based Radiomic Features for Glioblastoma Overall Survival Prediction

Ankit Das[1], Kee Yen Cheng[2], Yong Liu[1], Rick Siow Mong Goh[1], Feng Yang[1]
[1]*Institute of High Performance Computing*
Agency for Science, Technology and Research
Singapore 138632, Republic of Singapore
[2]*Temasek Polytechnic*, Singapore
[1]{dasak, liuyong, gohsm, yangf}@ihpc.a-star.edu.sg
[2]yencheng109@gmail.com

***Abstract*—This study aims to prognosticate the overall survival (OS) of patients afflicted with gliomas undergoing treatment. The predictive models are constructed through diverse combinations of feature selection techniques and machine learning algorithms. MRI scans (T1 and T2 FLAIR) obtained from pre-treatment, one week post-treatment, and two months later yielded 112 features. Delta radiomics, derived from pre and post-treatment features are used for feature selection using Mutual information, Chi-squared, and F-test. Six machine learning methods (RF, Logistic Regression, KSVM, MLP, Gradient Boosting, XGBoost) were used to build a classification model predicting Overall Survival (OS). F-test outperformed, yielding the highest AUC, ACC, TPR, and TNR in conjunction with Gradient Boosting.**

***Index Terms*—Survival prediction, Machine learning, Classification, Regression, Feature Selection.**

I. Introduction

Gliomas are the predominant form of primary malignant tumors originating in the brain. They constitute the major percentage of intracranial tumors. Gliomas are categorized by the World Health Organization according to their malignancy into low-intensity and high-intensity [10]. The distinct categorization of gliomas display distinct levels of invasiveness and prognoses, presenting a substantial risk to human health. Despite the application of comprehensive treatment approaches, relapse is nearly universal among patients. The challenges arise from factors like spatial and temporal intratumor heterogeneity, as well as the extent and location of the tumors, rendering them difficult to resect and, in certain instances, inoperable. The considerable challenges posed in removing tumors through surgery and the restriction in passing the drugs to the brain significantly contribute to the absence of efficacious treatments and an unfavorable prognosis for individuals facing these issues.

Magnetic resonance imaging (MRI) serves as a widely utilized technique for acquiring images of brain tumors, typically incorporating various modes. Scientists have recognized that MRI offers unique information capable of predicting survival [8], [9], irrespective of pathological and clinical data. The technique involves extracting diverse quantitative features, considering factors like intensity, volume and shape etc., from MRI images. Subsequently, models are devised to establish the correlation between these extensive features and the patient's survival and overall outcome. This approach is commonly referred to as radiomics [5].

Radiomics, a sophisticated methodology utilizing characterization algorithms, facilitates the extraction of a myriad of features from medical images [1]. These extracted features find versatile applications in various analyses, ranging from predicting overall survival and tumor staging to tumor classification. In the context of this research, the emphasis is placed on the prognostication of overall survival. Distinguishing between single time-point radiomics, which involves features extracted from a solitary time point, and delta radiomics [11], encompassing features extracted from the discrepancy between two distinct time points, is crucial to the study. The primary aim is to meticulously evaluate and juxtapose the effectiveness of delta radiomics in contrast to the features derived from single time points. The study's outcomes underscore a noteworthy superiority in performance. This superiority is evident when comparing the delta radiomics features to their counterparts derived from single time points. The implications of these findings suggest the potential significance of incorporating delta radiomics in medical imaging analyses, particularly when predicting overall survival in a clinical setting. This study acquires T1 and T2 fluid-attenuated inversion recovery (FLAIR) MRI images at pre-treatment, one week post-treatment, and two months. From these images, 112 features are meticulously extracted. Additionally, delta radiomics are computed using features from pre-treatment and post-treatment scans. Feature selection techniques, namely Mutual Information, Chi-squared test, and F-test, are employed to identify pertinent features for the subsequent classification model.

