Brain tumor classification with interpretable, transferable lasso-based group-wise feature selection

Nicholas Nuechterlein, Erin Wilson, Xingfan Huang

Abstract

Short-term survivors of glioblastoma (GBM), the most aggressive brain tumor, die before it is possible to enroll them in experimental clinical trials. Thus, it is of critical importance to identify these patients when their disease is first discovered so they may receive experimental therapy immediately. In this project, we predict the presence of genomic aberrations thought to characterize short-term survivors in 46 GBM patients using magnetic resonance (MR) scans from the The Cancer Imaging Atlas (TCIA) database. Because GBM MR scans are extremely high-dimensional, heterogeneous, and scarce, we apply traditional transfer learning techniques to learn discriminative imaging features from more straight-forward, related classification tasks on the Brain Tumor Segmentation (BraTS) dataset, a larger MR dataset of brain tumor patients (N = 285). We extract over 30,000 features from the MR brain scan data using classical computer vision techniques before developing a method which combines lasso-based grouped feature selection with dimensionality reduction to settle on only the most essential features for each classification task in the larger BraTS dataset. We then use these features in small-scale machine learning classifiers on the smaller, original TCIA dataset to predict the presence of abnormal genomic features. We leverage the natural semantic grouping of the lasso-selected features to interpret which regions of the MR images have the most influence on machine learning classifiers, and we use SHapley Additive exPlanations (SHAP) to interpret individual predictions. We compare the results of linear models, support vector machines (SVMs), gradient boosted decision trees, and other non-linear models using classification accuracy. Our method outperforms traditional machine learning models on all classification tasks on the larger BraTS dataset and improves human performers on the genomic-aberration classification task by 40%.

1 Introduction

Glioblastoma (GBM) is the most common and most aggressive primary brain tumor [8]. Patients with this disease live on average only 12-15 months. However there exists a subgroup of extremely short-term survivors who usually live less than six months. These patients are critical to identify so they can be treated with experimental therapies. Recent work has shown that the genomes in the short-survivor tumors exhibit extreme changes in their DNA code. These changes are known as tumor copy-number alterations–duplication or deletions of long sections of DNA–and may be able to delineate the short-survivor subgroup [6]. Currently, dangerous, costly, and often unsuccessful brain surgery is necessary to acquire such genomic information for patients. Since magnetic resonance (MR) scans are non-invasive and comparatively inexpensive, we are interested in predicting these types of major genomic changes in the tumors from magnetic resonance (MR) scans in the place of surgery.

MR scans are a rich source of volumetric data composed of up to hundreds of image slices of a patient's brain from different angles and orientations. At every clinical visit, GBM patients receive

Preprint. Under review.

the four different standard kinds of MR scans–T1, T1ce, T2, and FLAIR–known as MR modalities (Figure 1). Each MR modality is useful for capturing different aspects of the brain tumors. For example, T1ce can show active tumor cells while the T2 and FLAIR images are sensitive to water content, which exposes areas of fluid accumulation in the brain. Unfortunately, these MR scans are often quite noisy due to variations in scanner manufacturers, scan protocols, and even simple patient body movements during the scan. However, these scans each reveal important features of the tumor's extent and composition and, when used together, provide non-redundant information that characterize tumor features and differences. For a single patient, we typically have $\sim$ 50,000,000 voxels of information.

We possess a dataset of 46 GBM patients from The Cancer Imaging Atlas (TCIA) with MR data. Additionally, we have corresponding genome information from The Cancer Genome Atlas (TCGA) which bin the patients into two genomic aberration classes (present and not present). However, visual tumor classification is not a trivial task, even for experts. When presented with the tumor scan data and asked to perform the genome aberration classification by eye, an expert neuroradiologist at the University of Washington Medical Center (UWMC) was unable to do better than chance (50%), even after reviewing the images and their labels in advance. This underscores the difficulty of GBM prediction tasks.

Here, we aim to use computer vision techniques on the tumor image data to predict genomic aberration class, which has been suggested to be a proxy for patient survival class (short vs long) [6]. Well-known image classification datasets, such as ImageNet, need millions of images to learn how to identify every day objects that trained humans are able to classify accurately. In our case, we have few data points all of which are difficult to parse by human experts and thus we must adapt our classification strategy accordingly (see Methods).

1.1 Previous Work

The origin of the genomic aberration classification task lies in recent genomic work by Cimino et al., in which they stratify GBM survival by the presence of major genomic changes (copy-number alterations) [6]. To the best of our knowledge, no one has attempted to predict the copy-number alteration groups delineated in their work from imaging data. Recently, Kong et al. demonstrated that unsupervised clustering methods can group MR imaging features into known subtypes of GBM that have different survival rates, but they explicitly concluded that they found no evidence suggesting specific image features are indicative of major genomic changes [11].

