Report

Group 04-B

F. Abdi, A. Berger, Q. Tedjasukmana and V. Velauthampillai

The code repository for this project can be found on github and is accessible through the following link:

https://github.com/auberger/FoDS23_G04-B.git.

Abstract

Glioblastoma is a highly aggressive and malignant tumor with limited treatment options. This project aims to develop a predictive model for overall survival in patients diagnosed with glioblastoma de novo. The dataset consists of clinical and radiomic features collected from magnetic resonance imaging scans of 611 patients. Data preprocessing steps were performed, including handling missing values and feature selection. Four classifiers, including support vector machine, logistic regression, random forest, and k-nearest neighbors, were employed for model training and performance evaluation. Evaluation metrics such as accuracy, precision, recall, F1 score, and the area under the ROC-curve were used. Results showed potential correlations between age, gross tumor resection, and overall survival. Optimal numbers of features were determined for each classifier. The findings highlight the potential of radiomics and machine learning approaches in improving patient outcomes in glioblastoma de novo.

1 Introduction

Glioblastoma (GBM) is a highly aggressive and complex primary tumor of the central nervous system with limited treatment options. Despite advancements in surgery, radiotherapy, and chemotherapy, patient overall survival remains at around 16-20 months following standard care therapy, with a 5-year survival rate of only 10%. The heterogeneous nature of GBM at the molecular and micro-environmental levels poses a significant challenge in extending patient survival [2]. Genetic alterations, including mutations frequently observed in genes such as the isocitrate dehydrogenase 1 (IDH1), the isocitrate dehydrogenase 2 (IDH2) and the tumor protein 53 (TP53) play a significant role in the development and progression of glioblastoma de novo. Mutations in the IDH1 and IDH2 genes result in the production of a mutant form of the IDH enzyme that converts alpha-ketoglutarate ($alpha$-KG) into 2-hydroxyglutarate (2-HG). These mutations are disrupting cellular metabolism and contributing to tumor aggressiveness [4]. TP53 is a crucial tumor suppressor gene that regulates cell division and prevents cancer formation. Mutations in TP53 are frequently found in glioblastoma de novo, leading to loss of TP53's function in controlling cell growth and DNA repair [6].

Glioblastoma de novo presents a significant challenge in the field of oncology due to its poor prognosis and limited treatment options. With a median survival of 16-20 months and existing therapies offering only limited success, there is an urgent need to explore novel approaches that can improve patient outcomes. The emerging field of radiomics, which enables the extraction and analysis of quantitative features from medical imaging, holds great potential in developing predictive models for overall survival and treatment response in glioblastoma de novo. By integrating radiomic features with genetic information and clinical data, these models can enhance our understanding of the disease and guide personalized treatment decisions. This shift towards personalized medicine not only offers the potential for improved therapeutic outcomes but also considers the impact on patient quality of life. This project aims to analyze radiomic features derived from magnetic resonance imaging (MRI) scans to develop a predictive model for overall survival in patients diagnosed with glioblastoma de novo. The goal is to enhance the accuracy of overall survival predictions, optimize treatment strategies, and improve patient outcomes in glioblastoma de novo.

2 Methods

For the development of predictive classifiers, publicly available data derived from The University of Pennsylvania glioblastoma (UPenn-GBM) cohort was used [2]. The dataset used in this study comprises 8 clinical features and 4751 radiomic features collected from 611 patients with glioblastoma. To prepare the data for analysis, a redundant column named "SubjectID" and an irrelevant feature called "Time_since_baseline_preop" were removed. Handling missing values in the clinical features dataset involved replacing specific values, such as "Not Available", "Unknown", "Indeterminate", and "NOS/NEC" with NaN values. The datasets were carefully examined for duplicate values, which were confirmed to be absent.

Subsequently, missing values and data types were evaluated in both the clinical and radiomic features datasets. The clinical features dataset was analyzed to calculate the number of missing values per variable and their original data types. Similarly, the radiomic features dataset was assessed to determine the number of missing values per variable. Furthermore, individuals with the highest number of missing rows were identified in the radiomic features dataset. The datasets were then merged into a new dataset named "df_full". Selected columns such as "PsP_TP_score", "KPS", "MGMT," and "IDH1" were removed from the dataset due to high numbers of missing values. The mutation states "MGMT" and "IDH1" of the tumor did not get further considered since the investigation of blood samples and biopsies is often not given and this project aims to build models using radiomic and demographic data only. Furthermore, rows with missing values in the label column "Survival_from_surgery_days" and the "GTR_over90percent" column were dropped, hence 159 and 28 samples were dropped respectively. Considering the unknown importance of the "GTR_over90percent" column at this stage, this feature was retained without being dropped. It was assumed that the percentage of the tumor been resected at the first surgery could have a remarkable effect on the overall survival, but this hypothesis must be confirmed. To assess the importance of the gross tumor resection (GTR) column, a boxplot was used to compare the overall survival in days for two GTR groups (> 90% and <= 90%). The Wilcoxon signed-rank test confirmed a significant

difference (p-value = 0.000001), indicating the importance of the GTR column, which justified its retention.