The classification model, designed for OS prediction, incorporates six distinct machine learning algorithms: Random Forest (RF), Logistic Regression, Kernel Support Vector Machine (KSVM), Multi-layered Perceptron (MLP), Gradient Boosting, and Extreme Gradient Boosting (XGBoost). The models' performance is meticulously evaluated using prominent metrics, including the Area under the Receiver Operating Characteristic Curve (AUC), Accuracy Score (ACC), True Positive Rate (TPR), and True Negative Rate (TNR). Notably, the F-test for feature selection emerges as the standout performer among the evaluated methods, surpassing Mutual Information and Chi-

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00164

squared. The model featuring F-test for feature selection coupled with the Gradient Boosting classifier attains the highest AUC, ACC, TPR, and TNR scores. These findings underscore the efficacy of this specific combination in predicting OS for patients with recurrent malignant gliomas, offering valuable insights for clinical applications.

Section II presents the methodology and Section III represents the results followed by conclusions in Section IV.

II. Method

The initial section of our study provides a comprehensive description of the dataset. Subsequently, we delve into the processes of feature extraction, feature selection, and the implementation of classification and regression models.

A. Data

The data utilized in this research was sourced from a prior investigation conducted by the Duke university health system institutional review board [1]. This dataset comprises of 12 patients diagnosed with WHO grade III or IV recurrent malignant gliomas. The overall survival (OS) for the dataset ranged from 5.3 months to 29.4 months, with a demographic composition of 9 males and 3 females. The dataset includes MRI images for all 12 patients, along with a radiation therapy structure file available in the corresponding study on the same dataset [2]. The images were captured at three distinct time points: pre-treatment, one week post-treatment, and two months post-treatment. Each time point encompasses two types of MRI images, namely T1 and T2 Flair images. For every image type at each time point, 60 images are provided. These images are stored in Digital Imaging and Communications in Medicine (DICOM) format, with each DICOM file representing a slice of the patient's MRI in 2D. Combining the 60 images and the RT structure file enables the creation of a 3D image of the brain. To visualize this 3D model, the software 3D slicer is employed, highlighting specific regions of interest as outlined in the radiation therapy structure file.

B. Feature Extraction

Features are derived from T1 and T2 images captured across all three time points. The Pyradiomics python library [3], employed for image data feature extraction, necessitates the input image to be in Neuroimaging Informatics Technology Initiative (NIFTI) format. Given that DICOM images are initially in the form of slices—a 2D representation of a brain section—the conversion to NIFTI file format is undertaken to encompass a 3D image of the patient's brain. This conversion process is facilitated by the dcmrtstruct2nii Python library [4]. For each patient and across various time points and MRI image types, an individual image file is generated along with an accompanying mask file for each specified region of interest outlined in the RT structure file. The region of interest corresponding to the gross tumor volume is specifically employed in the feature extraction. Subsequently, the image and mask file for the gross tumor volume are utilized for feature extraction. The Pyradiomics python library is employed for this task, utilizing a function that requires the image and the region of interest mask file as input. A total of 120 features are extracted for each time point and MRI image type, falling into eight distinct classes: 1) First Order Statistic. 2) Shape-based (3D). 3) Shape-based (2D). 4) Gray Level Co-occurrence Matrix. 5) Gray Level Run Length Matrix. 6) Gray Level Size Zone Matrix. 7) Neighboring Gray Tone Difference Matrix. 8) Gray Level Dependence Matrix.

C. Post-processing for longitudinal features

The extracted data undergoes post-processing, involving two main steps. Initially, the T1 and T2 data features are amalgamated into a single CSV file using the open-source Python library, pandas. Subsequently, delta radiomics is computed between the features obtained at the three timepoints. Specifically, three delta radiomics features are calculated: Delta-1 and Delta-2, following the methodology outlined in a distinct study [2], and the addition of another delta radiomic, Delta-3. The data from pre-treatment, one-week post-treatment, and two-months post-treatment is consolidated into a single CSV file. A similar consolidation is applied to the delta features data, resulting in a total of eight datasets: 1) Pre-treatment (Fpre). 2) One-week post-treatment (Fpost). 3) Two-months post-treatment (Fpost2). 4) Delta-1 (ΔF1). 5) Delta-2 (ΔF2). 6) Delta-3 (ΔF3). 7) Fpre + Fpost + Fpost2. 8) ΔF1 + ΔF2 + ΔF3.

