# Radiomic Prediction of Tumor Grade and Overall Survival from the BraTS Glioma Dataset: An Exploratory Analysis of Dimensionality Reduction Techniques and Machine Learning Classifiers

Kareem Wahid
Colen Lab, MD Anderson Cancer Center

## Abstract

Radiomics obtains quantitative features from medical images that reveal novel information about tumor phenotype. However, before radiomics can be applied in a clinical setting, highly accurate and reliable machine learning models must first be explored with open-source tools and data. The BraTS Challenge provides a standardized multi-modal magnetic resonance imaging dataset complete with expert segmentations for lower-grade and higher-grade glioma cases; overall survival data is provided for a subset of higher-grade glioma cases. Herein, we investigate two machine learning classification tasks using radiomic features extracted from this dataset: i. prediction of tumor grade (higher-grade vs. lower-grade gliomas), and ii. prediction of overall survival in higher-grade gliomas (< 12 months vs. > 12 months). These tasks are attempted using machine learning models constructed with various classifier methods and dimensionality reduction techniques. Models are assessed in terms of their predictive performance and stability using a bootstrap approach. We show that the choice of classifier and dimensionality reduction technique plays a significant role in model performance and stability for both these tasks. As expected, metrics are much higher for grade classification. It is also shown that certain models work well for either task. In addition, variability analysis indicates the choice of classification method is the most dominant source of performance variation for both tasks. Through the use of publically available data and open-source radiomics tools/machine learning algorithms, we hope to provide a benchmark of comparison for future radiomics studies on brain tumors.

## Introduction

Radiomics is a budding new field of medical informatics that seeks to extract mathematically defined quantitative features such as shape, intensity, 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 the classification and prognosis of disease. Normally in the radiomics pipeline, regions of interest are manually or automatically segmented from medical images, whereupon radiomic features are extracted. Subsequently, machine learning models are constructed using these features in order to predict clinical outcomes.

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 grading and prognosis difficult [3,4]. Radiomic analysis of glioma medical imaging can provide additional information about a patient's grade, prognosis, and likely survival outcomes [5-7].

Though significant research has been conducted on the application of machine learning algorithms to radiomic features for grading and prognostic prediction [8-17], there is still much that is unknown about which models are best due to lack of standardization in the field. Moreover, many studies utilize proprietary or in-house software in their radiomic feature extraction/analysis pipeline, severely limiting the community from making advances. Coupled with medical images that are protected under patient confidentiality laws, this makes results extremely difficult, if not impossible, to reproduce. Therefore, it is crucial to utilize publically available datasets and open-source tools in order to expand the radiomics field.

In this study, supervised machine learning models were trained on radiomic features extracted from publically available glioma magnetic resonance images (MRI) obtained from the 2017 BraTS Challenge to predict tumor grade and overall survival outcomes. An open-source radiomics toolbox, Pyradiomics, was used for radiomic feature extraction, coupled with the popular open-source machine learning python library, scikit-learn, for model building and analysis.

## Methods

*Dataset:* We utilized the 2017 BraTS Challenge Training Dataset [18], i.e. multimodal (T1, T1ce, T2, FLAIR) pre-operative MRI scans of lower-grade glioma (LGG) and higher-grade glioma aka glioblastoma (GBM), each with corresponding manual segmentations. Segmentation annotations comprise of the following tumor phenotypes aka subtypes: Necrotic/non-enhancing tumor (NCR), peritumoral edema (ED), and Gd-enhancing tumor (ET). The data was pre-processed by registering to the same anatomical template, interpolated to 1 mm$^3$ resolution and skull stripped. In addition, the overall survival data was available for a subset of the GBM scans. More information on the dataset can be found in Supplementary A. Hereafter all samples used for the grade prediction classification task will be referred to as the Grade Dataset and all samples used for the overall survival prediction classification task will be referred to as the Survival Dataset. For the Grade Dataset, GBM was considered the negative class (n = 210) while LGG was considered the positive class (n = 75). The Survival Dataset was stratified into binary classes based on median survival rates for GBM [19]; cases that died before 12 months from diagnosis were considered the negative class (n = 81), while cases that died after 12 months were considered the positive class (n = 82).