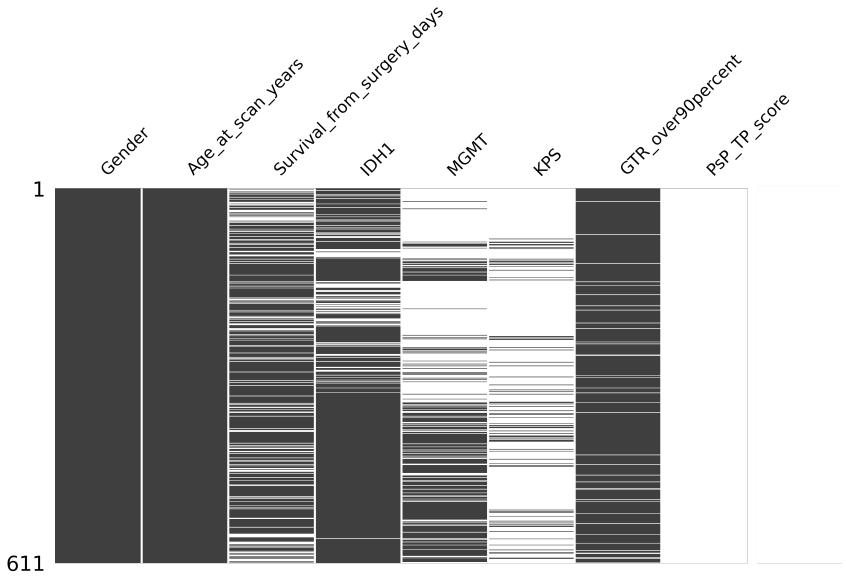

Figure 1: Missing values of clinical features

A threshold analysis was then conducted to determine the optimal number of features to drop based on missing values. The used threshold represents the minimum required non-missing values per column. In an iterative procedure, all possible thresholds got tested and columns with a certain number of missing values were dropped, while monitoring the number of patients remaining. Given the limited number of patients available, a higher threshold of 417 was set to maximize the patient cohort, resulting in 416 remaining patients.

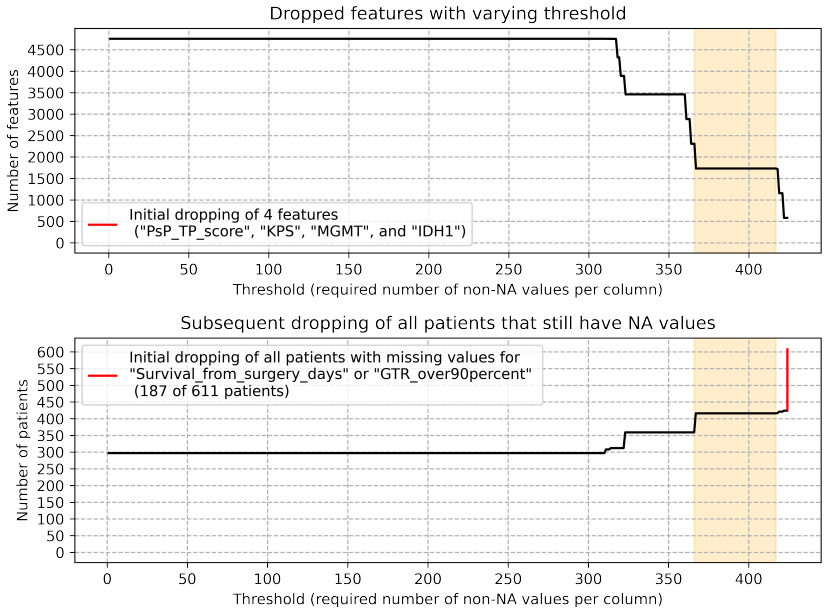

Figure 2: Number of dropped radiomic features (top) and dropped samples (bottom) with varying threshold

Next, the dataset was divided into a training set and a testing set, based on a 80:20 split. A filter-based feature selection approach was applied to the training set, serving as a first dimensionality reduction to reduce computational costs and improve interpretability. Particularly, 96 features with very low variance ($< 0.1\%$) were dropped because they fail at explaining much variation in the outcome. Additionally, highly correlated features were identified using the Spearman's correlation coefficient which is not vulnerable to outliers and capable of detecting non-linear monotonous relationships. 1302 features with a correlation coefficient greater than 0.8 got removed, reducing redundancy and enhancing the interpretability of the model. After all data preprocessing steps, the dataset consisted of 416 patients and 333 features. Descriptive statistics were computed for the dataset, and the Shapiro-Wilk test was used to assess the normality assumption for numeric variables. Given that the target variable exhibited a non-normal distribution, a binary classification approach was adopted using the median value of the "Survival_from_surgery_days" column (366 days) as the threshold for binary classification. One class represented survival of less than one year and the other class survival exceeding one year. The resulting dataset was well-balanced according to the label.