D. Feature selection

Feature selection is a crucial step in optimizing the classification or regression pipeline by identifying the most relevant features. This method aims to trim down the number of input features, emphasizing those that contribute the most to the output's variance. In this study, three feature selection methods are employed, and we will compare their impact on model performance. The three methods include mutual information, chi-squared, and the f-statistical test. Mutual information assesses the relationship between variables, measuring the reduction in uncertainty for one variable (overall survival) given another variable. A higher value indicates a more significant impact of the variable on patients' overall survival. Chi-squared is employed to evaluate the independence of two variables. A higher chi-square value suggests a lesser degree of independence between the two variables. For the chi-squared test in this study, a different normalized feature is utilized, opting for minimum-maximum normalization instead of z-score normalization. This choice is made because the chi-squared calculation requires non-negative input values. The F-test is a statistical method used for comparing models through hypothesis testing. A higher F-test value signifies a stronger correlation between the variables. Each feature selected was given a ranking in terms of its significance in affecting the OS of the patients. The feature ranking results were recorded down in a csv file for further analysis of the results. The selected features were then input used to build the classification model. The f statistical test performed better than the other 2 feature selection method thus it was used for the regression model.

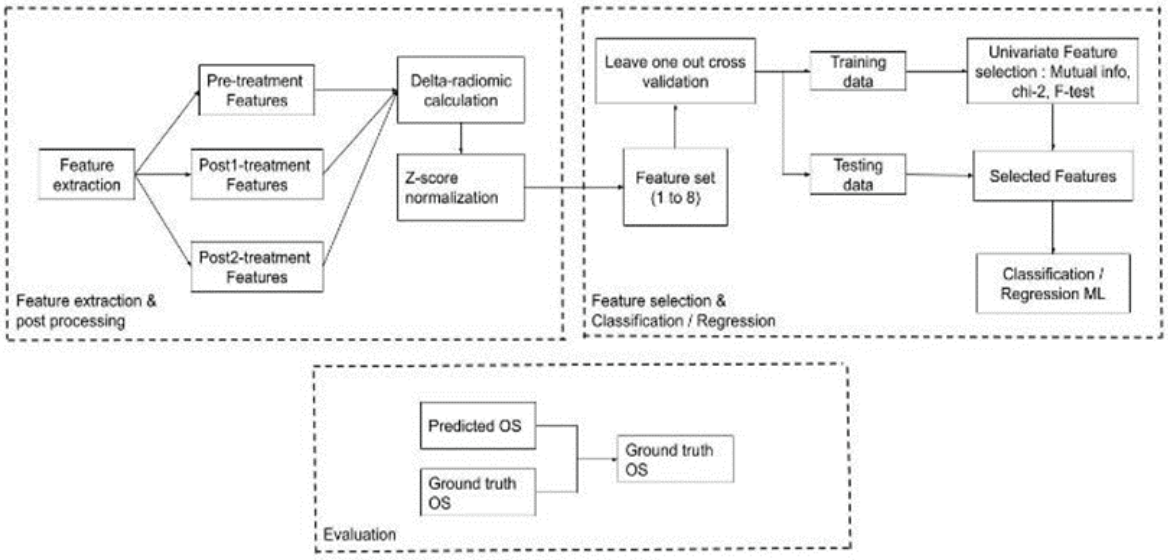


Fig. 1. Pipeline

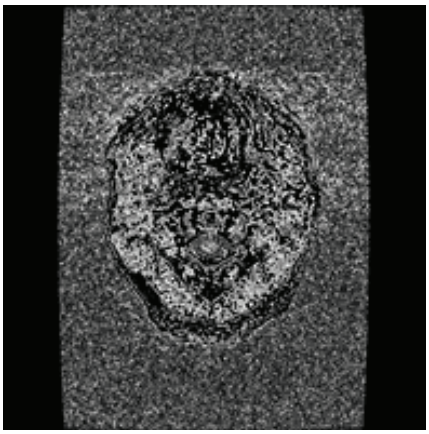


Fig. 2. Sample slice of image data.

The feature selection method is done on the training dataset after the splitting of data.