Also relevant to our work are recent attempts to predict GBM survival directly from MR brain scan data, which we attempt to do as part of our feature selection pipeline. In both the 2017 and 2018 BraTS GBM survival prediction challenges, 163 patients with MR data were provided with survival labels. Each competition team was evaluated on the accuracy of their predictions of short-term survivors (<10 months), mid-survivors (between 10 and 15 months), and long-term survivors (> 15 months). Shboul et al., the 2017 winners, used a Univariate Cox regression model to select forty features which they fed into a random forest classifier to obtain their final predictions, recording 58% accuracy [15]. The second place team, Jungo et al., used extensive cross-validation for feature selection and a support vector machine (SVM) with an RBF kernel for final predictions, obtaining 57% accuracy, [10]. Perhaps surprisingly, the 2018 winners, Feng et al., achieved 62% accuracy using only a linear model and nine different features. As pointed out in a recent competition review, the poor results on survival prediction point to the difficulty of dealing with large, heterogeneous data, and the absence of popular deep learning models from the scoreboard highlights the fact that traditional machine learning approaches are better suited for small training sets [4].

1.2 Data

Primary Dataset (TCIA): We download tumor MR image data from 46 GBM patients from the TCIA data portal [7, 14]. Each patient contains the four standard MR modalities (T1, T1ce, T2, and FLAIR) taken at diagnosis as well as genomic aberration labels downloaded from The Cancer Genome Atlas (TCGA). We used the FLIRT algorithm from the neuroimaging package FSL to process and align image slices from each patient's four modalities [9]. We then segmented the tumor images using an in-house CNN segmentation model as a first pass and then manually corrected the segmentation voxel-by-voxel to the satisfaction of an expert neuroradiologist at the University of

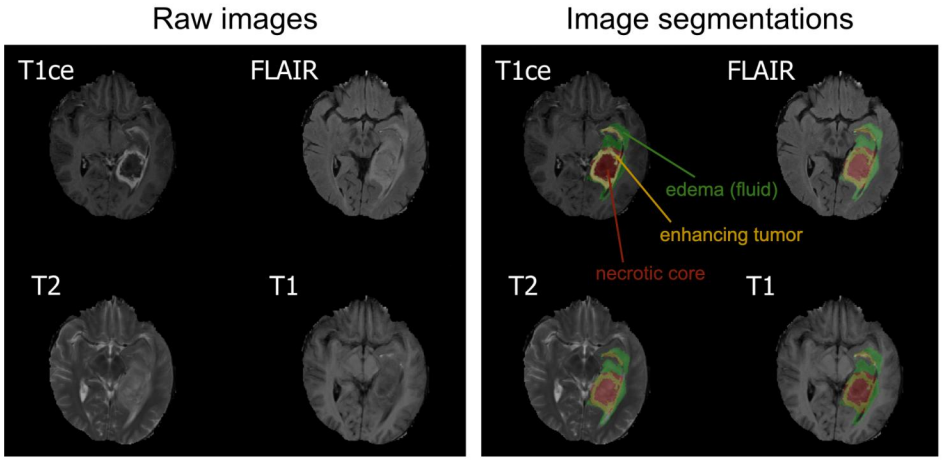

Figure 1: The T1ce sequence shows the contrast enhancing portion of the tumor (shown in yellow in the segmentation mask) and necrotic (dead) region (shown in red); FLAIR best indicates the edema region (green); T2 reinforces the FLAIR signal; the difference between T1 and T1ce indicates tumor presence.

Washington Medical Center. Each segmentation mask marks various tumor characteristics (e.g., enhancing tumor (yellow), peritumoral edema (green), and necrotic core (red)) (Figure 1).

Secondary Dataset (BraTS): We download data from the 2018 Multimodal Brain Tumor Segmentation Challenge (BraTS), which provides 285 MR tumor scans, each consisting of the 4 standard scan types (T1, T1ce, T2, and FLAIR) [1, 2, 3, 5, 13]. Of these scans, 210 are GBM tumors (high grade tumors) and 75 are low grade brain tumors. Tumor grade is an indication of the aggressiveness of a tumor; higher grade tumors have poorer survival times. Each patient's scans are annotated with the same voxel segmentation labels as in the primary TCIA dataset, processed and aligned to the same resolution, and are generally cleaner than the MR scans in the primary TCIA dataset. This dataset provides ground-truth segmentations manually drawn and approved by expert neuroradiologists. The BraTS dataset also has survival labels (in days) for 163 of the 210 GBM subjects as well as an indicator for whether or not the patient had surgery for 58 of the 210 GBM subjects.

2 Methods

2.1 Feature Generation and Selection

To compensate for the low sample count in our primary TCIA MR dataset, we first consider the larger, secondary BraTS MR dataset. We consider three separate classification tasks on the BraTS dataset: (1) directly predict patient **survival** (under a year or over a year), (2) predict the tumor **grade** (high grade or low grade), and (3) predict **resection** (whether a tumor is surgically removable or not). Our goal is to learn discriminative features for these BraTS tasks which we can then transfer to classifiers on the smaller TCIA dataset.