For specific feature selection, a wrapper method combining feature selection, model training, and performance evaluation using 5-fold cross-validation was employed. In an iterative procedure, all possible numbers of features got tested on each classifier individually similar to a step forward feature selection. The univariate feature selection (UVFS) method was used to select the top k features in every of the 333 iterations. The number of features that produced the best performance on the validation data was selected for all classifiers, including support vector machine (SVM), logistic regression (LR), random forest (RF), and k-nearest neighbors (KNN). This optimal number of features for every classifier ensures optimal performance while minimizing the risk of overfitting.

SVM offer robustness against overfitting and outliers, while providing versatility for handling various classification tasks, making them a powerful and flexible machine learning algorithm. LR combines the advantages of interpretable results, efficient computation, probabilistic interpretation, and robustness to noise, making it a versatile and reliable algorithm for binary classification tasks across different domains. RF's two key advantages are its robustness against overfitting, which allows for better generalization to new data, and its effective estimation of feature importance, providing insights into the relative contribution of features in making accurate predictions. Regarding the choice of models, the k-nearest neighbors (KNN) algorithm was selected for this study. It was chosen due to its interpretability and its compatibility with relatively small datasets, which aligns with the characteristics of our dataset. The KNN algorithm involves selecting the number of neighbors (K), predicting the query data point by considering the labels of its nearest neighbors, calculating the distances of K neighbors, and assigning the new data point to the majority class among its neighbors. To determine the optimum value of K, possible K values ranging from 1 to 60 were evaluated using a 5-fold cross-validation approach. The value of K with the minimum error rate, often referred to as the elbow method, was selected as the optimum K value for achieving good performance. By employing these methods, the aim is to assess the performance of the KNN algorithm in predicting the survival outcomes based on the selected features in our glioblastoma dataset [7].

The evaluation metrics used in this study included accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (ROC AUC). These metrics were chosen to provide a comprehensive assessment of the model's performance. Accuracy measures the overall correctness of the model's predictions, indicating the proportion of correctly classified instances. Precision represents the model's ability to correctly classify positive instances, emphasizing the relevance of correctly identifying true positives and minimizing false positives. Recall, also known as sensitivity or true positive rate, measures the model's ability to identify all positive instances, highlighting the importance of minimizing false negatives. The F1 score is the harmonic mean of precision and recall, providing an overall measure of the model's performance that balances both metrics. The ROC AUC metric evaluates the model's ability to discriminate between positive and negative instances across different probability thresholds. The corresponding plot describes the true positive rate against the false positive rate. A bigger area under the curve (ROC AUC score) indicates a better ability of the model to distinguish between the classes. In addition to these metrics, a confusion matrix was utilized to gain further insight into the model's performance. The confusion matrix provides a tabular representation of the predicted labels against the true labels, enabling the calculation of metrics such as true positive, true negative, false positive, and false negative rates. It offers a detailed understanding of the model's performance in terms of different types of errors and correct predictions.

Since the investigated label was balanced, no further considerations regarding the evaluation metrics had been necessary.

3 Results

Figure 3 (left) reveals that male and female patients exhibit a similar mean age, ranging from approximately 65 to 70 years. However, there appears to be an under-representation of female patients in the data set. Notably, a few female outliers exhibit a considerably young age. Overall, the distribution of patient age follows a normal distribution pattern.Upon closer examination, Figure 3 (middle and right) demonstrates that the median age of patients with an overall survival exceeding one year is lower compared to patients with an overall survival of less than one year. This finding suggests a potential correlation between age and survival outcomes. The age of patients spans a range from 20 to 88 years, with a small number of outliers falling between 20 and 30 years of age. Of particular interest is the observation that only five subjects younger than 30 years experience an overall survival shorter than one year.