E. Classification Model

The eight feature categories (F_{pre}, F_{post}, F_{post2}, $\Delta F1$, $\Delta F2$, $\Delta F3$. $F_{pre+post+post2}$, $\Delta F_{(1+2+3)}$) will undergo separate analysis for feature selection and classification models. Given the limited number of patients in the dataset, Leave-one-out cross-validation (LOOCV) is employed to mitigate overfitting by iteratively splitting the data, leaving one observation for validation while the rest forms the training set. This process is repeated for each observation. Following this, univariate feature selection is applied to identify optimal features based on univariate statistical tests. The selection involves one of the following univariate tests: chi-square test, f-test, or mutual information. The chosen features are selected based on their significance in influencing overall survival. Subsequently, six machine learning algorithms are employed for binary classification: (1) Random Forest (RF) classifier using the scikit-learn ensemble (scikit-learn [7]); (2) Logistic Regression implemented through the scikit-learn linear model (scikit-learn [7]); (3) Kernel Support Vector Machine (KSVM) using the scikit-learn svm module (scikit-learn [7]); (4) Multi-layered Perceptron (MLP) employing the scikit-learn neural network module (scikit-learn [7]), with the MLP comprising one hidden layer with 10 hidden units; (5) Gradient Boosting (GB) implemented through the scikit-learn ensemble (scikit-learn [7]); (6) Extreme Gradient Boosting (XGBoost) implemented using the xgboost Python library [6]. During each iteration of LOOCV, the model is fitted and utilized to predict observations, implemented using the scikit-learn library.

F. Regression Model

The regression model adopts the same data splitting method as the classification model, employing Leave-one-out cross-validation (LOOCV) to alleviate overfitting. In the regression model, feature selection involves univariate techniques such as mutual information, chi2, and F-test. For the regression model, six machine learning algorithms are utilized: (1) RF Regression, implemented through the scikit-learn ensemble (scikit-learn). (2) Decision Tree, implemented using the scikit-learn linear model. (3) Kernel Support Vector Machine (KSVM), implemented through the scikit-learn svm. (4) Multi-layered Perceptron (MLP), implemented using the scikit-learn neural

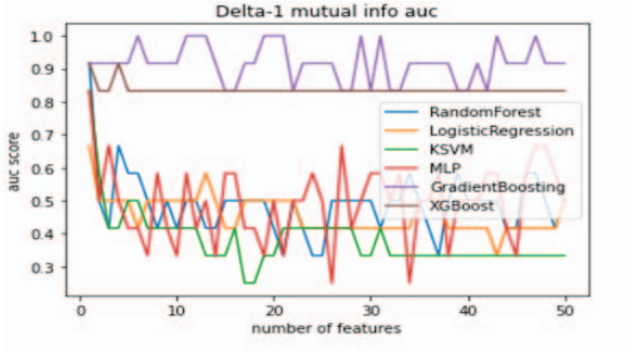
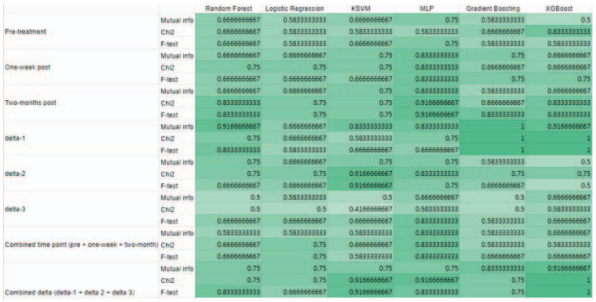