For the three BraTS classification tasks, we first extract semantically interpretable shape features, such as tumor volume, maximum diameter, and surface area. We also use the python package "pyradiomics" to extract more complex texture features such as histogram statistics, Gray-Level Co-Occurrence Matrix (GLCM) features, Grey-Level Run Length Matrix (GLRLM) features, Gray Level Size Zone Matrix (GLSZM) features, and other standard computer vision derived matrix features which we shall refer to as *groups of features*.

As a part of our pre-processing steps, we will add two additional tumor compartment masks: the whole tumor and the tumor core, which we create by merging all segmentation labels (necrotic + enhancing tumor + edema) and the enhancing and necrotic labels, respectively. We will perform several separate transforms on the original scans, including local binary patterns (LBP) and wavelet transforms. These groups of features will be calculated separately for each transformation, for each of

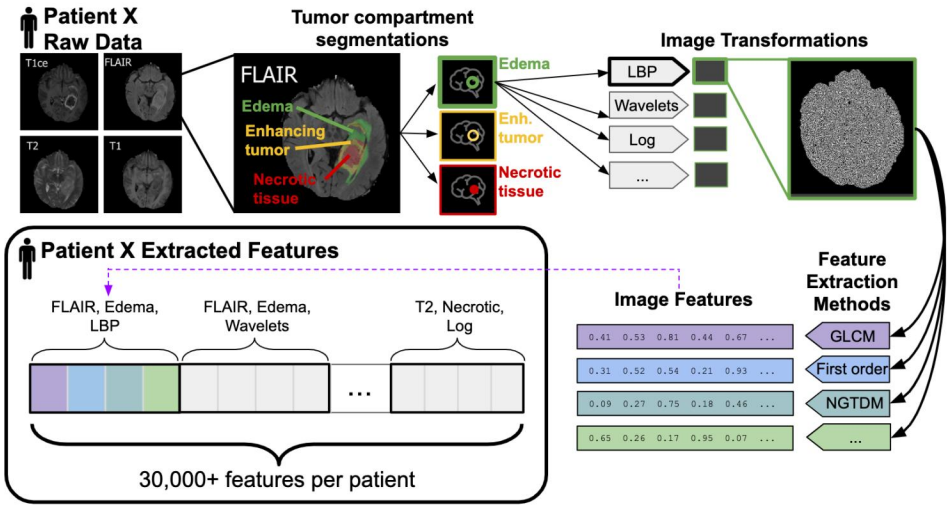

Figure 2: Overview of how 30,000 image features are generated for each brain tumor patient. Each of the 4 MR scan modalities is segmented into tumor compartments (e.g., edema). Each segmented image is transformed by various computer vision methods (e.g., LBP). Transformed images are then fed through various feature extraction methods which gather statistics about the image features (e.g., GLCM) and ultimately produce a small vector of real-valued numbers. We consider these numbers as features. Combinatorially, we end up with about 30,000 features in total.

the four MR modalities, and for each of the five tumor compartments (including the whole tumor and tumor core). For example, the GLCM features on the edema region of a tumor on a LBP transformed FLAIR modality will be considered a family of features. All of these groups will be calculated for both the smaller TCIA dataset and the larger BraTS dataset. Ultimately, we will end up with about 30,000 image features partitioned into around 2000 groups per patient, a substantial reduction from 50,000,000 voxel values obtained from raw scans, but still far too many for our limited number of patients (Figure 2).

In order to reduce our set of features further, we use both dimensionality reduction and feature selection methods. We settle on the following pipeline empirically, after performing extensive experiments as shown in Table 1.

Method 1: We reduce the number of feature groups from $\sim$2000 to 124 by merging all groups which share an MR modality, tumor compartment, and feature extraction method. We then apply the traditional lasso to each group to select individual features representative of the group, reducing the total number of features from $\sim$30,000 to 4,000. Further, we record the lasso cross-validated accuracy score for each group and keep only the groups with the top fifty accuracy scores. To benchmark our method, we use PCA to perform two-dimensional group-wise dimensionality reduction to reduce the number of features to 100. We pass them through seven simple ML classifiers as shown in Table 1. Method 1 is illustrated in Figure 4.

Next, we take the features selected by Method 1 on each BraTS classification tasks and use them to train classifiers for the genomic aberration classification task on the TCIA dataset. Specifically, we pass low-dimensional PCA projections of these features to lasso, RBF kernel support vector machines (SVM), multilayer perceptrons (MLP), XGBoost, decision trees (DT), random forests (RF), and logistic regression (LR) classifiers to obtain a final classification on the TCIA dataset. Our pipeline is shown in Figure 3. We use accuracy as an evaluation metric because the survival and copy number prediction tasks are well-balanced and false positives and false negatives are equally bad.

Method 2: As in Method 1, we apply lasso to each group of features and rank all groups by lasso's accuracy score. Unlike Method 1, instead of using PCA on only the lasso-selected features in each group, we apply PCA to the entire group the selected features were selected from.

Method 3: Unlike Methods 1 and 2, we first use PCA to project each group to two dimensions and second use lasso to select a subset of the PCA reduced groups on which to train our classifiers.