*Radiomic Features:* The newly released open-source radiomics toolbox, Pyradiomics v1.2.0, was utilized for feature extraction [20]. 16 shape, 19 first-order statistics, 27 gray level co-occurrence matrix, 16 gray level size zone matrix, and 16 gray level run length matrix features were extracted from each phenotype region of interest with the following image:mask combinations: T1ce:NCR, FLAIR:ED, T1ce:ET. In addition, a coiflet wavelet transform filter (8 decompositions) was applied to each image; for each phenotype, intensity-based features were calculated for each of the decompositions. This combination of shape, first-order, texture, and wavelet features led to 718 features extracted for each phenotype, i.e. 2154 features in total for each sample. More details on the radiomic features are available in Supplementary B. Before extracting features, voxel intensity values were normalized using the Pyradiomics normalization function (Z-score whole brain normalization), discretized with a binwidth of 0.1, and constrained to an intensity value range of 3 standard deviations from the mean (outliers removed). For the Grade Dataset, several LGG samples did not contain ET segmentations. Therefore, in order to keep the number of features equal regardless of tumor grade, these samples were discarded from the analysis. In addition, several image:mask combinations suffered from geometry mismatches and were likewise discarded. The removal of these samples from the Grade Dataset led to a total of 44 LGG samples and 191 GBM samples remaining for the analysis. Similarly, after removal of inappropriate samples, the Survival Dataset was left with a total of 73 GBM samples with survival < 12 months and 77 GBM samples with survival > 12 months.

*Feature Reduction Methods:* Radiomics leads to the creation of several informative features for use in predictive modeling. However, many of these features can be highly redundant or uninformative. Consequently, machine learning models built with radiomic features often suffer from the curse of dimensionality, i.e. models with more features often require exponentially more data to make accurate predictions. Therefore, feature reduction is often a helpful component of radiomic analysis in order to increase performance and decrease computational costs. There are many methods to reduce the feature space for machine learning models. Popular categories of feature selection methods include filter, wrapper, and embedded methods. Moreover, filter methods provide an advantage over wrapper and embedded methods by being classifier independent and computationally efficient [21]. Surprisingly, past studies have found univariate filter methods that ignore interactions between variables can be just as effective as multivariate methods that consider these interactions [8,9]. A possible alternative to feature selection is dimensionality reduction, whereby a high-dimensional space is transformed into a space of fewer dimensions via linear or nonlinear mappings, often taking into account complex interactions between variables [22,23]. It was recently suggested that unsupervised dimensionality reduction techniques could have higher predictive performance then filter methods in radiomic studies [24]. In this study, 4 unsupervised dimensionality reduction methods were utilized in building machine learning models: principal component analysis (PCA), kernel PCA (KPCA), independent component analysis (ICA), and factor analysis (FA). We chose these methods due to their simplicity, computational efficiency, and easily available implementation. More information on the dimensionality reduction methods can be found in Supplementary C. In addition, these methods were compared with a univariate filter technique, ANOVA F-score with the top 30 features selected (FILT), and maximum 2D diameter features from each phenotype (DIAM). DIAM was chosen to investigate how our radiomic methods would compare against a commonly utilized prognostic radiological metric [4].

*Classifiers:* Classification is considered a supervised method in machine learning where models approximate a target function from underlying labeled data [22]. Training data consists of examples represented by a set of input features (radiomic features) and an output value (tumor grade or overall survival class). Once a machine learning model is built from labeled data using a classifier and feature reduction technique, the class of an unlabeled sample may then be predicted. In this study, 9 classifier methods from different classifier families were selected for comparison: Decision Trees (DT), Random Forrest (RF), Bagging (BAG), Boosting (BST) Gaussian Naïve Bayes (NB), Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), Logistic Regression (LR), and K-nearest Neighbors (KNN). These methods were chosen for their widespread use in radiomic studies and simple implementation. More information on the classifier methods can be found in Supplementary D. To improve manuscript readability, all acronyms related to machine learning model building are displayed in Table 1. While classifier hyperparameter tuning, such as through cross validation, can often increase model performance

[25], in this study classifiers were used with their default hyperparameter settings to maintain simplicity and reduce computational cost. Machine learning models were built from combinations of feature reduction methods and classifier methods.