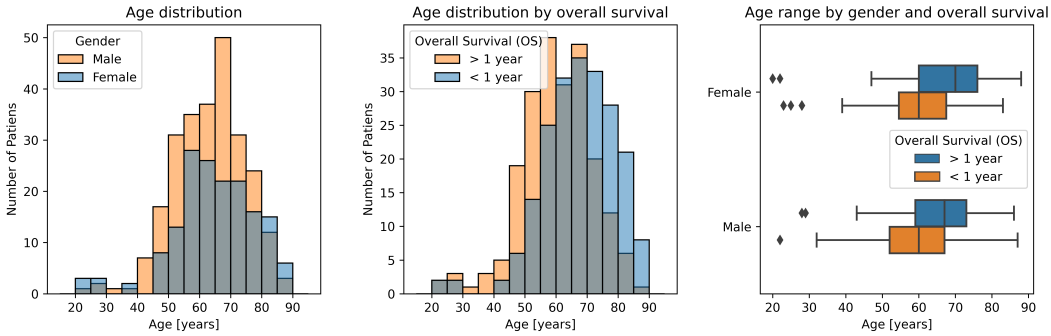

Figure 3: Age distribution by gener (left), age distribution by overall survival (middle) and age split by gender and overall survival (right)

Next, we examined the potential relationship between gross tumor resection (GTR) and overall survival to justify the inclusion of this feature. Figure 4 (left) illustrates a slight difference in median survival days after surgery among different GTR groups. Given the non-normal distribution of the label, we employed the Wilcoxon signed rank test, which detected a significant difference in the mean overall survival among GTR groups (p-value = 0.000001). This finding underscores the importance of the GTR feature as a predictor of survival outcomes. Figure 4 (right) further supports this observation, revealing that the group with less than 90% GTR has a higher likelihood of surviving less than one year (64%) compared to the group with more than 90% GTR (40.9%).

In our dataset, 252 patients underwent a gross tumor resection of more than 90%, while 164 patients had a gross tumor resection of less than 90%. Patients who underwent GTR with a resection greater than 90% exhibited a tendency to survive longer than those without.

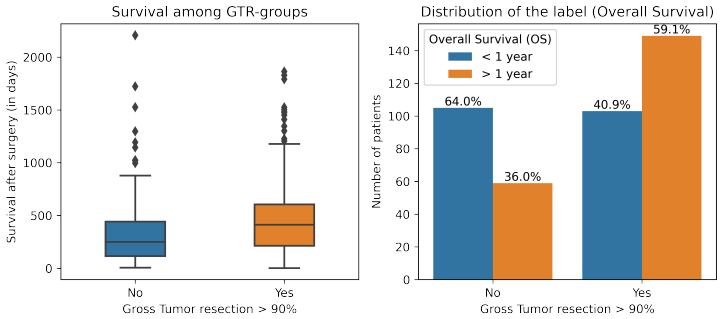

Figure 4: Overall survival (in days) among GTR-groups (left) and number of patients among GTR-groups dependent on overall survival (right)

After the preprocessing steps, our dataset consisted of 416 samples and 333 features, with a majority of the features exhibiting a non-normal distribution. Similarly, the label (overall survival) also displayed a non-normal distribution, as depicted in Figure 5. Notably, the label demonstrated a right skew, indicating a small number of subjects who had significantly longer survival durations. The range of overall survival durations varied from a minimum of 3 days to a maximum of 2000 days.

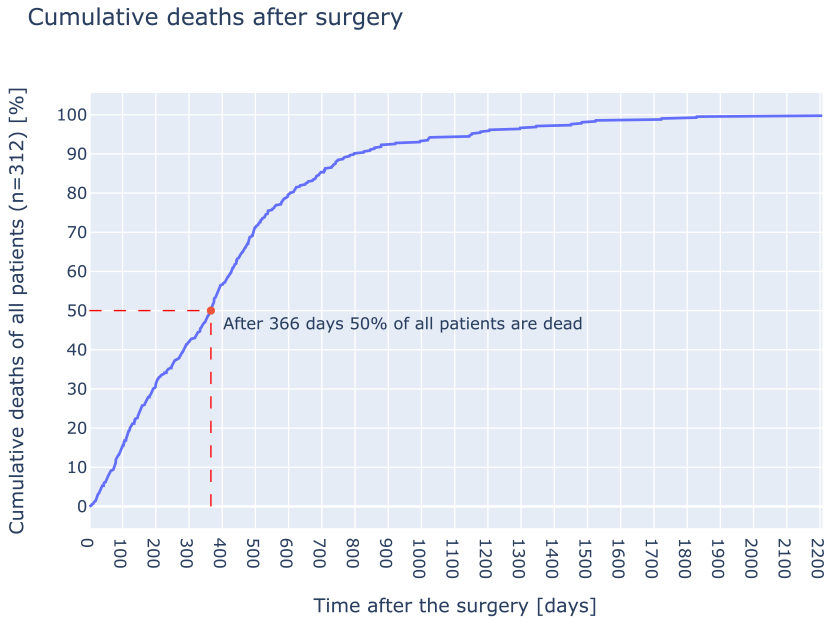

Figure 5: Cumulative death after surgery

Figure 5 also presents the cumulative percentage of patient deaths over time. The curve initially shows a gradual linear increase in cumulative deaths for the first 500 days, after which it approaches 100% in an asymptotic manner. This trend signifies that the majority of deaths occur within the first 500 days following surgery. Additionally, the median overall survival days after surgery is represented by the red line in Figure 5 and is observed to be 366 days.