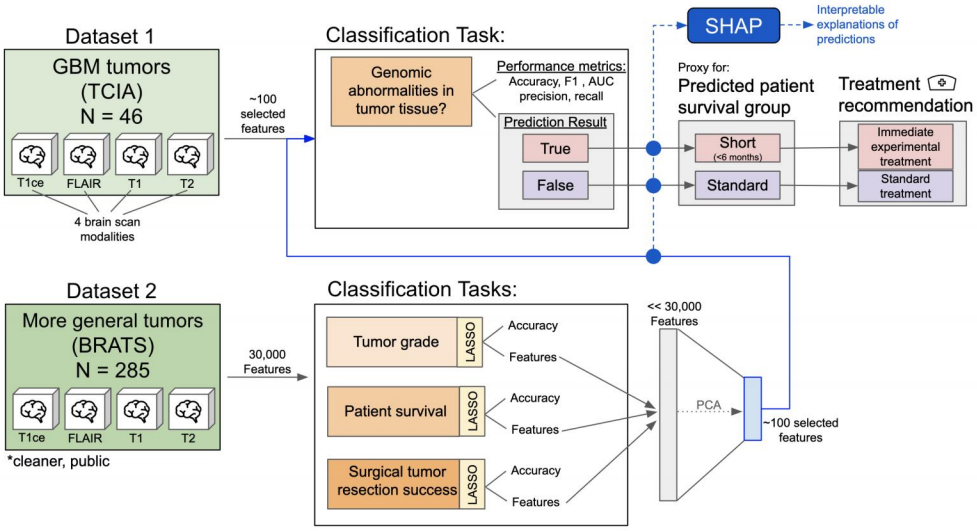

Figure 3: Computational Pipeline: Features are selected from the BraTS classification tasks, compressed using dimensionality reduction, and used as input for the TCIA genomic aberration classification task. SHAP is used to interpret the importance of these features to individual predictions.

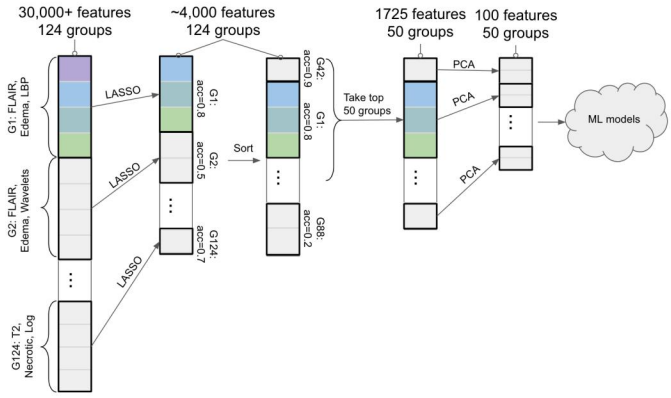

Figure 4: Method 1: lasso is applied group-wise to select 4000 features in 124 groups; groups are sorted by their lasso accuracy score; the top 50 groups are reduced to 2 dimensions and fed to the machine learning classifiers.

Method 4: In this method, we consider a group to be selected if any feature in that group is selected when lasso is applied to the entire dataset. After using lasso to select groups, we performed group-wise PCA on these selected groups and feed the PCA vectors into our machine learning classifiers. Of note, Method 4 selected 49 groups, 49 groups, and 50 groups from the 3 BraTS classifications tasks, respectively, which we used as justification for choosing 50 as the threshold for selecting accuracy-ranked groups in Method 1.

Method All Features: For comparison, we feed the entire dataset into the machine learning classifiers without performing any prior feature selection or dimensionality reduction.

We experimented with other forms of dimensionality reduction, such as UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction), t-SNE, and MDS, but the results were not significantly different. We chose PCA because linear methods are the most interpretable, and possibly more stable when transferring to different tasks on different datasets. While these tasks seems well suited for group lasso, we found in practice that group lasso was too computationally expensive to run when the number of groups exceeded 30 and decided to develop methods to use the standard lasso to preform group selection, such as Method 1 and Method 4.

2.2 Modeling and Evaluation

With each of the 5 feature selection methods described above and for each of the 3 BraTS and 1 TCIA classification tasks, we compared a variety of standard classification models from python's sci-kit learn library. In particular, we evaluated Lasso, Support Vector Machines (SVM), Multi-layer Perceptron (MLP), XGBoost, Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR).

While we report accuracy as our primary performance metric in Tables 1 and 2, we also compute the F1 score, AUC, precision, and recall to account for the class imbalance in the BraTS grade and resection tasks, but we observed no significant difference from accuracy in the order or magnitude of the other metrics.

We use 10-fold cross validation for all metrics. All of the machine learning classifiers use default hyperparameters expect for the SVM which we preform a grid search to find the best hyperparameters because it under preformed in early experiments. For feature selection, we use a python implementation of the R package glmnet.