**Table 1.** Acronyms related to machine learning models.

| **Classifier Methods** | **Dimensionality Reduction Methods** | **Feature Selection Methods** |
|---|---|---|
| Decision Trees (DT) | Principal Component Analysis (PCA) | ANOVA F-score (FILT) |
| Random Forest (RF) | Kernel PCA (KPCA) | Max 2D Diameter (DIAM) |
| Bagging (BAG) | Independent Component Analysis (ICA) | - |
| Boosting (BST) | Factor Analysis (FA) | - |
| Naïve Bayes (NB) | - | - |
| Multi-Layer Perceptron (MLP) | - | - |
| Support Vector Machine (SVM) | - | - |
| Logistic Regression (LR) | - | - |
| K-Nearest Neighbor (KNN) | - | - |

## Analysis

All analysis was performed using Python 2.7 with scikit-learn v0.18.1 [26] on Mac OS Sierra v10.12.5.

*Model building and Evaluation:* For each dataset (Grade Dataset n = 245, Survival Dataset n = 150), data was randomly split into training and testing sets with a test size = 0.2, yielding training sets containing 196/120 samples and testing sets containing 49/30 samples respectively. To prevent class imbalances from affecting classifiers, synthetic minority upsampling (SMOTE) [27] was applied to the Grade training set; previous radiomics studies have shown SMOTE to be effective at improving classification predictive performance when classes are imbalanced [24]. SMOTE was not applied to the Survival Dataset since classes were already balanced. Z-score normalization, an often necessary pre-processing step in dimensionality reduction, was used to standardize features with respect to the training set.

In order to investigate and compare different dimensionality reduction and classification methods, we constructed a three-dimensional parameter grid for analysis. For each of the 4 dimensionality reduction methods, we incrementally selected the number of dimensions (e.g. principal components) ranging from 1 to 15 in steps of 2 (n = 1, 3, 5, …, 15). These subsets of dimensions were evaluated using each of the 9 machine learning classifiers and training data to build a predictive machine learning model. The model was then evaluated on the test set by calculating the area under the receiver operating curve (AUC) score. This was repeated 100 times for each combination with different random splits through a bootstrap approach. The mean of the AUC values ($\mu_{AUC}$) over all iterations was calculated to determine the final AUC value for a given model. By calculating the mean over 100 iterations we are able to ensure a more representative value for each model. Similarly, an empirical metric for stability, relative standard deviation (*RSD*) was previously defined as [8]:

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

where $\sigma_{AUC}$ and $\mu_{AUC}$ were the standard deviation and mean of the 100 AUC values respectively. It should be noted that higher stability corresponds to lower RSD values. A diagrammatic representation of the model building process is outlined in Fig. S1.

*Experimental Factors Affecting Prediction:* There are 3 main experimental factors in our study which can affect the radiomics based prediction: classifier method (RF, NB, DT, BAG, BST, SVM, LR, MLP, KNN), dimensionality reduction method (PCA, KPCA, ICA, FA), and number of dimensions selected (1,3,5,…,15). Multifactor analysis of variance (ANOVA) was utilized to quantify the variability in AUC scores contributed by these factors and their interactions for each classification task. In order to compare the variability contributed 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.

## Results

To investigate machine learning approaches for glioma grade prediction and GBM survival prediction, a total of 2154 features were extracted from the segmented tumor regions of the pre-treatment MRI scans from the BraTS Glioma Dataset. For the Grade Dataset, the output classes were LGG or HGG while for the Survival Dataset the output classes were < 12 months or > 12 months survival. For both classification tasks, feature reduction and classification training was done using the training set, whereas the testing set was used to assess performance and stability. An outline of the radiomic workflow for the glioma grade classification task is shown in Figure 1. A similar scheme was used for the overall survival classification task.