By utilizing the wrapper method feature selection, we were able to identify the ideal number of features for each classifier. While Figure 6 specifically displays the validation and training scores for the LR classifier, the same procedure was followed for the SVM, RF, and KNN classifiers, although their results are not depicted here.

In Figure 6 (left), the x-axis represents the number of features used, ranging from one to 333. However, it is notable that the validation accuracy result remain relatively constant throughout this range, lacking a clear peak that would facilitate the determination of the best number of features.

To address this challenge, we introduced a criterion represented by the red line in Figure 6 (left). We selected the number of features that yielded the highest sum of the four evaluation metrics: accuracy, precision, recall, and ROC AUC. For the LR classifier, this resulted in the selection of 12 features as the optimal number.

Similarly, for the SVM, RF, and KNN classifiers, the best number of features determined through this criterion were 18, 24, and 18, respectively. However, it is worth mentioning that for all classifiers, as the number of features increased, the training accuracy approached a value of 1, indicating potential overfitting and a lack of further improvement in the validation scores.

Notably, for the KNN classifier, both the training and validation scores exhibited lower values for recall compared to the other classifiers, suggesting a potential limitation of the model in correctly identifying positive instances.

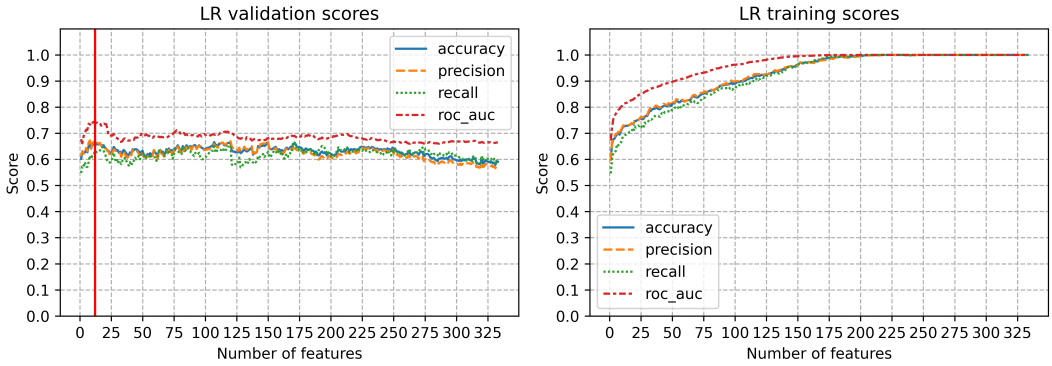

Figure 6: LR validation score (left) and training score (right) for various numbers of features

To perform manual hyperparameter tuning for the number of neighbors (K) of KNN, we utilized an elbow plot, presented in Figure 7. By employing a 5-fold cross-validation, we assessed the performance of the KNN algorithm across different values of the number of neighbors (K). As commonly observed, the error rate initially decreases with increasing K, reaches a stable point, and then starts to rise again. Figure 7 clearly illustrates this trend in the error rate. We identified the optimal K value by selecting the point where the error rate reached its minimum, which occurred at K = 41.

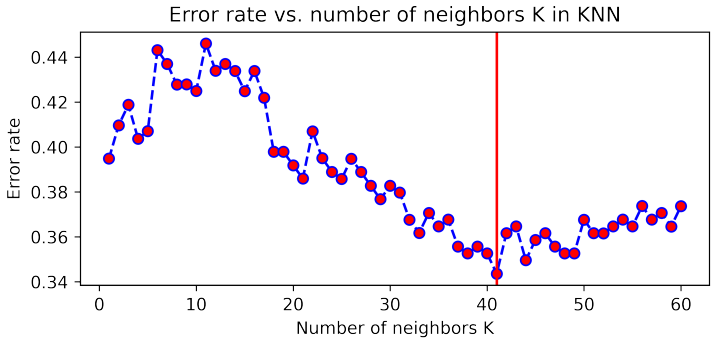

Figure 7: Error rate vs. number of neighbors K in KNN

During the final phase of evaluation, the classifiers were used to predict the overall survival on the testing data. The evaluation encompassed the analysis of the confusion matrices (Figure 8) and the ROC curves (Figure 9).

