Predicting Glioblastoma Survival Outcomes from Radiomic Features

Udacity Machine Learning Engineer Capstone
Kareem A. Wahid
June 15th, 2017

Abstract

Radiomics extracts quantitative features that provide information about tumor phenotype from medical images. Accurate and reliable machine learning methods can aid in the clinical applications of radiomics. Herein, 10 classification methods are examined in terms of their performance and stability for predicting glioblastoma patient overall survival. In addition, different feature selection methods and number of features used are varied for each classifier to examine their effect upon predictive performance. Upon comparison with other similar studies, classifier performance is suggested to be highly dependent on the tumor type and imaging modality used in the radiomic analysis.

I. Definition

Project Overview

Radiomics is a budding new field of medical informatics that seeks to extract mathematically defined quantitative features such as statistics, shape, and texture from medical images [1]. Medical imaging up until recently could only provide visual qualitative information to a clinician. Radiomics allows hidden quantitative information in medical images to be deciphered and analyzed, with the potential to aid in prognosis of disease.

Cancer imaging in particular is a heavily researched area of radiomics. Tumors are often spatially and temporally heterogeneous. This frequently requires multiple tissue biopsies to be performed in order to capture the molecular heterogeneity of the tumor, which can be dangerous for the patient. Radiomics provides a non-invasive window into probing the heterogeneity of a tumor [2]. Gliomas are the most common variety of primary brain malignancies and have a high degree of intrinsic heterogeneity. This heterogeneity is apparent in their appearance and shape upon imaging, making prognosis difficult [3]. Radiomic analysis of glioma medical imaging can provide additional information about a patient's prognosis and likely survival outcomes [4-6].

Though significant research has been conducted on the application of machine learning algorithms to radiomic features for prognostic prediction [7-10], there is still much that is unknown about which models are best due to lack of standardization in the field. Herein, supervised machine learning algorithms were trained on radiomic features extracted from glioblastoma magnetic resonance images obtained from the 2017 BraTS Challenge to predict patient overall survival outcomes. The popular Python machine learning library scikitlearn was used for all calculations. Classifier performance was evaluated using the area under the receiver operator curve metric and compared to previously published data.

Problem Statement

The Multimodal Brain Tumor Segmentation (BraTS) Challenge [11] is an annual competition that seeks to employ the brightest minds in computational radiology to develop the following:

1. Accurate segmentation algorithms of gliomas in magnetic resonance images.
2. *Prediction of patient overall survival through radiomics and machine learning algorithms.*

It is the goal of this project to address the second listed objective of the BraTS Challenge. In this study we will determine which supervised machine learning classification models are the most suitable for predicting prognostic information from radiomic features of glioblastoma magnetic resonance imaging (MRI) scans acquired from the 2017 BraTS Challenge. A variety of supervised classification methods trained on radiomic feature data and known prognostic outcomes will be compared through quantifiable metrics to determine which method most accurately predicts the prognostic class for a set of new patients unseen radiomic feature inputs. In addition, different

feature selection methods and the number of features used for training classifiers will be varied to examine their effect on classifier predictive performance.

Metrics

The Receiver Operating Characteristics (ROC) curve is a commonly utilized metric to evaluate binary classifier output performance [12]. ROC curves typically display true positive rate on the Y-axis and false positive rate on the X-axis as discrimination threshold is varied. The area under the ROC curve, AUC, can be used to quantify the degree to which a model is able to accurately classify data. AUC values close to 0.5 are worse, signifying random guessing, while AUC values close to 1.0 are better, signifying perfectly accurate classification. Herein, we will use AUC values to measure the predictive performance of our classification method/feature selection method combinations, and compare them to the results in Parmar et al. [7] (detailed in the benchmark section of Analysis).

Desirable models should be stable in response to small perturbations to the input space. Classifier stability will be measured via the relative standard deviation (RSD); a metric described in Parmar et al. RSD can be defined as:

$$RSD = \frac{\sigma_{AUC}}{\mu_{AUC}} X\ 100$$

where σ_{AUC} and μ_{AUC} are the standard deviation and mean of the AUC values generated from repeated subsamples of the training data using a bootstrap approach. In this study, train test split will be iterated 10 times with different random states to generate σ_{AUC} and μ_{AUC}. These calculations in relation to our dataset are demonstrated in the Implementation section of Methodology.

II. Analysis

Data Exploration

The BraTS Challenge provides a standardized MRI dataset of patients diagnosed with gliomas from the publically available Cancer Imaging Archive; tumor volumes have been segmented by expert radiologists and are included in the dataset. An example of a high-grade glioblastoma (GBM) MRI scan and corresponding tumor segmentation from the BraTS dataset is visualized in Figure 1. In addition the BraTS Challenge provides overall survival data for patients corresponding to these MRI scans. The MRI dataset and patient survival data from the 2017 BraTS Challenge was obtained for this project by entering the competition. More details about the BraTS Challenge can be found at: http://braintumorsegmentation.org/.

