Foundations of Data Science

Project 2: Brain tumor imaging features

Authors
Enola Heinzer
Kaan Irmak
Pamina Lenggenhager
Catherine Rohner

Abstract

Glioblastoma is a challenging brain tumor. This study used the UPenn-GBM dataset to identify imaging biomarkers for survival prediction with machine learning models. The Logistic Regression and Support Vector Machine with L1 regularization achieved the highest accuracy at 63.7%. Issues of overfitting and limited feature selection were noted, and no significant performance differences were found across gender and age groups. Future work should focus on larger datasets and integrating multi-modal data to improve accuracy and clinical applicability.

1 Introduction

Glioblastoma is the most prevalent and aggressive form of primary malignant brain tumor, significantly impacting many patients. These tumors are typically located in the frontotemporal area, posing substantial challenges for treatment and management. (Grochans et al., 2022) Numerous studies highlight limitations in current public datasets, such as small sample sizes and inconsistent protocols, which can affect the robustness of findings (Bakas et al., 2022). The "University of Pennsylvania Glioblastoma Imaging, Genomics, and Radiomics" (UPenn-GBM) dataset addresses these issues by providing a comprehensive dataset from 611 patients, including advanced MRI scans and extensive clinical, demographic, and molecular information (Bakas et al., 2022).

1.1 Background Glioblastoma

Glioblastoma arises from the pathological proliferation of altered astrocytes, a type of glial cell in the brain. It is recognized as the most common and complex primary adult tumor of the central nervous system. The genetic alterations responsible for this proliferation are primarily due to mutations in tumor suppressor genes. These mutations lead to uncontrolled cell growth and the formation of malignant tumors. Additional causative genetic alterations are the subject of ongoing research. (Lan et al., 2024) Current treatment options for glioblastoma involve a combination of approaches to improve patient outcomes:

Operative Therapy: The primary goal of surgery is to reduce tumor mass significantly, striving for maximal safe resection to alleviate symptoms. However, this procedure is highly invasive and carries risks, including potential damage to surrounding brain tissue, which can affect neurological function. (Davis et al., 2016)

Radiochemotherapy: This combined approach uses radiation and chemotherapy to enhance treatment efficacy by targeting the tumor through different mechanisms. Radiation aims to destroy cancer cells and shrink tumors, while chemotherapy uses drugs to kill cancer cells or stop them from growing. This dual approach can be more effective but also increases the risk of side effects, such as fatigue, nausea, and susceptibility to infections. (Davis et al., 2016)

Tumor Therapy Fields: Introduced in 2014, this innovative treatment disrupts tumor cell division using electric fields. While promising, the exact mechanism remains partially understood, posing challenges for optimization. This non-invasive method involves placing electrodes around the tumor site to deliver low-intensity electric fields, potentially reducing tumor growth with fewer side effects than traditional therapies. However, patient adherence and proper device usage can be challenging. (Davis et al., 2016)

Despite expanded treatment options, overall survival for glioblastoma patients has seen little improvement over the past two decades. The invasive nature of surgeries and the severe side effects of radiochemotherapy highlight the need for less invasive, more effective treatments. (Taylor et al., 2019)

1.2 Objective of this study

This study aims to identify imaging biomarkers that can provide crucial information about glioblastoma, including the tumor's mutation status and patient survival. Traditionally, such information requires invasive biopsies. By leveraging imaging biomarkers, we aim to enhance patient outcomes and understanding of the disease, potentially establishing a new therapeutic modality and advancing medical practices. Imaging biomarkers can help in early detection, monitor treatment response, and predict patient outcomes more accurately, thereby personalizing and improving the overall management of glioblastoma.

2 Methods

2.1 Data-set Description

The dataset utilized in this project was derived from multi-phase MRI scans of glioblastoma patients, collected by the University of Pennsylvania and provided by the National Cancer Institute. Data from 611 patients were included and divided into two separate datasets. The first dataset contained 10 clinical features, including age, gender, survival days from surgery, and other glioblastoma-related biomarkers. The second dataset comprised radiomic features extracted from MRI images, including 145 characteristics describing tumor appearance in various imaging contrasts (T1, T2, FLAIR, diffusion imaging). These two datasets were merged using the patient ID present in both, resulting in a comprehensive dataset with 4762 features per data point.

The objective was to predict the survival days from surgery, using this as the label. Initially, the data were examined for any imbalances in age or gender. The dataset displayed a slight male dominance (244 females and 367 males) and a normal age distribution across genders, as shown in Figure 1.