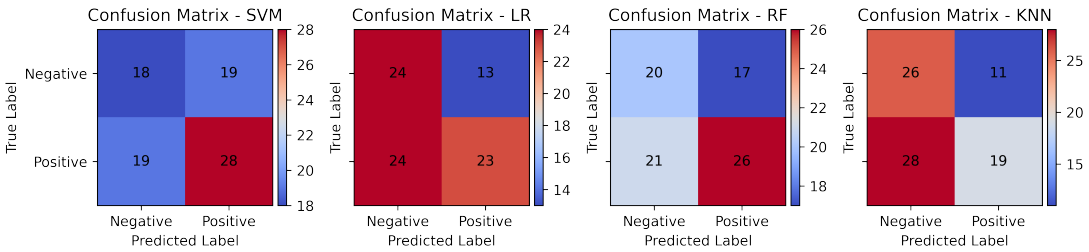

Figure 8: Confusion Matrix of SVM, LR, RF and KNN

Upon analyzing the confusion matrices (Figure 8), it became evident that the classifiers exhibited moderate to poor sensitivity in correctly identifying true positive cases. LR and KNN displayed the lowest performance in this regard. However, LR and KNN demonstrated better specificity in detecting true negatives.

The ROC curves (Figure 9) demonstrated that LR, RF, SVM, and KNN slightly outperformed a random classifier. Nevertheless, the ROC AUC scores remained relatively low and comparable among the classifiers, averaging around 0.6.

In terms of sensitivity, SVM and RF exhibited superior performance compared to LR and KNN, while LR and KNN demonstrated better specificity. Accuracy and ROC AUC scores appear to be comparable across all classifiers.

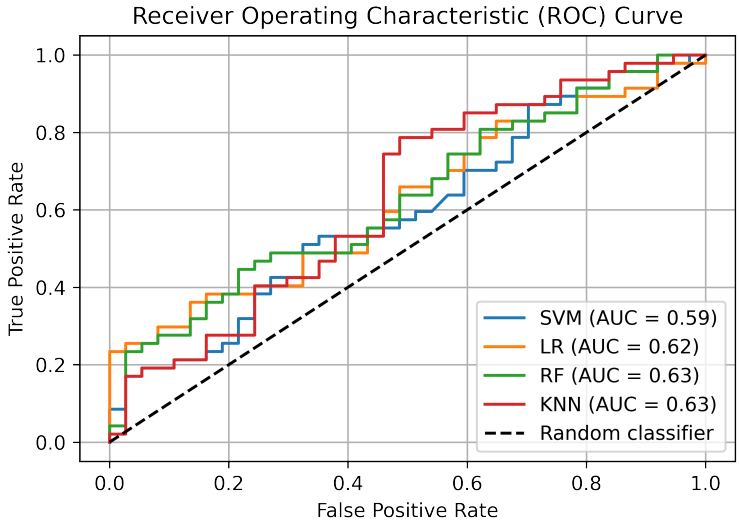

Figure 9: Receiver Operating Characteristic (ROC) Curve of SVM, LR, RF and KNN

To investigate the key features that contribute to the variation in the outcome, a feature importance analysis was conducted using the LR model (Figure 10). As anticipated, age and GTR were found to have a significant impact on overall survival and were thus included in the LR model. When considering the radiomic features, interpreting their importance becomes challenging. However, it is evident that features with names ending in "ShortRunEmphasis" appeared frequently, with totally three features. These particular features relate to the identification of short runs, which refer to

neighboring pixels with similar intensity values or other spatial relationships. Consequently, these three features, which capture textural and structural aspects of the tumor image, emerge as particularly important in predicting the outcome.