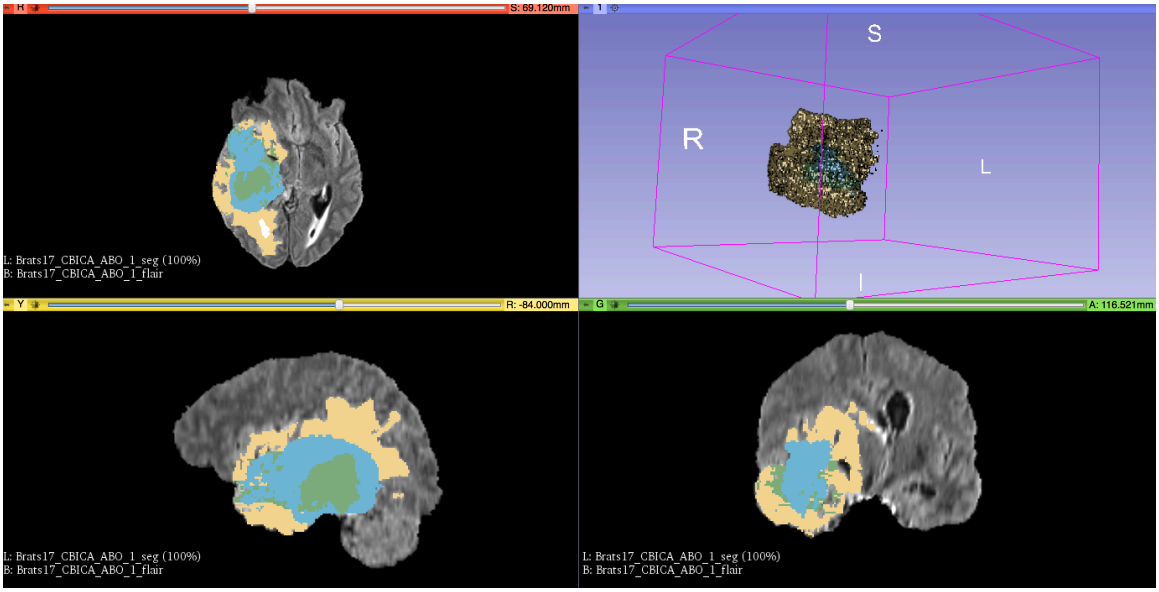

Figure 1: Visualization of GBM FLAIR MRI scan with corresponding tumor segmentation obtained from BraTS dataset. Tumor segmentation represented by colored areas. Color-coding represents subregions of tumor but will not be considered separately in this study. Tumor segmentation 3D volume is visualized in top right corner. Visualization performed in 3D Slicer [13].

The BraTS Challenge offers the number of days survived after diagnoses (overall survival) for 163 GBM patients. The patient overall survival data are continuous numerical values. The distribution of the patient survival data is plotted in Figure 2a. The survival data has a leftward skew with a mean survival centered around 380 days. Since the patient survival is our output, it is necessary to transform the continuous values to categorical values for use in classification. The data has been partitioned into two categorical values using a cutoff time of one year, which is based on median survival rates for glioblastoma [5]; patients who live less than one year are assigned a 0, patients who live longer than one year are assigned a 1. These patients will be referred to as short-term and long-term survivors throughout this manuscript. These transformed overall survival categories will serve as the output for our models. Partitioning the survival data using a cutoff time of one year yields approximately equal instances of short-term and long-term classes, which makes subsequent analysis more manageable (Fig 2b).

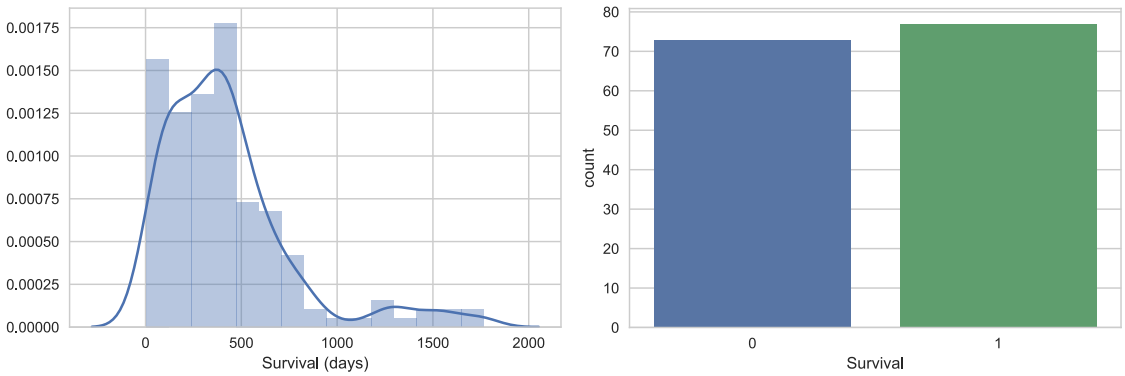

Figure 2: Survival outcome distribution of BraTS Challenge GBM dataset before (a) and after (b) partitioning.