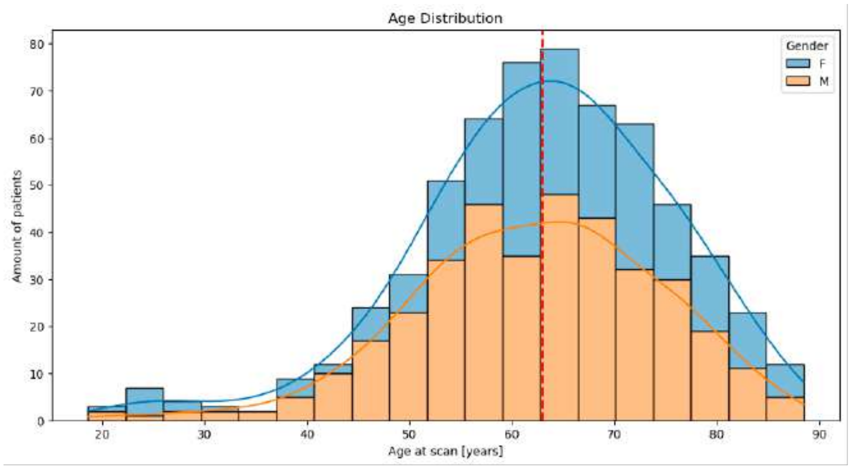

Figure 1: Plot of Gender Distribution within our Dataset. The red dotted line indicates the median.

Upon further inspection, it was found that 159 patients lacked the survival days label, leading to the removal of these data points. This decision was made to avoid skewing results due to the dataset's high variability. Predicting exact survival days proved challenging given the dataset's size and feature complexity. Therefore, survival days were split into two groups: 0 (short survival) and 1 (long survival), based on the median value of 422.0 days, as indicated by the red dotted line in Figure 2. This split resulted in 265 patients (58.6%) with long survival and 187 patients (41.4%) with short survival, making the dataset relatively balanced.

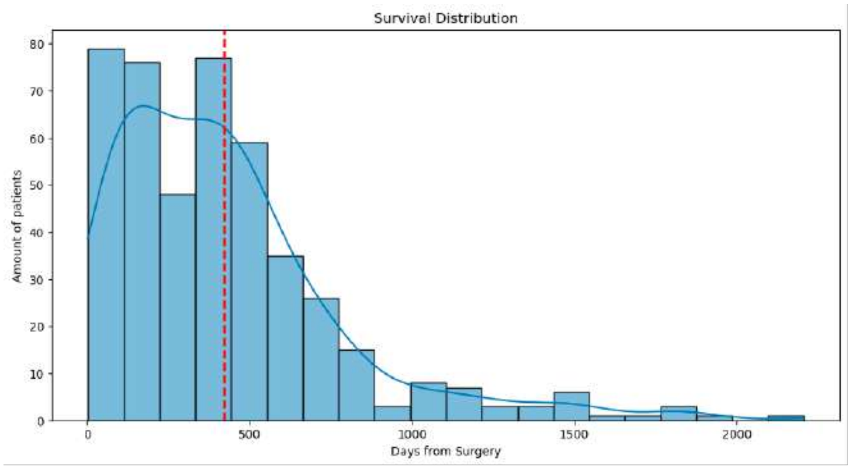

Figure 2: Plot of Age Distribution within our Dataset. The red dotted line indicates the median.

2.2 Data processing

First, the extent of missing data was assessed, leading to the decision to drop features with more than 20% missing values, as they likely lacked predictive value. The subject ID (used for merging datasets) and the survival days label were removed from the feature set, reducing the number of features to 3462.

For the remaining missing data, categorical values were filled with the mode, and label encoding was used to transform categories into integers. Missing numerical values were imputed with the column median. To further reduce dimensionality, a variance threshold of 0.01 was applied, leaving 2389 features.

An 80-20 train-test split was performed, resulting in a training set of 361 data points and a test set of 91 data points. Numerical values, excluding encoded categories and the labels, were standardized. Additionally, a balanced subset was created to test if this could improve model accuracy, considering the original dataset's slight imbalance.

2.3 Model Selection and Implementation

Four machine learning models were employed to categorize the data into 0 or 1: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting Machine (GBM). Each model's hyperparameters were optimized through grid search with 5-fold cross-validation to ensure the best performance. The balanced subset was also utilized to compare results and potentially improve accuracy.