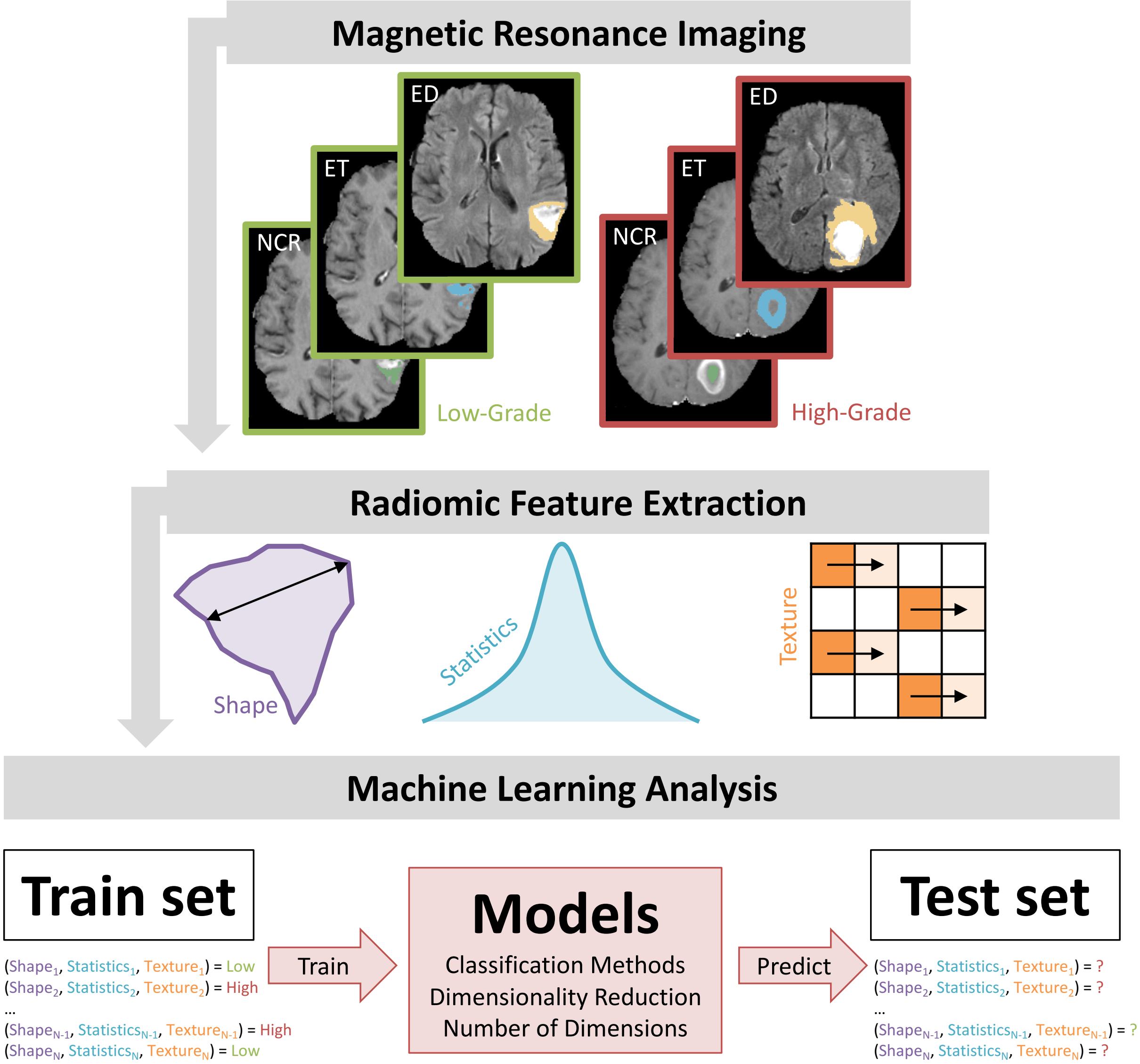
Magnetic Resonance Imaging
ED
ET
NCR
Low-Grade
ED
ET
NCR
High-Grade
Radiomic Feature Extraction
Shape
Statistics
Texture
Machine Learning Analysis
Train set
(Shape1, Statistics1, Texture1) = Low
(Shape2, Statistics2, Texture2) = High
...
(ShapeN-1, StatisticsN-1, TextureN-1) = High
(ShapeN, StatisticsN, TextureN) = Low
Train
Models
Classification Methods
Dimensionality Reduction
Number of Dimensions
Predict
Test set
(Shape1, Statistics1, Texture1) = ?
(Shape2, Statistics2, Texture2) = ?
...
(ShapeN-1, StatisticsN-1, TextureN-1) = ?
(ShapeN, StatisticsN, TextureN) = ?