94 radiomic features corresponding to statistics (19), shape (16), and texture (59) were extracted from each of the 163 GBM MRI images/segmentations by implementing a standardized open-source radiomics python library, PyRadiomics [14], with default parameter configuration. The inputs to the radiomics algorithm were FLAIR MRI scans as "Images" and corresponding tumor segmentations as "Masks". 13 of the image/mask combinations suffered from geometry mismatch, so these samples were discarded. Statistical information for five of the radiomic features extracted is shown in Table 1. A detailed list of the features extracted can be found in the PyRadiomics documentation at http://pyradiomics.readthedocs.io/en/latest/.

			Textural Features		
	Variance (statistics)	**Volume (shape)**	**Autocorrelation (GLCM)**	**Small Area Emphasis (GLSZM)**	**Short Run Emphasis (GLRLM)**
count	150	150	150	150	150
mean	8643.94	11856.81	190.71	0.55	0.81
std	14328.38	14355.17	215.69	0.08	0.10
min	20.58	47.00	5.11	0.33	0.31
25%	2034.99	3274.75	67.45	0.50	0.75
50%	4273.07	7820.50	118.01	0.56	0.84
75%	8120.41	15919.75	240.79	0.61	0.88
max	98968.98	91299.00	1329.48	0.74	0.96

Table 1: Sample set of radiomic features derived via PyRadiomics for the 5 feature categories. GLCM = Gray Level Co-occurrence Matrix, Gray Level Size Zone Matrix = GLSZM, Gray Level Run Length Matrix = GLRLM are textural features.

The radiomic feature data and categorical survival data have been collated to yield the final dataset that was used for this project. The dataset contains 150 rows corresponding to the number of patients, and 95 columns corresponding to the radiomic features and the overall survival category (short-term: 0, or long-term: 1). The features are the input to our models while the overall survival category is the output of our models. A shortened version of the dataset with the first 5 samples showing the survival data, first 2 features, and last 2 features is shown in Table 2.

	Survival	**Maximum 3D Diameter**	**Compactness 2**	**...**	**Zone Entropy**	**Small Area Low Gray Level Emphasis**
0	0	72.36712	0.022687	...	5.630367	0.017143
1	0	51.584882	0.032489	...	5.748432	0.03035
2	0	63	0.038263	...	6.093996	0.014043
3	1	34.655447	0.397477	...	6.182728	0.00722
4	1	21.470911	0.076978	...	4.051376	0.120535

Table 2: Shortened list of final dataset used for project showing overall survival, first two features (Maximum 3D Diameter, Compactness 2) and last two features (Zone Entropy, Small Area Low Gray Level Emphasis) of the total set of 94 features.

Exploratory Visualization

One may intuit that larger tumor size correlates to a worse prognostic outcome for GBM. The reality is not as clear-cut as our intuition would lead us to believe. To examine this concept, 2 radiomic shape features corresponding to tumor size (Volume and Maximum 3D Diameter) are plotted for each sample with corresponding survival data in different colors in Figure 3. On average short-term survivors tend to have slightly larger volume and diameter when compared to long-term survivors. However, no obvious separation boundary exists between short-term and long-term survivors; survival classes are intermixed between high and low values for tumor volume and diameter. This highlights the fact that GBM is spatially heterogeneous and survival outcome would be difficult to predict from visual inspection of tumor size alone. The combination of statistic, shape, and texture based radiomic features should lead to the generation of a multi-dimensional decision boundary that can differentiate short-term survivors from long-term survivors.

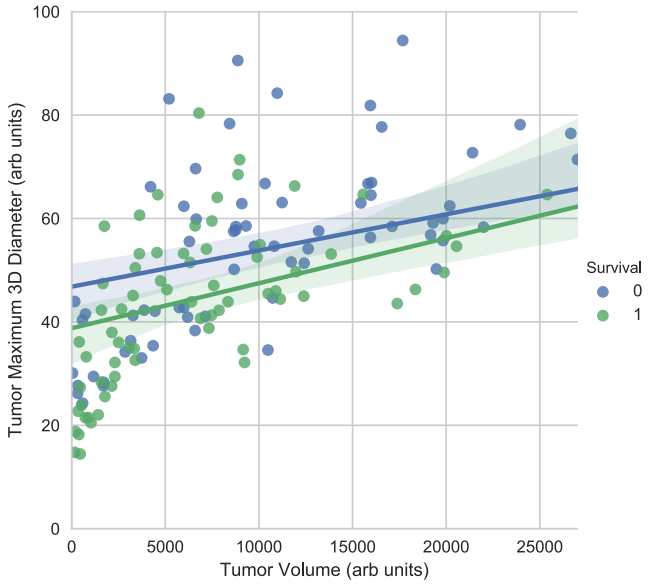

Figure 3: Relationship between two shape features (Volume, Maximum 3D Diameter) and survival classes.

Algorithms and Techniques

10 supervised machine learning classification algorithms from the popular python machine learning library scikitlearn (sklearn) were selected for examination. These algorithms were chosen in accordance with those studied in Paramer et al and are listed in

column 1 of Table 3. These algorithms use the radiomic features as inputs and survival classes as outputs to build a predictive model.