2.3.1 Logistic Regression

Logistic Regression (LR) was selected for its inherent suitability for binary classification problems. An extended hyperparameter grid was defined to optimize performance, including varying the regularization parameter, testing different solvers, and experimenting with multiple tolerances to determine stopping criteria. L1 and L2 regularization techniques were applied to enhance model performance. L1 regularization helped in selecting the most relevant features by reducing the coefficients of less important features to zero, which is particularly useful in our high-dimensional dataset. L2 regularization prevented overfitting by penalizing large coefficients, ensuring better generalization to new data.

2.3.2 Support-Vector Machines

Support vector Machines (SVM) were chosen for their effectiveness in high-dimensional spaces and their ability to ensure a clear margin of separation between classes. An extensive hyperparameter grid was defined to optimize performance, including variations in the regularization parameter, different kernel functions, kernel coefficients, polynomial degrees, and coefficients for polynomial and sigmoid kernels. Both L1 and L2 regularization techniques were applied to improve model performance.

2.3.3 Random Forest

The Random Forest (RF) method was employed due to its ability to handle large datasets and model complex interactions between features. Hyperparameters such as the number of trees, maximum depth, minimum samples split, and minimum samples required at each node were tuned using grid search to enhance performance and reduce overfitting. As Random Forests use tree-specific parameters to manage complexity and prevent overfitting, L1 and L2 regularization are not applicable here.

2.3.4 Gradient Boosting Machine

Gradient Boosting Machine (GBM) was selected for its strength in sequentially building models to correct errors from previous iterations. Hyperparameters such as the number of estimators, learning rate, and maximum depth were fine-tuned using grid search. Similar to Random Forests, L1 and L2 regularization do not apply to decision trees and therefore were not used here.

2.4 Performance Selection

To assess the performance of the machine learning models, several evaluation metrics were utilized.

Accuracy: The proportion of correctly classified instances among the total instances was measured to provide a general sense of the model's performance.

Precision: The ratio of true positive predictions to the total predicted positives was calculated. In this context, high precision reduces the risk of false positives, which could lead to inappropriate treatment decisions by predicting a long survival when the actual survival is short.

Recall: Also known as sensitivity, recall is the ratio of true positive predictions to actual positives. It measures the model's ability to capture all relevant instances, ensuring that most patients who will indeed survive longer are identified.

Specificity: Specificity is the ratio of true negative predictions to the actual negatives. High specificity ensures that patients predicted to have short survival indeed have short survival.

F1 Score: The F1 score, which is the harmonic mean of precision and recall, was used to provide a single metric that balances both concerns. This balance is particularly useful in the context where both types of errors have significant consequences for patient care.

ROC-AUC: The ROC curve and the Area Under the Curve (AUC) were used to summarize the model's ability to discriminate between patients with long and short survival times across all possible classification thresholds. A high AUC indicates that the model is robust in distinguishing between the two classes, which is essential for reliable clinical decision-making.

3 Results

The results of the study indicate varied performance among the four machine learning models (Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting Machine) used to predict survival days for glioblastoma patients. The performance metrics evaluated include accuracy, precision, recall, specificity, F1 score, and ROC AUC. For the visual evaluation, the ROC curve was chosen.

3.1 Logistic Regression

Listed here are the performance metrics of the Logistic Regression models.

	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
LR (test)	0.604396	0.602683	0.645833	0.601986	0.602041	0.626453
LR (train)	1.0	1.0	1.0	1.0	1.0	1.0
LR selected (test)	0.604396	0.602853	0.6875	0.599564	0.598529	0.647771
LR selected (train)	1.0	1.0	1.0	1.0	1.0	1.0
LR L1 (test)	0.637363	0.644809	0.791667	0.628391	0.622596	0.709302
LR L1 (train)	0.803324	0.814262	0.921659	0.773329	0.783194	0.88252
LR L2 (test)	0.538462	0.532697	0.729167	0.527374	0.5125	0.606589
LR L2 (train)	0.783934	0.793868	0.912442	0.75136	0.760463	0.874296
LR balanced subset (test)	0.55	0.55	0.55	0.55	0.55	0.5423
LR balanced subset (train)	1.0	1.0	1.0	1.0	1.0	1.0

Table 1: Performance Metrics for Various Logistic Regression Models

3.2 Support Vector Machine

Listed here are the performance metrics of the Support Vector Machine models:

	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
SVM (test)	0.56044	0.55721	0.6875	0.553052	0.548163	0.51187
SVM (train)	0.67313	0.657794	0.746544	0.654522	0.655802	0.616615
SVM UVFS (test)	0.461538	0.438462	0.666667	0.449612	0.428113	0.539486
SVM UVFS (train)	0.98892	0.99095	1.0	0.986111	0.988392	0.999968
SVM L1 (test)	0.626374	0.649297	0.854167	0.61313	0.595873	0.705426
SVM L1 (train)	0.722992	0.747368	0.926267	0.671467	0.673126	0.748688
SVM L2 (test)	0.527473	0.514706	0.9375	0.503634	0.39957	0.611919
SVM L2 (train)	0.66759	0.821958	1.0	0.583333	0.534554	0.998432
SVM balanced subset (test)	0.58	0.580808	0.63	0.58	0.578947	0.605
SVM balanced subset (train)	1.0	1.0	1.0	1.0	1.0	1.0

Table 2: Performance Metrics for Various Support Vector Machine Models

3.3 Random Forest

Listed here are the performance metrics of the Random Forest models:

	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
RF (test)	0.549451	0.549648	0.8125	0.534157	0.502334	0.608043
RF (train)	0.99169	0.993182	1.0	0.989583	0.991304	1.0
RF balanced subset (test)	0.53	0.530048	0.51	0.53	0.529812	0.5249
RF balanced subset (train)	1.0	1.0	1.0	1.0	1.0	1.0

Table 3: Performance Metrics for Various Random Forest Models

3.4 Gradient Boosting Machine

Listed here are the performance metrics of the Gradient Boosting Machine models:

	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
GBM (test)	0.604396	0.608732	0.770833	0.594719	0.586364	0.584787
GBM (train)	1.0	1.0	1.0	1.0	1.0	1.0
GBM balanced subset (test)	0.49	0.489936	0.53	0.49	0.489183	0.4889
GBM balanced subset (train)	1.0	1.0	1.0	1.0	1.0	1.0

Table 4: Performance Metrics for Various Gradient Boosting Machine Models

3.5 Logistic Regression and Gender

Listed here are the performance metrics of the Logistic Regression model trained on gender-split datasets:

	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
LR female (test)	0.555556	0.5375	0.583333	0.541667	0.532468	0.486111
LR female (train)	1.0	1.0	1.0	1.0	1.0	1.0
LR male (test)	0.490909	0.478979	0.655172	0.481432	0.469697	0.486737
LR male (train)	1.0	1.0	1.0	1.0	1.0	1.0

Table 5: Performance Metrics for Logistic Regression L1 Separated by Gender

3.6 Logistic Regression and Age

Listed here are the performance metrics of the Logistic Regression model trained on age-split datasets:

	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
LR young (test)	0.30303	0.31015	0.263158	0.31015	0.30303	0.323308
LR young (train)	1.0	1.0	1.0	1.0	1.0	1.0
LR old (test)	0.568966	0.521104	0.771429	0.516149	0.503934	0.627329
LR old (train)	1.0	1.0	1.0	1.0	1.0	1.0

Table 6: Performance Metrics for Logistic Regression L1 Separated by Age

3.7 ROC Plots

Listed here are the Reciever Operating Curves (ROC) of the four chosen models:

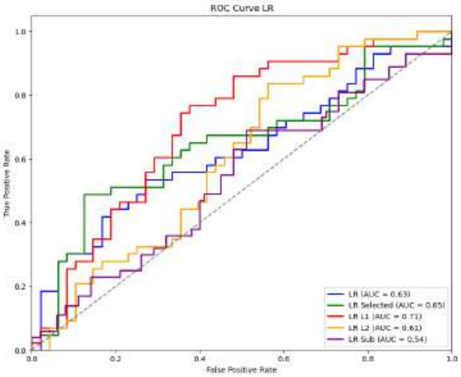

Figure 3: ROC curve for LR

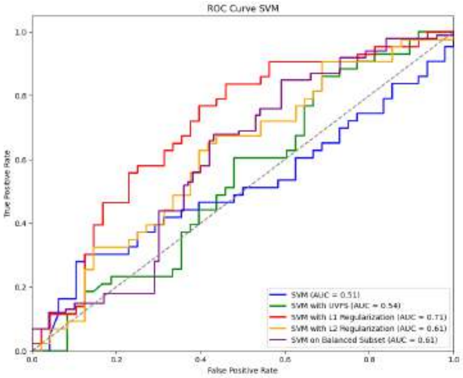

Figure 4: ROC curve for SVM

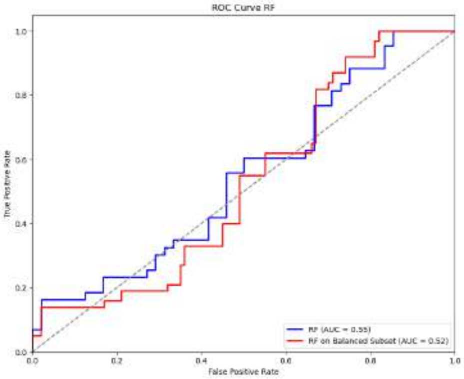

Figure 5: ROC curve for RF

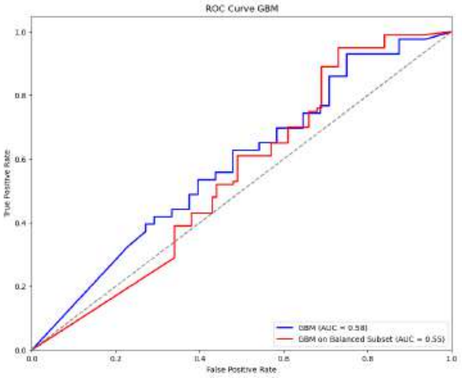

Figure 6: ROC curve for GBM

4 Discussion

The study's results indicate varied performance among the four machine learning models (Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting Machine) in predicting survival days for glioblastoma patients. The highest prediction accuracies in the test sets were achieved by both the Logistic Regression with L1 regularization and the Support Vector Machine with L1 regularization. The highest accuracy of 0.637 was attained by the LR, indicating that the model will provide a correct prediction in approximately 63.7% of cases and a ROC curve with an area of 0.709.

It is noteworthy that after performing a grid search on these models, only a very small number of features, sometimes fewer than 10, were selected as important for prediction. This suggests that much of the patient data was not taken into account, which could potentially be crucial for making an accurate life expectancy prediction. Life expectancy is influenced by many factors, such as overall health and lifestyle, which are not considered in our data.

In a final step, one of the best performing models was employed to investigate potential differences between female and male patients, as well as between young and old patients. The logistic regression model with L1 regularization was utilized for this analysis. The model was tested under different conditions, such as datasets separated by gender or age groups, to observe any performance differences.

For gender, the ROC curves covered roughly the same area, with an accuracy of 0.55 on the test set of females and 0.49 on the test set of males. This difference was considered too small to indicate a significant change in prediction accuracy. When the data were separated by age, grouping patients into those below and above 60 years of age, the logistic regression model yielded an accuracy of 0.303 for the younger group and 0.569 for the older group. This significant difference should be viewed in light of the data distribution, as there were more elderly patients, providing more data for model training, which could influence the differences in prediction accuracy.

However, these models do not perform better on age or gender groups, likely due to the limited dataset size and the complex, multifactorial nature of survival outcomes, which are not fully captured by the available features.

A significant problem encountered was the overfitting of the models. Even when balancing the datasets, which should normally help reduce overfitting, this issue was still observed in some of the training sets. This problem could potentially be mitigated if the model were trained on a larger dataset. (Johnson et al., 2019)

4.1 Comparison to previous work

In earlier studies machine- and deep learning has already been used to get an estimate of the patient's survival. Some of them also with exceptional results, such as an AUROC of approximately 0.9, and accuracies of 0.86 and 0.81 depending on if the survival outcome was a short or long timespan (Samara et al., 2021). They used a SEER dataset which was also provided by the National Cancer Institute (SEER Incidence Data, n.d.). A later study even tried using deep learning and multiclass machine learning (Babaei Rikan et al., 2024). What all the projects had in common was, that the feature which was credited as the most important in survival outcome prediction, was the age at diagnosis. This would be represented as the 'age at scan years' in our data set, which also was the most important feature in our results.

In comparison to this project, these other studies only used the clinical data of the patients. Also, they worked with approximately 34 features, which is a lot less than our dataset provided. But even then, they had access to a lot more patients, and so had a much bigger data volume. This, and their use of more detailed models, could explain the differences between our model performances.