**Figure 1.** Radiomics workflow for grade classification task. Survival classification task utilized a similar scheme.

*Predictive Performance of Models:* Predictive performance of machine learning models constructed from different dimensionality reduction and classification methods for both tasks was assessed using the AUC score; alternative scoring metrics for the grade classification task and the survival classification task are displayed in Tables S1 and S2 respectively. Figure 2 depicts the performance of dimensionality reduction (in rows) and classification methods (in columns) using 9 dimensions for both tasks; performance from models constructed using ANOVA F-score univariate filter method (FILT), and diameter features (DIAM) are also displayed for comparison. We repeated the above experiment by varying the number of dimensions; heatmaps corresponding to 1,3,5,7,9,11,13, and 15 dimensions for each dimensionality reduction method are displayed in Fig. S2 and S3.

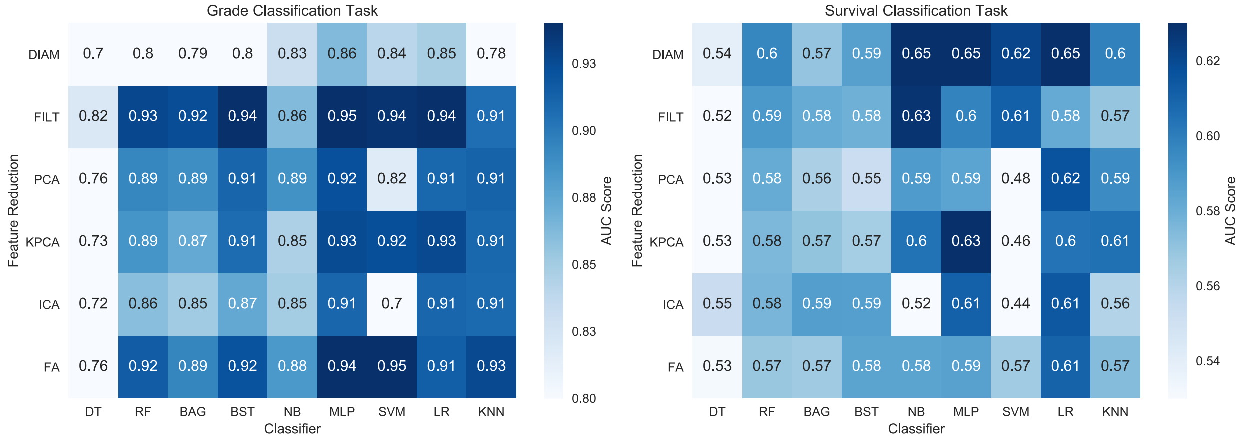

**Figure 2.** Heatmaps depicting predictive performance (AUC) of feature reduction (rows) and classification (columns) methods for grade classification task (left) and survival classification task (right).

For the grade classification task (Fig. 2, left), the best results among the four dimensionality reduction techniques is often achieved by FA, while the worst results are often achieved by ICA. Moreover, FA has comparable results to FILT, which often has the highest predictive performance. Additionally, using diameter features alone scores much lower than any dimensionality reduction techniques. In terms of classifiers, most classifier methods show similar results with the exception of DT which is noticeably lower.