2.3 Interpretation with SHAP

In addition to careful feature selection, we explore other methods to add interpretability to our models. As part of the previously described feature selection methods, we used lasso to impose sparsity and retain only a subset of features that are related to our classification tasks. Lasso provides some interpretability as the retained set of features should be the ones that collectively distinguish the two classes in classification. However, lasso is unable to say how and how much each feature contributes to the classification result in each sample. Therefore, we use SHapley Additive exPlanations (SHAP), an approach that can help explain the output of any machine learning model by calculating the contribution of each feature to each prediction at test time [12].

For the three BraTS classification tasks, SHAP is computationally prohibitive to run on our most accurate model, SVM, with the model agnostic explainer; therefore we run SHAP with one of the more accurate mdoels, XGBoost, using the SHAP tree explainer with all features. We then compare SHAP values with lasso feature selection to confirm that SHAP values are in line with our lasso results, which is the case. For the TCIA classification task, we run SHAP on multiple machine learning models with features selected by Method 1 and report the groups of features with top SHAP values. We expect these groups of features to be important for distinguishing the patients with and without copy number alterations.

3 Results

3.1 Feature Selection

We first analyze the relative proportion of image features selected from each MR modality and tumor compartment. Figure 5 visualizes the distribution of features selected by Method 1 from the three BraTS classification tasks over each MR modality (Fig 5A), tumor compartment (Fig 5B), and combination of tumor compartment and MR modality (Fig 5C). We observe that a disproportionate number of features involve the T1ce modality in the BraTS tumor grade classification task. This is in line with our knowledge that a bright ring around the tumor is present in nearly all high-grade tumors (as visible in the T1ce image in Figure 2), but rarely in low-grade tumors. We also see that emphasis is placed on all T1ce features in Fig 5C, especially in those compartments which include the enhancing ring (T1ce enh, T1ce tumor core, T1ce whole tumor). Intriguingly, it seems that the necrotic (nec) areas of the tumor are also important to predicting tumor grade, which may be a sign that aggressive tumors cut off the oxygen to the center of the tumor, causing cell death.

We appreciate that these plots affirm prior knowledge, but we are most interested in developing novel conjectures from new observations. For example, it is unclear why the T2 modality is so influential to the resection and survival tasks when neuroradiologists and neuro-oncologists usually consult the FLAIR modality when evaluating patients. Also interesting is the apparent irrelevance of the enhancing compartment to tumor resection prediction. We hypothesize that neurosurgeons are more concerned with the outer margins of the tumor, encoded in the edema and whole tumor compartment, instead of the inner areas which are easier to resect. However, the fact that the necrotic region of the

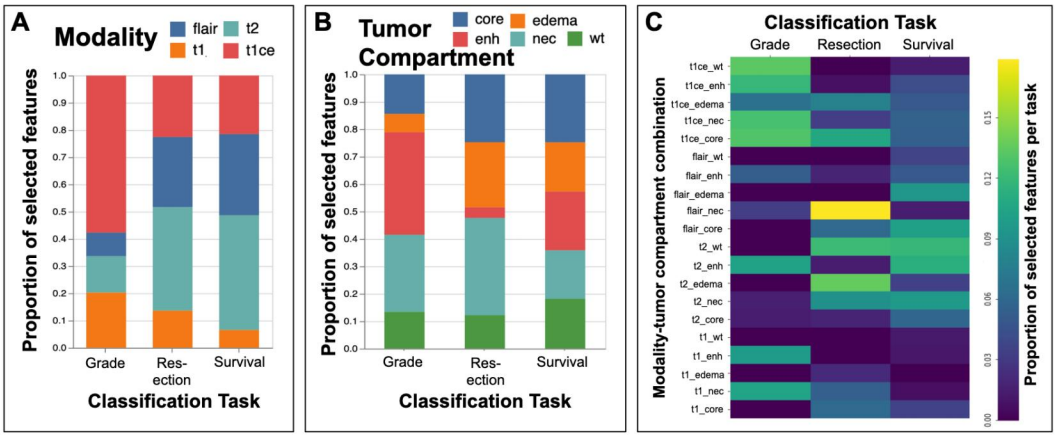

Figure 5: Relative proportion of selected features originating from interpretable regions of MR scans. A) Proportion of chosen features from each MR Scan modality (FLAIR, T1ce, T1, T2) for their relevance in each BraTS classification task (Tumor Grade, Tumor Resection, Patient Survival). B) Proportion of chosen features from each tumor compartment image segmentation for their relevance in prediction BraTS classification tasks. C) Heatmap depicting proportion of chosen features selected from combined modality and tumor compartment images.

FLAIR modality is the most influential combination of MR modality and tumor compartment directly contradicts this conjecture and is a significant surprise.