The feature selection method SelectKBest with 3 different scoring functions was utilized to select the top 10, 20, 30, and 40 features in a systematic fashion for each classifier. SelectKBest is a univariate feature selection method that utilizes a scoring function to assign values to each feature and removes all but the K highest scoring features from the feature set. It selects the top features that have the most relevance to the target variable, in this case survival outcome. The selection process is classifier independent so can be considered a filter method [15]. The 3 scoring functions chosen are f_classif which represents the ANOVA F-score, mutual_info_classif which represents mutual information, and chi2 which represents a chi-squared statistic. Each classifier was then trained using the selected feature subsets.

Hyperparameter tuning for each classifier/feature selection combination was accomplished through a stratified 5-fold cross validation grid search via GridSearchCV. GridSearchCV systematically performs an exhaustive search over a constructed hyperparameter grid to select the best hyperparameters for each classifier/feature selection method. Stratification creates folds that preserve the percentage of samples for each class and leads to models with lower bias and variance when compared to regular cross validation [15]. The hyperparameter tuning values for each classifier were chosen in accordance with Parmar et al and are shown in column 2 of Table 3. Classifiers without hyperparameters such as Gaussian Naïve Bayes and Quadratic Discriminant Analysis skipped the tuning step. It should be noted that in order to expedite computational time, some hyperparameter tuning values were shortened from Parmar et al. For example, the n_splits parameter for Random Forrest in Parmar et al was set to values of 2 to 30 with step size of 3, but in our study we only utilized values of 2, 20, and 30.

Classifiers	**Hyperparameters**
Random Forrest	n_estimators: 500 min_samples_split: [2,20,30]
Gaussian Naïve Bayes	None
Decision Tree	min_samples_split: [2,20,30]
Multi Layer Perceptron	hidden_layer_sizes:[(100,),(1,),(9,)] beta_1: [0.9, 0.1, 0.0001]
Bagging	n_estimators: [5,10,100]
Gradient Boosting	n_estimators: [1,10,20,100]
Support Vector Machines	C: [0.1,1.0,10.0] kernel: ['poly', 'rbf']
Logistic Regression	C: [0.1,1.0,10.0], penalty: ['l1','l2']
K Nearest Neighbors	n_neighbors: [5,10,15,20]
Quadratic Discriminant Analysis	None

Table 3: List of sklearn classifiers used in this study (column 1) along with their hyperparameter tuning values (column 2).

Benchmark

Parmar et al. published a landmark study comparing fourteen feature selection methods and twelve classification methods in terms of their performance and stability for predicting overall lung cancer patient survival [7]. This study contains AUC values for each of the feature selection and classification combinations, which can be directly compared with our projects AUC values. In addition, they developed a method for evaluating classifier stability, RSD, which can also be directly compared with our classifiers.

Though the models in Parmar et al. were trained using lung cancer computed tomography imaging and a different survival classification threshold (2 years instead of 1), it can be inferred that underlying radiomic principles are similar regardless of the tumor type studied or imaging methodology. Therefore, it is logical to predict our classification methods will follow comparable predictive performance trends. However, this may not necessarily be the case, as it has been shown in the past that different classifiers will work better for different cancer types and imaging modalities [8]. It should be noted that Parmar et al. also utilizes a larger training set than what is available to us (310 vs. 120 samples), so our models may not be as accurate, stable, or generalizable as theirs.

III. Methodology

Data Preprocessing

As previously discussed, 13 of the 163 samples displayed geometry mismatches, leading to the generation of 'NaN' values in the resulting dataset in place of feature values. These samples were removed from the dataset leading to a remaining total of 150 samples in the dataset. Furthermore, survival data was mapped from continuous values to categorical values of 0 and 1 for short and long-term survivors, respectively. Additionally, feature values were normalized with respect to the training set, described in more detail in the implementation section. No outliers were removed from this dataset since the BraTS Challenge has ensured the correct segmentation and prognostic outcome of each patient, making every data point valuable for the machine learning process. Different feature selection methods coupled with different numbers of top features selected were implemented and are detailed in the implementation section.

Implementation

The dataset of 150 samples, each with 94 features and a corresponding output class, was split into training and testing sets with a test size = 0.2, yielding a training set with 120

samples and a testing set with 30 samples. MinMaxScaler was used to normalize features with respect to the training set features and subsequently applied to the test set features. For each classifier, hyperparameters were tuned using GridSearchCV. 12 different feature selection combinations were applied to each classifier corresponding to the 3 scoring functions (ANOVA F-score, Mutual Information, Chi-sqr) with 10, 20, 30 or 40 top features selected via SelectKBest. AUC scores were then computed for each of the classifier/feature selection combinations. Since 10 classifiers were investigated in this study, a total of 120 AUC scores corresponding to each classifier/feature selection combination were generated.

The entire process outlined above was iterated 10 times with train test split in different random states. The mean of the AUC values over all iterations was calculated to determine the final AUC values for each classifier/feature selection combination. Averaging AUC values after repeat iterations of train test split were necessary because our models were observed to be somewhat unstable upon perturbation. By calculating the mean over 10 iterations we are able to ensure a more representative value for each classifier/feature selection combination. Similarly, the standard deviation of the AUC values over all iterations was calculated to determine the final AUC standard deviation for each classifier/feature selection combination.