4.2 Limitations

Despite the promising results, several limitations were identified in this study. First, the dataset's size of 611 patients may not be sufficient to capture the full variability of glioblastoma characteristics, potentially limiting the generalizability of the findings. The high dimensionality of the data (4762 features) poses challenges for model training and leads to overfitting despite the use of regularization techniques. Additionally, the removal of data points with missing survival days could introduce bias, as these missing values might be non-random. Simplifying the prediction by categorizing the survival period into just two categories is also problematic. Even if the prediction accurately indicates a shorter lifespan, the range would still vary significantly from just 1 day to 14 months, which would not be particularly helpful for patients or doctors.

4.3 Outlook

It has been demonstrated that machine learning models, particularly Logistic Regression with L1 regularization have the potential to predict glioblastoma patient survival based on imaging and clinical features, showing the importance of integrating advanced computational methods into clinical decision-making processes.

However, the initial goal of creating a model to differentiate between shorter and longer life expectancy without needing invasive procedures has not been fully realized. The accuracy of the models is not yet sufficient for clinical application, and the simplification of the prediction into just two categories (short and long survival) is not helpful for patients or doctors. Future work should focus on expanding the dataset to include more patients and exploring the integration of multi-modal data, such as genetic information and patient history, to enhance predictive accuracy. Additionally, employing more advanced models like deep learning could further improve performance. (Huang et al., 2023)

5 Bibliography

[1] Babaei Rikan, S., Sorayaie Azar, A., Naemi, A., Bagherzadeh Mohasefi, J., Pirnejad, H., Kock Wiil, U. (2024). Survival prediction of glioblastoma patients using modern deep learning and machine learning techniques. Scientific Reports, 14, 2371. https://doi.org/10.1038/s41598-024-53006-2

[2] Bakas S, Sako C, Akbari H, Bilello M, Sotiras A, Shukla G, Rudie JD, Santamaría NF, Kazerooni AF, Pati S, Rathore S, Mamourian E, Ha SM, Parker W, Doshi J, Baid U, Bergman M, Binder ZA, Verma R, Lustig RA, Desai AS, Bagley SJ, Mourelatos Z, Morrissette J, Watt CD, Brem S, Wolf RL, Melhem ER, Nasrallah MP, Mohan S, O'Rourke DM, Davatzikos C. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. Sci Data. 2022 Jul 29;9(1):453. doi: 10.1038/s41597-022-01560-7. PMID: 35906241; PMCID: PMC9338035.

[3] Samara, K. A., Al Aghbari, Z., & Abusafia, A. (2021). GLIMPSE: a glioblastoma prognostication model using ensemble learning-a surveillance, epidemiology, and end results study. Health Inf Sci Syst, 9, 5. https://doi.org/10.1007/s13755-020-00134-4

[4] SEER Incidence Data. (n.d.). Retrieved June 17, 2024, from https://seer.cancer.gov/data/

[5] Grochans S, Cybulska AM, Simińska D, Korbecki J, Kojder K, Chlubek D, Baranowska-Bosiacka I. Epidemiology of Glioblastoma Multiforme-Literature Review. Cancers (Basel). 2022 May 13;14(10):2412. doi: 10.3390/cancers14102412. PMID: 35626018; PMCID: PMC9139611.

[6] Taylor OG, Brzozowski JS, Skelding KA. Glioblastoma Multiforme: An Overview of Emerging Therapeutic Targets. Front Oncol. 2019 Sep 26;9:963. doi: 10.3389/fonc.2019.00963. PMID: 31616641; PMCID: PMC6775189.

[7] Davis ME. Glioblastoma: Overview of Disease and Treatment. Clin J Oncol Nurs. 2016 Oct 1;20(5 Suppl):S2-8. doi: 10.1188/16.CJON.S1.2-8. PMID: 27668386; PMCID: PMC5123811.

[8] Johnson, Justin & Khoshgoftaar, Taghi. (2019). Survey on deep learning with class imbalance. Journal of Big Data. 6. 27. 10.1186/s40537-019-0192-5.

[9] Huang, Y., Li, J., Li, M. et al. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. BMC Med Res Methodol 23, 268 (2023)630 . https://doi.org/10.1186/s12874-023-02078-1

[10] Lan Z, Li X, Zhang X. Glioblastoma: An Update in Pathology, Molecular Mechanisms and Biomarkers. International Journal of Molecular Sciences. 2024; 25(5):3040. https://doi.org/10.3390/ijms25053040