Fig. 3. Classification Results



		Random Forest	Logistic Regression	KSVM	MLP	Gradient Boosting	XGBoost
Pre-treatment	Mutual info	0.6666666667	0.5833333333	0.6666666667	0.75	0.5833333333	0.5
	Chi2	0.6666666667	0.5833333333	0.5833333333	0.5833333333	0.6666666667	0.8333333333
	F-test	0.6666666667	0.5833333333	0.6666666667	0.75	0.5833333333	0.5833333333
One-week post	Mutual info	0.6666666667	0.6666666667	0.75	0.8333333333	0.75	0.6666666667
	Chi2	0.75	0.75	0.75	0.8333333333	0.6666666667	0.6666666667
	F-test	0.6666666667	0.6666666667	0.6666666667	0.8333333333	0.75	0.75
Two-months post	Mutual info	0.6666666667	0.6666666667	0.75	0.8333333333	0.5833333333	0.6666666667
	Chi2	0.8333333333	0.75	0.75	0.9166666667	0.6666666667	0.8333333333
	F-test	0.8333333333	0.75	0.75	0.9166666667	0.8333333333	0.8333333333
delta-1	Mutual info	0.9166666667	0.6666666667	0.8333333333	0.8333333333	1	0.9166666667
	Chi2	0.75	0.6666666667	0.5833333333	0.75	1	1
	F-test	0.8333333333	0.5833333333	0.6666666667	0.6666666667	1	1
delta-2	Mutual info	0.75	0.6666666667	0.75	0.75	0.5833333333	0.5
	Chi2	0.75	0.75	0.9166666667	0.8333333333	0.75	0.75
	F-test	0.6666666667	0.6666666667	0.9166666667	0.75	0.6666666667	0.5
delta-3	Mutual info	0.5	0.5833333333	0.5	0.6666666667	0.5	0.6666666667
	Chi2	0.5	0.5	0.4166666667	0.5833333333	0.5	0.5833333333
	F-test	0.6666666667	0.6666666667	0.6666666667	0.8333333333	0.5833333333	0.6666666667
Combined time point (pre + one-week + two-month)	Mutual info	0.5833333333	0.5833333333	0.5833333333	0.8333333333	0.5833333333	0.6666666667
	Chi2	0.6666666667	0.75	0.6666666667	0.8333333333	0.5833333333	0.5833333333
	F-test	0.6666666667	0.75	0.5833333333	0.8333333333	0.6666666667	0.6666666667
Combined delta (delta-1 + delta 2 + delta 3)	Mutual info	0.75	0.75	0.75	0.75	0.8333333333	0.9166666667
	Chi2	0.75	0.75	0.9166666667	0.9166666667	0.75	1
	F-test	0.8333333333	0.6666666667	0.9166666667	0.8333333333	0.75	1

Fig. 4. Comparison of Classification Results.

network with one hidden layer and 10 hidden units. (5) Gradient Boosting (GB), implemented via the scikit-learn ensemble. (6) Extreme Gradient Boosting (XGBoost), implemented using the xgboost Python library [6].

III. Results

A. Classification Results

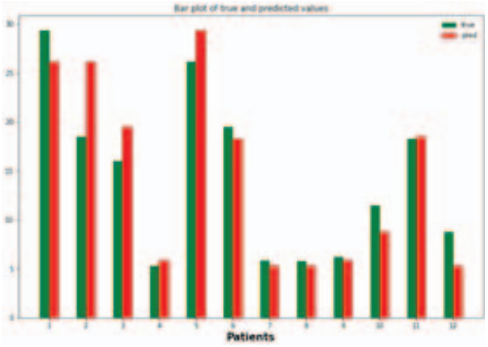


Fig. 6. Regression Results.

Four evaluation metrics are employed assess the model's performance: area under the receiver operating characteristic curve (AUC), accuracy score (ACC), true positive rates (TPR), and true negative rate (TNR). The AUC scores were notably superior for the model constructed with f-test feature selection compared to other feature selection methods. Fig 3 illustrates the impact of machine learning algorithms on model performance. The most effective model combines f-test feature selection and the gradient boosting algorithm, achieving perfect scores of 1 in AUC, ACC, TPR, and TNR with the selection of three features. Fig 3 shows the performance of the model for delta 1 dataset. The delta-1 dataset performed significantly better than the other dataset which conforms with the findings as stated in [2]. Delta features generally get a better performance than single time point features. We also compare with different datasets in Figure 4. It can be seen from the figure that delta features generally get a better performance than single time point features.

B. Regression Results

Five evaluation metrics are utilized to assess the performance of the regression models: Mean Square Error (MSE), Root Mean Square Error (RMSE), Coefficient of determination (R2), adjusted R2, and Mean Absolute Percentage Error (MAPE). Similar to the classification model, the delta features outperformed the single time point features. The most effective model emerges from a combination of the delta-1 dataset using f-test for feature selection, and the decision tree as the regression algorithm, achieving an R2 score of 0.85, adjusted R2 of 1.023, MSE of 9.508, RMSE of 3.08, and MAPE of 0.16. Fig 5 and 6 shows the plot for the best results, the regression model was able to predict the patients more accurately with shorter overall survival (OS $<$1 year) compared to patients with a longer overall survival (OS $>$1 year).

C. Discussion

The results obtained by the different classification and regression models are inconsistent with an increasing number of features. The expected outcome would be a linear relation between the number of features selected and the performance of the model. The model should perform better as the number of features increases. However, this is not the case in this study.