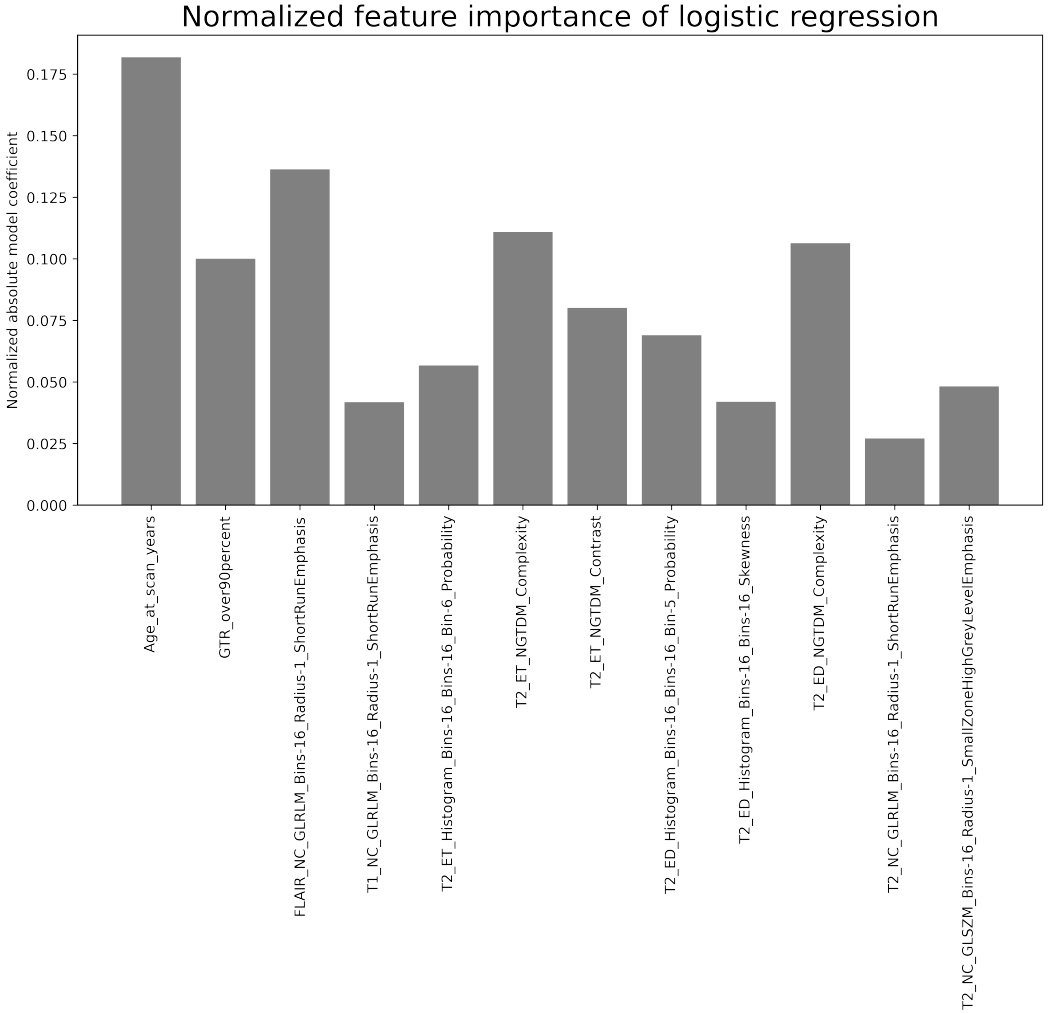

Figure 10: Feature importance of the LR classifier

4 Discussion

The results of our project provide insights into the relationship between various clinical as well as radiomic features and overall survival in patients diagnosed with glioblastoma de novo. Our findings indicate that age and gross tumor resection (GTR) are linked to survival outcomes. Specifically, we observed that younger patients had a higher likelihood of surpassing the one-year survival mark. This could be attributed to genetic factors that make glioblastoma more responsive to certain treatments or less aggressive in younger individuals [10]. Additionally, patients who underwent GTR of more than 90% exhibited longer survival rates. This observation can be attributed to the enhanced probability of eliminating cancer cells and reducing the likelihood of tumor recurrence by removing a larger portion of the tumor. [8].

Our objective with the application of the wrapper method on four different machine learning algorithms was to identify the optimal number of features that would maximize test accuracy. According to theoretical expectations, the train accuracy tends to increase, while the test accuracy typically reaches a peak before declining [5]. To achieve this, we further split the training data into a validation

set and the remaining training set. However, we observed that the validation scores of each classifier did not exhibit a clear peak, deviating from the theoretical test accuracy. On the other hand, the training accuracy aligned with the expected trend, approaching near-perfect accuracy.

Regarding the receiver operating characteristic (ROC) curve, an ideal curve represents excellent performance in distinguishing between negative and positive instances. In our analysis, the resulting ROC curve demonstrated moderate classification performance. For instance, LR achieved an area under the curve (AUC) of 0.62, indicating slight improvement over random guessing but limited accuracy in classifying negative and positive instances. RF and KNN showed similar moderate performance, while SVM performed slightly poorer in comparison.

Given the extremely low accuracy and ROC-AUC values, it is evident that all classifiers lack clinical relevance and are not suitable for practical applications. Ideally, the ROC curve should approach the left corner of the positive y-axis and continue vertically upwards, representing perfect sensitivity. Similarly, the curve should extend horizontally along the x-axis, indicating perfect specificity. The farther the distance between the ROC curve and the diagonal representing random guessing, the better the classifier's performance would be. However, in our case, the classifiers' performance fell far short of these ideal expectations. [3]

Our study has several limitations that may have contributed to the observed low accuracy. Firstly, the artificial creation of our label, which categorizes patients based on a specific time threshold, does not reflect the continuous nature of survival durations. For instance, individuals who survive only slightly less or slightly more than one year are placed in separate groups, despite their similarities. Conversely, patients who survive just over one year and those who survive more than five years are grouped together, despite their substantial differences in survival times. This artificial categorization poses challenges in accurately predicting individual patient outcomes.