The mean of the final AUC values for each classifier/feature selection combination for a given classifier was calculated to determine the μ_{AUC} for the given classifier. The mean of the final standard deviation values for each classifier/feature selection combination for a given classifier was calculated to determine the σ_{AUC} for the given classifier. RSD for each classifier was calculated by dividing μ_{AUC} by σ_{AUC} for that classifier then multiplying by 100 as described previously in the Metrics section of Definition. Python code for the entire implementation is documented in the accompanying Jupyter notebook.

Refinement

Initially only 4 classification methods described in Parmar et al. were utilized (Random Forest, Naïve Bayes, Decision Trees, Multi Layer Perceptron). The number of classifiers used was increased to 10 (Table 3, column 1) in order to more closely model Parmar et al. Additionally, only untuned models with default hyperparameters were implemented at first. To increase classification performance, recreations of the tuning parameters from Parmar et al. (Table 3, column 2) were implemented via k fold cross validation as described in the Algorithms and Techniques section of Analysis. Hyperparameter tuning marginally increased classification performance of all applicable models but greatly increased computational time. Moreover, μ_{AUC} for Random Forrest, Decision Tree, Bagging, Gradient Boosting, and Support Vector Machines increased from 0.584, 0.578, 0.593, 0.607, and 0.533 to 0.613, 0.587, 0.611, 0.615, and 0.557 respectively.

IV. Results

Model Evaluation and Validation

An AUC value mean, median, and standard deviation list of the 10 iterations for each classifier/feature selection combination is shown in table 4. A full table for each classifier/feature selection combination can be found in the accompanying Jupyter notebook.

	Classifier	Feature Selection Method	Feature number	Mean	Median	Standard Deviation
0	RandomForestClassifier	ANOVA F-score	10	0.58	0.58	0.08
1	GaussianNB	ANOVA F-score	10	0.56	0.59	0.09
2	DecisionTreeClassifier	ANOVA F-score	10	0.59	0.59	0.07
3	MLPClassifier	ANOVA F-score	10	0.58	0.58	0.06
4	BaggingClassifier	ANOVA F-score	10	0.58	0.53	0.10
5	GradientBoostingClassifier	ANOVA F-score	10	0.58	0.59	0.11
6	SVC	ANOVA F-score	10	0.58	0.55	0.08
7	LogisticRegression	ANOVA F-score	10	0.57	0.56	0.06
8	KNeighborsClassifier	ANOVA F-score	10	0.57	0.58	0.05
9	QuadraticDiscriminantAnalysis	ANOVA F-score	10	0.50	0.51	0.08
10	RandomForestClassifier	ANOVA F-score	20	0.60	0.62	0.06
…	…	…	…	…	…	…
110	RandomForestClassifier	Chi-sqr	40	0.62	0.62	0.07
111	GaussianNB	Chi-sqr	40	0.51	0.50	0.06
112	DecisionTreeClassifier	Chi-sqr	40	0.56	0.55	0.07
113	MLPClassifier	Chi-sqr	40	0.57	0.57	0.06
114	BaggingClassifier	Chi-sqr	40	0.62	0.60	0.09
115	GradientBoostingClassifier	Chi-sqr	40	0.64	0.65	0.08
116	SVC	Chi-sqr	40	0.56	0.56	0.09
117	LogisticRegression	Chi-sqr	40	0.55	0.55	0.06
118	KNeighborsClassifier	Chi-sqr	40	0.54	0.57	0.08
119	QuadraticDiscriminantAnalysis	Chi-sqr	40	0.52	0.50	0.07

Table 4: Shortened AUC statistics list for each classifier/feature selection method combination.

Figure 4a depicts a heatmap of the performance of feature selection in rows and classification methods in columns. The AUC value for each combination was obtained by averaging over the feature number variations, i.e. mean of 10, 20, 30, 40 top features for each classifier. Similarly, Figure 4b depicts a heatmap of the performance of feature number in rows and classification methods in columns where the AUC value for each combination was obtained by averaging over the feature method variations, i.e. mean of ANOVA F-score, Mutual Information, and Chi-squared for each classifier. The heatmaps demonstrate noticeable difference between classifiers but minimal difference upon altering the feature selection method (Fig. 4a) or number of features used (Fig. 4b).