Table 1: BraTS Experimental Results (Accuracy)

Task	Method	lasso	SVM	MLP	XGBoost	DT	RF	LR
Survival	1*	**0.700**	0.700	0.675	**0.681**	**0.626**	**0.650**	0.700
	2	0.650	0.589	0.540	0.577	0.571	0.583	0.589
	3	0.681	**0.736**	**0.687**	0.656	0.564	0.607	**0.736**
	4	0.663	0.613	0.601	0.620	0.577	0.564	0.613
	All Features	0.607	0.632	0.650	0.620	0.583	0.546	0.632
Resection	1*	**0.843**	**0.855**	**0.795**	**0.759**	**0.711**	**0.723**	**0.855**
	2	-	-	-	-	-	-	-
	3	-	-	-	-	-	-	-
	4	-	-	-	-	-	-	-
	All Features	0.639	0.711	0.687	0.711	0.687	0.639	0.711
Tumor Grade	1*	**0.933**	**0.926**	**0.930**	**0.944**	**0.891**	**0.930**	**0.926**
	2	0.923	0.923	0.923	0.930	0.853	**0.930**	0.921
	3	0.930	0.919	0.895	0.919	0.884	0.926	0.919
	4	0.910	0.919	0.905	0.940	**0.891**	0.905	0.919
	All Features	0.919	0.905	0.909	0.930	0.877	0.902	0.905

Table 2: TCIA Experimental Results (Accuracy)

Task	Training Features	lasso	SVM	MLP	XGBoost	DT	RF	LR
Copy-Number	Survival Features	0.696	0.783	0.739	0.739	0.652	0.739	0.783
	Resection Features	0.717	**0.804**	**0.783**	**0.804**	**0.761**	**0.761**	**0.804**
	Tumor Grade Features	**0.913**	**0.804**	0.587	**0.804**	**0.761**	0.696	**0.804**
	All Features	0.761	0.761	0.587	0.761	**0.761**	0.739	0.761

3.2 BraTS accuracy

We next examine model performance on the BraTS classification tasks for each of the feature selection methods outlined in the Methods section. We find that Method 1 performs the best, notably because it was the only method able to select features on the BraTS resection task and because it outperformed all other methods on the BraTS tumor grade classification task. We are somewhat surprised by the size of the margin between the performance of Method 1 and Method 2 on this task—the large disparity in performance indicates a significant amount of information is lost when using PCA to project an entire feature group to two dimensions vs. projecting only the lasso-selected features of a feature group to two dimensions.

Method 3 tests whether the order of applying lasso and applying PCA makes a significant difference in performance. Our results also show that Method 3, which applies PCA first, is not clearly better than Model 2, which applies lasso first. This is affirming because Method 1 also applies lasso first.

Method 4 tests the difference between (1) performing group selection by thresholding the group-specific lasso accuracy scores (as in Method 1 and Method 2) and (2) performing group selection by considering a group selected if at least one feature in the group is selected when lasso is applied to the entire dataset. Since both Method 2 and Method 4 use PCA on the entirety of each group, they serve as a fair comparison. The results of this experiment are inconclusive as Method 4 tends to outperform Method 2 on the BraTS survival task but tends to be outperformed by Method 2 on the BraTS tumor grade task. However, Method 4 has the advantages of yielding sparser solutions.

Our results indicate that predicting the tumor grade is by far the easiest BraTS classification task, achieving the highest accuracies across all models (Table 1). This is in line with our expectations as neuroradiologists can reliably predict tumor grade.

The blank rows in BraTS tumor resection task in Table 1 show that Method 2, Method 3, and Method 4 were unable to select any features. This remains a mystery to us, especially since these methods are able to select features on the BraTS survival task which, ostensibly, is a more difficult task.

3.3 TCIA accuracy

For our primary TCIA genomic-aberration classification task, we found that using the best features transferred from BraTS classification tasks performed better than using all features. In particular, features from tumor grade and resection prediction achieved the highest classification accuracy across most ML models we applied (Table 2). We are able to outperform an expert neuroradiologist's 50% classification accuracy by over 40% by training lasso on the features Method 1 extracted from the BraTS tumor grade classification task. We are also able to exceed the performance of vanilla machine learning algorithms on the TCIA genomic aberration classification task without custom-selected BraTS features (Table 2, All Features) by 15%.

3.4 Interpretability

In Figure 6 we show comparisons between SHAP results and lasso feature selection on the three BraTS classification tasks. We can immediately observe from plot 5A that SHAP values agree with lasso feature selection proportions, especially on the grade classification task, indicating that SHAP values are also consistent with our prior knowledge. When considering combinations of tumor compartments and MR modalities, SHAP values are high for the enhancing (enh) region of the T1ce modality, again matching prior knowledge. Another note is that SHAP did not assign a high importance value to the necrotic region of the FLAIR modality, indicating that the algorithm can help us refine our previous hypotheses from lasso. SHAP values and lasso selections are less consistent on the harder classification tasks (survival and resection prediction), which shows that we should use caution when interpreting results before multiple models and algorithms have been consulted for consistency. In general, we are confident that SHAP values are viable for evaluating feature importance.