Furthermore, the limited sample size of patients in our study may have impacted the generalizability and reliability of our findings. The presence of bias, such as selection bias or confounding variables, could also influence the accuracy of our classifiers. Additionally, the dropping of missing values and the presence of unreliable or incomplete feature extraction may have further contributed to the overall low accuracy observed in our results.

These limitations highlight the need for larger and more diverse datasets, as well as robust and standardized data collection methods, to improve the accuracy and clinical relevance of predictive models in this domain.

Considering the artificial label used in our study, one potential avenue for future research is to explore a multi-classification approach instead of binary classification. This would involve categorizing patients into multiple classes, such as short, middle, and intermediate survival durations, providing a more comprehensive representation of individual outcomes.

Additionally, the presence of missing values, which we addressed by dropping them due to the lack of additional information, presents another area for improvement. Future studies could consider employing advanced imputation techniques or exploring alternative strategies to handle missing data, thereby maximizing the available information and potentially improving the accuracy of the classifiers.

Furthermore, the use of convolutional neural networks (CNNs) could be of great interest in future investigations. CNNs possess the ability to simultaneously perform feature extraction and classification, potentially capturing complex patterns within radiomic data and enhancing predictive performance. Numerous studies have demonstrated the promising performance of CNN-based approaches in various medical imaging applications, suggesting their exploration in the context of glioblastoma survival prediction.[9] [1]

In summary, the utilization of a multi-classification framework, addressing missing data more effectively, and exploring the potential of convolutional neural networks present exciting directions for future research, aiming to enhance the accuracy and clinical relevance of predictive models in glioblastoma prognosis.

References

[1] Vinayak K Bairagi, Pratima Purushottam Gumaste, Seema H Rajput, and Chethan K S. Automatic brain tumor detection using CNN transfer learning approach. *Med. Biol. Eng. Comput.*, March 2023.

[2] Spyridon Bakas, Chiharu Sako, Hamed Akbari, Michel Bilello, Aristeidis Sotiras, Gaurav Shukla, Jeffrey D Rudie, Natali Flores Santamaría, Anahita Fathi Kazerooni, Sarthak Pati, et al. The university of pennsylvania glioblastoma (upenn-gbm) cohort: Advanced mri, clinical, genomics, & radiomics. *Scientific data*, 9(1):453, 2022.

[3] Jason Brownlee. Roc curves and precision-recall curves for imbalanced classification, Sep 2020.

[4] Sue Han, Yang Liu, Sabrina J Cai, Mingyu Qian, Jianyi Ding, Mioara Larion, Mark R Gilbert, and Chunzhang Yang. Idh mutation in glioma: molecular mechanisms and potential therapeutic targets. *British journal of cancer*, 122(11):1580–1589, 2020.

[5] Joshua Reini. A Practical Guide for Debugging Overfitting in Machine Learning - TruEra — truera.com. `https://truera.com/practical-guide-for-debugging-overfitting-in-ml/`. [Accessed 19-Jun-2023].

[6] Noa Rivlin, Ran Brosh, Moshe Oren, and Varda Rotter. Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis. *Genes & cancer*, 2(4):466–474, 2011.

[7] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[8] Walter Stummer, Hanns-Jürgen Reulen, Thomas Meinel, Uwe Pichlmeier, Wiebke Schumacher, Jörg-Christian Tonn, Veit Rohde, Falk Oppel, Bernd Turowski, Christian Woiciechowsky, Kea Franz, Torsten Pietsch, and ALA-Glioma Study Group. Extent of resection and survival in glioblastoma multiforme: identification of and adjustment for bias. *Neurosurgery*, 62(3):564–76; discussion 564–76, March 2008.

[9] Pallavi Tiwari, Bhaskar Pant, Mahmoud M Elarabawy, Mohammed Abd-Elnaby, Noor Mohd, Gaurav Dhiman, and Subhash Sharma. CNN based multiclass brain tumor detection using medical imaging. *Comput. Intell. Neurosci.*, 2022:1830010, June 2022.

[10] A Tortosa, Y Ino, N Odell, S Swilley, H Sasaki, D N Louis, and J W Henson. Molecular genetics of radiographically defined de novo glioblastoma multiforme. *Neuropathol. Appl. Neurobiol.*, 26(6):544–552, December 2000.