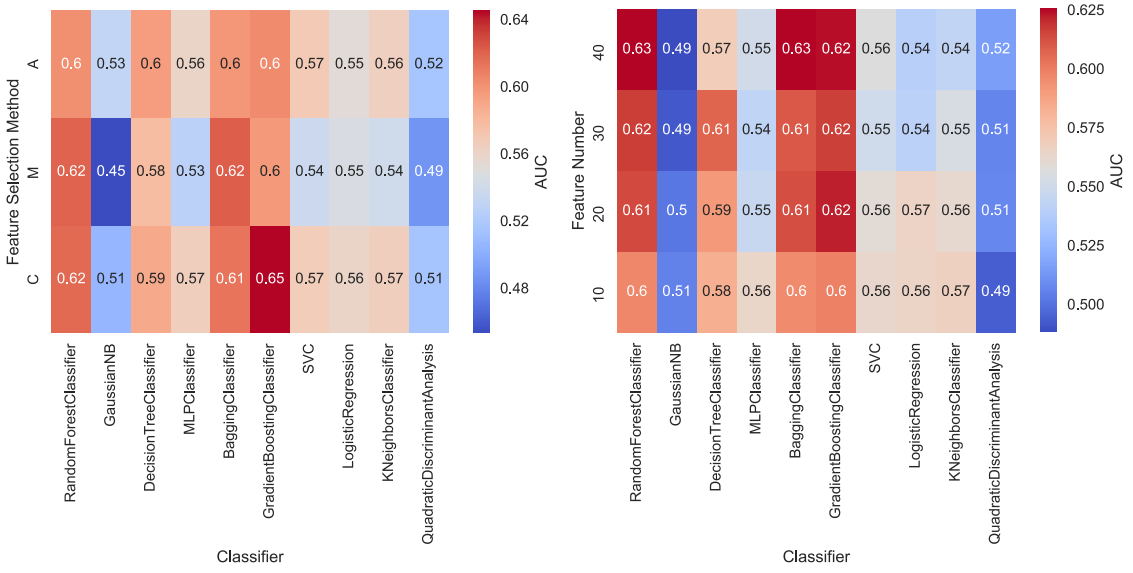

Figure 4: (a) AUC values of feature selection methods in rows and classifiers in columns. A = ANOVA F-score, M = Mutual Information, C = Chi-squared. (b) AUC values of feature numbers in rows and classifiers in columns.

By averaging over all feature selection combinations we obtain the final representative AUC value for each classifier (μ_{AUC}). In addition RSD values for each classifier can be calculated as described in previous sections. A full table of these values can be found in the accompanying Jupyter notebook. To help visualize predictive performance and stability of our classifiers a scatterplot of predictive performance (μ_{AUC}) vs. stability (RSD) for each classifier is shown in Figure 5.

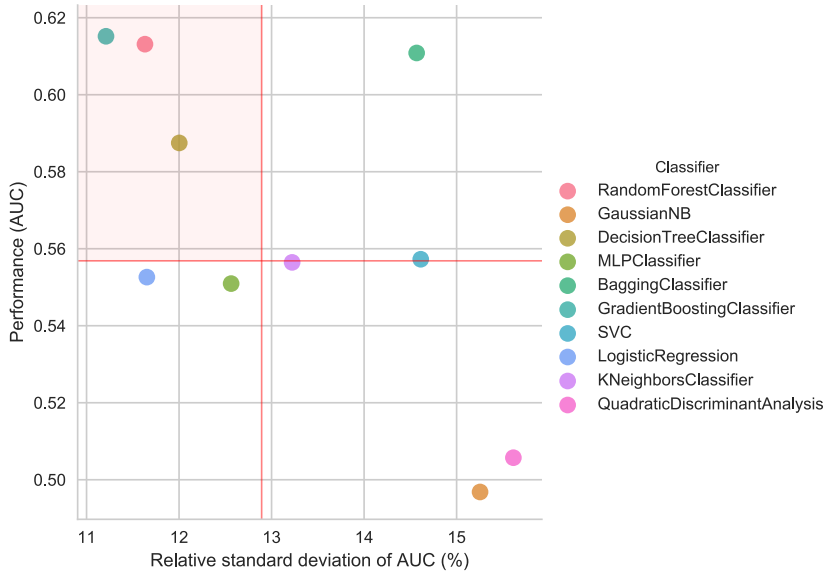

Figure 5: Scatterplot of AUC score vs. RSD score. Red lines represent median AUC (0.56) and median RSD (12.9) of all classifiers. Red area in upper left corner represents classifiers with high predicative performance and high stability.

Random Forest (AUC = 0.61, RSD = 11.6), Gradient Boosting (AUC = 0.62, RSD = 11.2), and Decision Tree (AUC = 0.59, RSD = 12.0) have predictive performance higher than the median values (AUC = 0.56) and RSD lower than the median values (RSD = 12.9). Therefore, these classifier methods should be considered to have relatively high predictive performance and stability when compared to the other classifiers studied. Gaussian Naïve Bayes (AUC = 0.50, RSD = 15.3) and Quadratic Discriminant Analysis (AUC = 0.51, RSD = 15.6) showed the worst predictive performance and stability with AUC values lower than the median and RSD values higher than the median. Some classifiers such as Bagging (AUC = 0.61, RSD = 14.6) showed predictive performance above the median AUC value but stability above the median RSD value, indicating high predictive performance but low stability. Oppositely, some classifiers such as Logistic Regression (AUC = 0.55, RSD = 11.7) and Multi Layer Perceptron (AUC = 0.55, RSD = 12.6) demonstrate predictive performance below the median AUC value and stability above the median RSD value, indicating low predicative performance but high stability.

Justification

Overall, classifiers in our study show lower predictive performance and higher RSD values when compared to their counterparts in Parmar et al. For example, Random Forest was shown to have an AUC and RSD of 0.66 and 3.5 respectively in Parmar et al, which is higher in predictive performance and stability when compared to our own Random Forest implementation. This is possibly due to a smaller training size utilized in our study when compared to Parmar et al.