For the TCIA classification task, we ran SHAP with lasso, SVM, and XGBoost using the features selected by Method 1 on the survival, resection, and tumor grade classification tasks to gain insight on the contributions of the input groups of features. We selected a diverse set of machine learning models to run SHAP on, aiming to probe the consistency of SHAP outputs across models. The results are shown in Figure 7. Overall we observe that the groups of features with the top SHAP scores

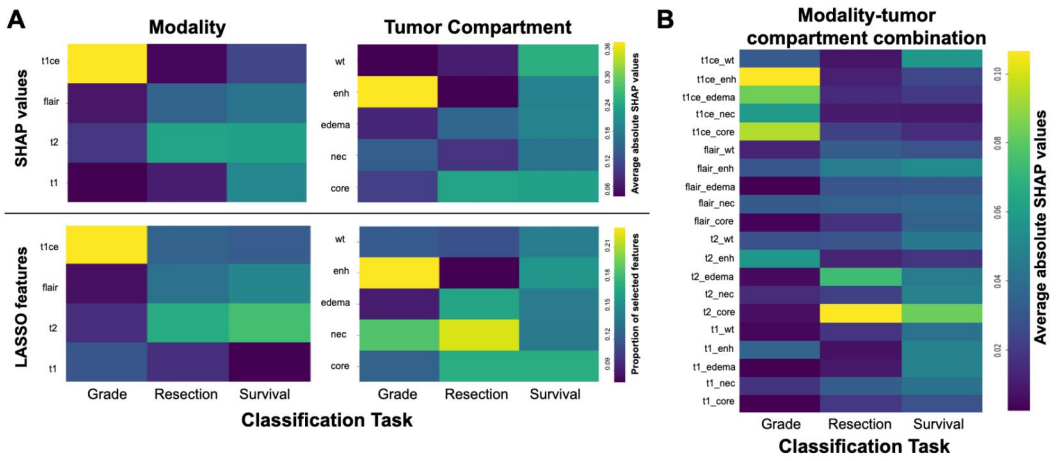

Figure 6: Comparing SHAP and lasso outputs on BraTS classification tasks. A) Heatmaps depicting average absolute SHAP values (top) and proportion of chosen features from lasso (bottom) from modality (left) and tumor compartment(right) images. B) Heatmap depicting average absolute SHAP values from combined modality and tumor compartment images.

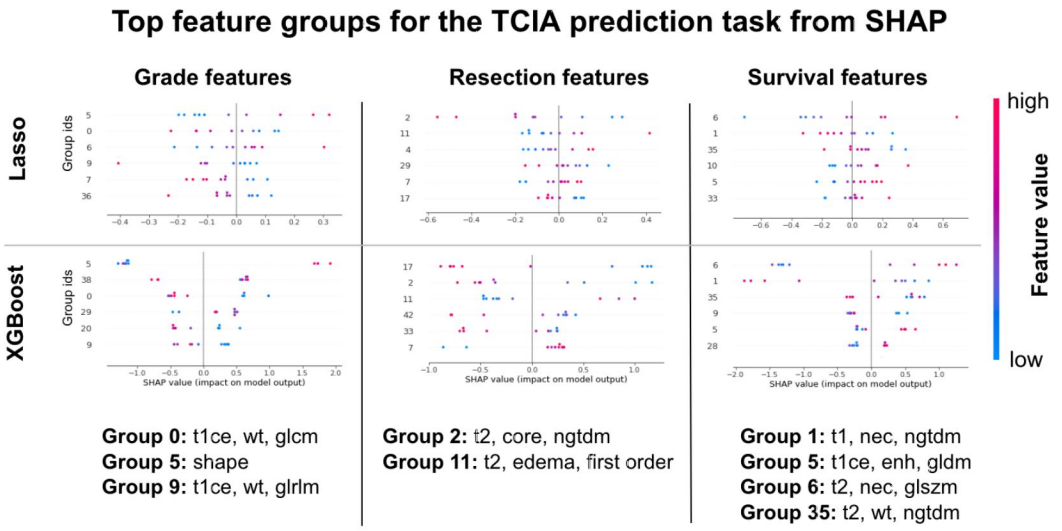

Figure 7: SHAP results on the TCIA prediction task using various machine learning models (rows) and input feature sets (columns). For each subplot, the x-axis is the SHAP value, the y-axis is the feature group id, and each point is a patient sample from the test set colored by the value of the feature. Points further from the mid-line have higher absolute SHAP values, indicating larger contribution to the classification. Top groups of features which received high average SHAP values across models are listed for each task.

are mostly consistent (across columns), but they do change when the same models are trained on sets of features Method 1 selected from different classification classes (across rows). When focusing specifically on the TCIA SHAP results from XGBoost with the features Method 1 selected from the tumor grade classification task, Group 5 has the highest average SHAP values beacause all SHAP values (points) fall far from the mid-line. Points with low group 5 feature values (blue) fall on the left hand side of the mid-line, meaning samples with low group 5 feature values were assigned negative SHAP values and encourage the XGBoost model to predict the 0 class (no copy number alteration).