The construction process for both the regression and classification models follows a similar path, utilizing the same dataset and feature extraction methods. Upon comparing the performance of the regression and classification models, it is evident that both models achieve optimal results when utilizing the delta-1 data and employing the f-test as the feature selection method. However, it's noteworthy that gradient boosting and extreme gradient boosting did not exhibit the same level of effectiveness in regression as they did in classification.

IV. Conclusion

In this paper, we focus on predicting the overall survival (OS) of glioma patients undergoing treatment by employing various feature selection techniques and machine learning algorithms. Utilizing MRI scans from different time points, 112 features were extracted. The application of delta radiomics,

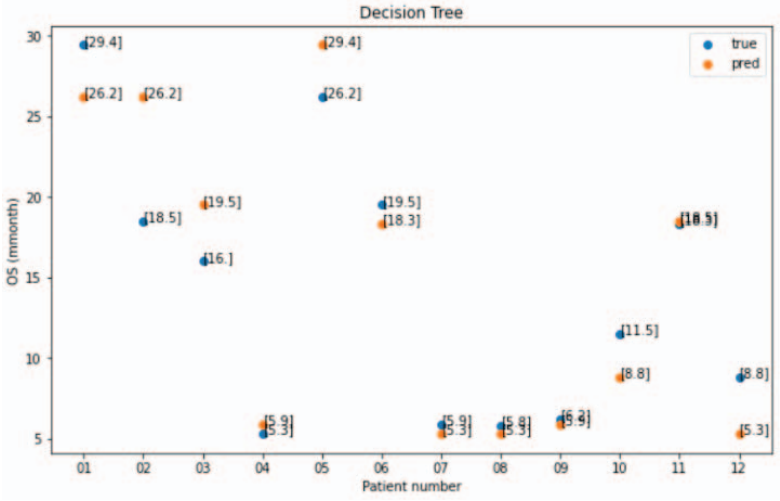


Fig. 5. Scatter Plots from Regression Results.

derived from pre and post-treatment features, involved feature selection methods like Mutual information, Chi-squared, and F-test. The classification model, built with six machine learning methods demonstrated that the F-test, particularly when combined with Gradient Boosting, outperformed other methods, achieving the highest AUC, ACC, TPR, and TNR for predicting Overall Survival (OS). This suggests the effectiveness of the chosen approach in enhancing prognostic capabilities for glioma patients.

ACKNOWLEDGEMENT

This research received support from the Agency for Science, Technology and Research (A*STAR) AME Programmatic Funds (Grant Number : A20H4b0141).

REFERENCES

[1] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563-77. doi: 10.1148/radiol.2015151169 . Epub 2015 Nov 18. PMID: 26579733; PMCID: PMC4734157.

[2] Chang Y, Lafata K, Sun Y, Wang C, Chang Z, Kirkpatrick JP, et al. (2019) An investigation of machine learning methods in delta-radiomics feature analysis. PLos ONE 14(12): e0226348

[3] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

[4] Phil, T. (2020). Sikerdebaard/dcmrtstruct2nii: v1.0.19 (v1.0.19) [Computer software]. Zenodo.

[5] Chaddad, A., Kucharczyk, M. J., Daniel, P., Sabri, S., Jean-Claude, B. J., Niazi, T., et al. (2019b). Radiomics in glioblastoma: current status and challenges facing clinical implementation. Front. Oncol. 9:374. doi: 10.3389/fonc.2019.00374

[6] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM.

[7] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830

[8] Huang, H., Zhang, W., Fang, Y., Hong, J., Su, S., Lai, X. (2021). Overall survival prediction for gliomas using a novel compound approach. Frontiers in Oncology, 11, 724191.

[9] Zhao, Rachel, and Andra Krauze. "Survival prediction in gliomas: current state and novel approaches." Exon Publications (2021): 151-169.

[10] Ye, Jianming, He Huang, Weiwei Jiang, Xiaomei Xu, Chun Xie, Bo Lu, Xiangcai Wang, and Xiaobo Lai. "Tumor Grade and Overall Survival Prediction of Gliomas Using Radiomics." Scientific Programming 2021 (2021): 1-11.

[11] Nardone, Valerio, Alfonso Reginelli, Roberta Grassi, Luca Boldrini, Giovanna Vacca, Emma D'Ippolito, Salvatore Annunziata et al. "Delta radiomics: A systematic review." La radiologia medica 126, no. 12 (2021): 1571-1583.