Random Forest was shown to be one of the best classifiers in Parmar et al and this held true for our data as well. Interestingly, some of the classifiers demonstrate the opposite behavior that was observed in Parmar et al. For example, Naïve Bayes was shown to be one of the best classifiers in terms of both predictive performance and stability in Parmar et al. but was shown to be the worst for both these values in our study. Additionally, Decision Tree and Boosting were shown to have relatively poor predictive performance and stability in Parmar et al., but were among the best models in our study. This is a result that has been observed in other studies by Parmar et al. [8]. This potentially highlights the idea that the best classification methods for radiomics studies are highly dependent on the imaging modality and the type of tumor being studied.

V. Conclusion

Free-Form Visualization

In this study there are three experimental factors that may affect the prediction of radiomics based survival outcomes: classification method, feature selection method, and

number of features used. In addition, the interactions between these factors also have a role in predictive performance. Multivariate analysis of variance (ANOVA) was performed to determine the variability in AUC scores contributed by each of the experimental factors and their interactions. The AUC scores used for each classifier/feature selection combination were the mean values after 10 iterations of the train test split procedure shown in table 4. In order to compare the variability contributed to predictive performance by each factor, the variance (sum of squares) calculated for each factor was divided by total variance and multiplied by 100 to yield the percent variance for each factor (Fig 6).

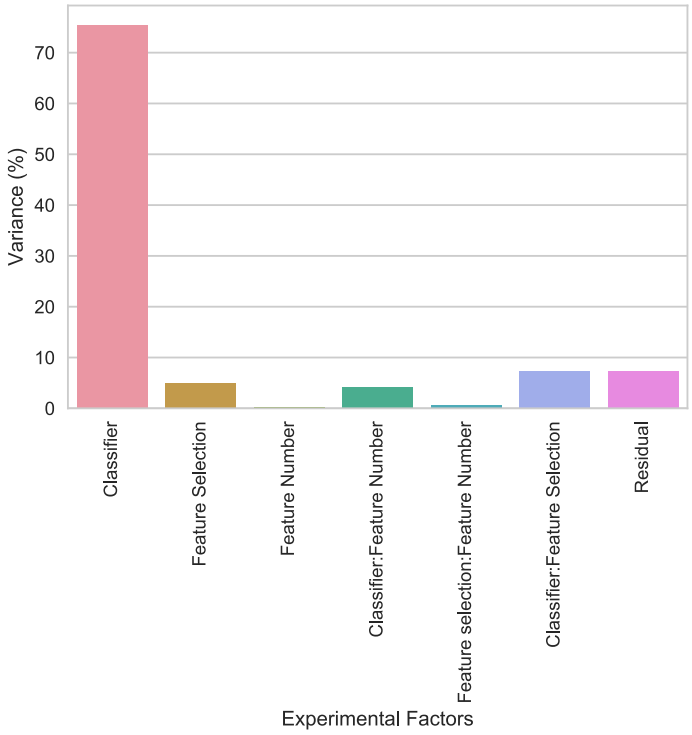

Figure 6: Variation of AUC explained by the experimental factors and their interactions.

Classification method had the largest contribution to variability accounting for 75.5 % of the total variance in AUC scores with a corresponding p-value of 8.0e-25 indicating it was a statistically significant factor contributing to variability ($p < 0.05$). Feature selection method accounted for 4.6 % of the variability with a corresponding p-value of 3.2e-6 indicating it was a statistically significant factor contributing to variability ($p < 0.05$). Finally, the number of features used in a model accounted for the least of the total variance at 0.3 % with a p-value of 5.2e-1, indicating it was not a statistically significant factor contributing to variability ($p > 0.05$). Interactions between factors followed similar trends. Classifier: feature number, feature selection: feature number, and classifier: feature selection accounted for 3.9, 0.9, and 7.3 % of the variability respectively with corresponding p-values of 4.5e-1, 4.1 e-1, and 1.5e-3 respectively. Since we used many more classifier methods than any other experimental factor these

results could be anticipated but were still interesting to see quantified. The results are in accordance with Parmar et al where similar trends were observed [7-8].

Reflection

Several classifier methods/feature selection combinations were evaluated for predictive performance and stability in GBM survival outcome prediction using AUC and RSD. Heatmaps show that the feature selection method and number of features used did not contribute to predictive performance as significantly as the classifier used; ANOVA demonstrates classifier identiy as accounting for 75.5 % of the total variance in AUC scores. The best classifiers for predicting GBM survival outcome were Random Forest, Gradient Boosting, and Decision Tree, while the worst classifiers were Gaussian Naïve Bayes and Quadratic Discriminant Analysis. Upon comparison with the results of Parmar et al., classifier performance is suggested to be highly dependent on the tumor type and imaging modality used in the radiomic analysis. This is an interesting result that highlights the need for independent measurements on classifier performance for different types of tumors and imaging modalities.