4 Discussion

We hypothesized that the classification of genomic aberrations on the TCIA dataset was intrinsically difficult, as evidenced by an expert neuroradiologist's failure to achieve better than chance. The primary contribution, and most surprising result, is that we managed to develop an accurate genomic aberration classifier. Given the radical improvement of our results–over 40% improvement over a trained human–we must wonder (1) whether the task was as difficult as we assumed it was and (2) whether our results are overly optimistic or (3) that we have made a significant breakthrough.

We suspect a combination of (1) and (2) are at play though we hope further validation will prove (3) true, even if only in part. We used only one neuroradiologist to estimate human performance. Moreover, the neuroradiologist, though presented with the MR images in review advance to study, was unfamiliar with any links between genomic aberrations and MR images. This did not seem problematic because no known link had been studied, but it possible that these aberrations present themselves in subtle but consistent ways on MR images. This is further evidenced by the 76% accuracy our basic machine learning classifiers achieved, already a 26% improvement over our expert.

On the other hand, it is possible our cross-validation techniques are over-optimistic. We did not use a holdout test set because we felt we already had too few samples. Results on a high-dimensional dataset of only 46 subjects should be taken cautiously. It is possible that there exists artifacts in these MR images to which the models are overfitting. Fortunately, a set of fifty patients with genomic aberration labels exists in the Fred Hutchinson Cancer Research Center's medical records on which we hope to validate our method as a follow up analysis.

Also related to potential over-optimism is our choice of accuracy as an evaluation metric. Accuracy is suited for well-balanced tasks to which false positives and false negatives are given the same importance, such the BraTS survival task and the TCIA genomic-aberration task. However, both the BraTS tumor grade classification task and BraTS tumor resection classification task are unbalanced, and reporting the F1 score seems more intuitive than accuracy. We choice accuracy over the F1 score because all feature selection in Method 1 is done by ranking feature groups by the accuracy of the lasso model used to select their features because the python implementation of the glmnet R package does not support the F1 score. Although we noticed no significant differences when we computed the F1 scores of the final classifiers, we will experiment with thresholding the top 50 groups in Method 1 with the F1 score in the future.

In machine learning we often face a tradeoff between model complexity/accuracy and interpretability. Interpretability is especially important for our task because our end goal is not only to have an accurate model, but also to understand how the model makes a certain classification and leverage that knowledge to help doctors make clinical decisions for incoming patients. In this project we attempted to add interpretability to our models using SHAP, a framework that explains machine learning model outputs by quantifying the importance of input values with SHAP values. We curated a list of groups of features that were assigned high SHAP values, indicating that these groups are potential markers for copy number alterations. We need to use external knowledge of the feature groups to validate our results in the future.

5 Team Member Contribution

Equal Contribution.

Erin Wilson: Extensive work on figures and report writing. Responsible for poster creation. Gave important methodology feedback.

Xingfan Huang: Responsible for the entire SHAP section of our project. Gave valuable feedback to other parts of the project.

Nicholas Nuechterlein: Project lead. Development and coding of methods, data acquisition and processing, project report drafting.

References

[1] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert

segmentation labels and radiomic features. *Sci Data*, 4:170117, 09 2017.

[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. *The Cancer Imaging Archive*, 2017.

[3] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. *The Cancer Imaging Archive*, 2017.

[4] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[5] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, B Menze, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[6] Patrick J Cimino, Lisa McFerrin, Hans-Georg Wirsching, Sonali Arora, Hamid Bolouri, Raul Rabadan, Michael Weller, and Eric C Holland. Copy number profiling across glioblastoma populations has implications for clinical trial design. *Neuro-oncology*, 20(10):1368–1373, 2018.

[7] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, Dec 2013.

[8] J. R. Fink, M. Muzi, M. Peck, and K. A. Krohn. Multimodality Brain Tumor Imaging: MR Imaging, PET, and PET/MR Imaging. *J. Nucl. Med.*, 56(10):1554–1561, Oct 2015.

[9] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

[10] Alain Jungo, Richard McKinley, Raphael Meier, Urspeter Knecht, Luis Vera, Julián Pérez-Beteta, David Molina-García, Víctor M Pérez-García, Roland Wiest, and Mauricio Reyes. Towards uncertainty-assisted brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 474–485. Springer, 2017.

[11] Doo-Sik Kong, Junhyung Kim, Gyuha Ryu, Hye-Jin You, Joon Kyung Sung, Yong Hee Han, Hye-Mi Shin, In-Hee Lee, Sung-Tae Kim, Chul-Kee Park, et al. Quantitative radiomic profiling of glioblastoma represents transcriptomic expression. *Oncotarget*, 9(5):6336, 2018.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[13] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015.

[14] L Scarpace, T Mikkelsen, S Cha, S Rao, S Tekchandani, S Gutman, and D Pierce. Radiology data from the cancer genome atlas glioblastoma multiforme [tcga-gbm] collection. *The Cancer Imaging Archive*, 11(4), 2016.

[15] Zeina A Shboul, Lasitha Vidyaratne, Mahbubul Alam, and Khan M Iftekharuddin. Glioblastoma and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 358–368. Springer, 2017.