For the survival classification task (Fig. 2, right), the best results among the four dimensionality reduction techniques is likewise also often achieved by FA, but otherwise performance results are more similar than in the grade classification task. Worthy of note, is that using diameter features alone often scores comparable or higher than any dimensionality reduction techniques. Again most classifier methods show similar results with the exception of decision trees and support vector machines (for PCA, KPCA, and ICA), which are noticeably lower. Additionally, AUC scores for the survival classification task are much lower (< 0.65) than for the grade classification task (> 0.80).

*Stability vs. Predictive Performance of Classifiers:* For each classification method, there are 4 AUC/RSD values corresponding to the different dimensionality reduction techniques (PCA, KPCA, ICA, FA). For each classification task, we used a median of all 4 AUC/RSD values as the representative AUC/RSD of a classifier. Scatterplots in Figure 3 assess the representative stability and predictive performance of classifiers for each classification task. For the grade classification task, it can be observed that MLP, LR, KNN, and BST should be preferred as their stability and predictive performance were higher than the corresponding median values across all classifiers. Similarly, for the survival classification task MLP, LR, and KNN should be preferred, with BST on the borderline of top performance and stability.

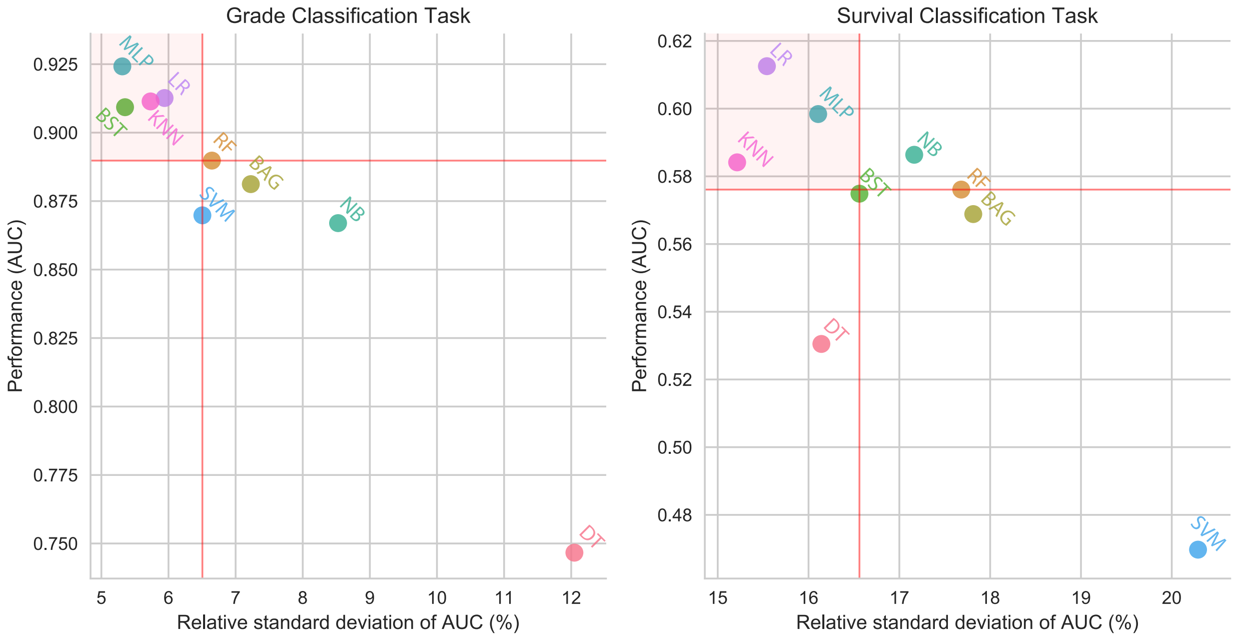

**Figure 3.** Scatterplots between representative stability and predictive performance of classification methods for grade classification task (left) and survival classification task (right). Classifier methods having RSD < median RSD and AUC > median AUC are considered as relatively reliable and accurate. Highly reliable and accurate methods are displayed in the red square region in the upper left corner.