Improvement

In order to more accurately calculate the representative AUC values of classifier/feature selection combinations, the number of iterations of train test split with different random states should be increased from 10 to 100. While 10 iterations have a very low probability of results being due to chance, 100 iterations would virtually ensure results are genuine. It should be noted that 10 iterations takes approximately 1 hour due to hyperparameter tuning, so 100 iterations will lead to very long computation times.

Additionally, hyperparameters can be further tuned to improve classifier performance, though it should be noted that for some classifiers such as Random Forest, parameter tuning may lead to significant computational slow downs. Moreover, additional classifiers can be studied to find if any have higher predictive performance for this set of data than Random Forest and Boosting.

Finally, more feature selection methods, or alternatives to feature selection, such as dimensionality reduction, could be implemented to see how it compares with our results. It is apparent that only a few features contribute significantly to the variability of the data, so principle component analysis could be an attractive alternative to feature selection. These methods should be further considered for future studies.

References

1. Gillies, R. J.; Kinahan, P. E.; Hricak, H., Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016,** *278* (2), 563-577.
2. Aerts, H. J. W. L.; Velazquez, E. R.; Leijenaar, R. T. H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; Hoebers, F.; Rietbergen, M. M.; Leemans, C. R.; Dekker, A.; Quackenbush, J.; Gillies, R. J.; Lambin, P., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **2014,** *5*, 4006.
3. Network, T. C. G. A. R., Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine* **2015,** *372* (26), 2481-2498.
4. Kotrotsou, A.; Zinn, P. O.; Colen, R. R., Radiomics in Brain Tumors: An Emerging Technique for Characterization of Tumor Environment. *Magnetic Resonance Imaging Clinics of North America* **2016,** *24* (4), 719-729.
5. Narang, S.; Lehrer, M.; Yang, D.; Lee, J.; Rao, A., Radiomics in glioblastoma: current status, challenges and potential opportunities. *Translational Cancer Research* **2016,** *5* (4), 383-397.
6. Chaddad, A.; Zinn, P. O.; Colen, R. R. In *Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM*, 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 16-19 April 2015; 2015; pp 84-87.
7. Parmar, C.; Grossmann, P.; Bussink, J.; Lambin, P.; Aerts, H. J. W. L., Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports* **2015,** *5*, 13087.
8. Parmar, C.; Grossmann, P.; Rietveld, D.; Rietbergen, M. M.; Lambin, P.; Aerts, H. J. W. L., Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Frontiers in Oncology* **2015,** *5*, 272.
9. Wang, J.; Wu, C.-J.; Bao, M.-L.; Zhang, J.; Wang, X.-N.; Zhang, Y.-D., Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. *European Radiology* **2017**, 1-9.
10. Ingrisch, M.; Schneider, M. J.; Nörenberg, D.; Negrao de Figueiredo, G.; Maier-Hein, K.; Suchorska, B.; Schüller, U.; Albert, N.; Brückmann, H.; Reiser, M.; Tonn, J.-C.; Ertl-Wagner, B., Radiomic Analysis Reveals Prognostic Information in T1-Weighted Baseline Magnetic Resonance Imaging in Patients With Glioblastoma. *Investigative Radiology* **2017,** *52* (6), 360-366.
11. Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; Lanczi, L.; Gerstner, E.; Weber, M. A.; Arbel, T.; Avants, B. B.; Ayache, N.; Buendia, P.; Collins, D. L.; Cordier, N.; Corso, J. J.; Criminisi, A.; Das, T.; Delingette, H.; Ç, D.; Durst, C. R.; Dojat, M.; Doyle, S.; Festa, J.; Forbes, F.; Geremia, E.; Glocker, B.; Golland, P.; Guo, X.; Hamamci, A.; Iftekharuddin, K. M.; Jena, R.; John, N. M.; Konukoglu, E.; Lashkari, D.; Mariz, J. A.; Meier, R.; Pereira, S.; Precup, D.; Price, S. J.; Raviv, T. R.; Reza, S. M. S.; Ryan, M.; Sarikaya, D.; Schwartz, L.; Shin, H. C.; Shotton, J.; Silva, C. A.; Sousa, N.; Subbanna, N. K.; Szekely, G.; Taylor, T. J.; Thomas, O. M.; Tustison, N. J.; Unal, G.; Vasseur, F.; Wintermark, M.; Ye, D. H.; Zhao, L.; Zhao, B.; Zikic, D.; Prastawa, M.; Reyes, M.; Leemput, K. V., The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **2015,** *34* (10), 1993-2024.
12. Fawcett, T., An introduction to ROC analysis. *Pattern recognition letters* **2006,** *27* (8), 861-874.
13. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; Buatti, J.; Aylward, S.; Miller, J. V.; Pieper, S.; Kikinis, R., 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging* **2012,** *30* (9), 1323-1341.
14. Joost JM van Griethuysen, A. F., Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, Hugo JWL Aerts, Computational Radiomics System to Decode the Radiographic Phenotype. *Submitted* **2017**.
15. James, G.; Witten, D.; Hastie, T.; Tibshirani, R., *An introduction to statistical learning*. Springer: 2013; Vol. 6.