*Experimental Factors:* To quantify the effect of the three experimental factors, i.e. classification methods, dimensionality reduction methods, and number of selected dimensions, multivariate ANOVA was performed on AUC scores (Fig. 4). We observed that all three experimental factors and their interactions are significant factors affecting the prediction performance for both classification tasks. Classification method was the most dominant source of variability as it explained 36 and 37 % of the total variance in AUC scores for grade and survival classification tasks respectively. The number of dimensions used was the second most dominant source of variability for both tasks as it explained 28 and 20 % of the total variance in AUC scores for grade and survival respectively. Dimensionality reduction method was the least dominant source of variability for both tasks as it explained 3 and 2 % of the total variance in AUC scores for grade and survival respectively. Interaction terms between the experimental factors followed similar trends. More information on ANOVA results for the grade classification task and survival task are shown in Tables S5 and S6 respectively.

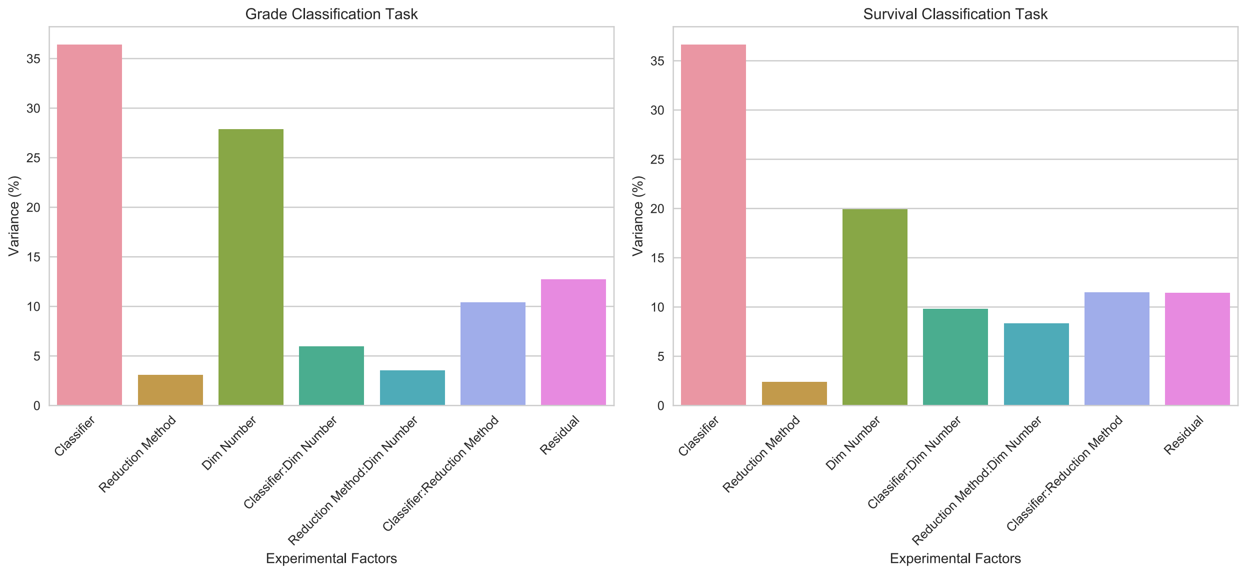

**Figure 4.** Variation of AUC explained by experimental factors and their interactions for the grade classification task (left) and the survival classification task (right).

## Discussion

Several studies have built radiomics based predictive models for various clinical factors such as tumor grade, prognostic outcome, treatment response, and more. However, to expand the radiomics community, studies utilizing open-source data, tools, and machine learning models, such as those used in our current investigation, are necessary.

In a series of papers by Parmar et. al, computed tomography (CT) radiomic machine learning models constructed with various feature selection filter methods and classifier methods were evaluated for predictive performance and stability [8,9]. They show that certain machine learning models perform differently depending on the cancer type, e.g. head and neck vs. lung. Therefore, it is important to test these methods in different cancer types and different imaging modalities. Additionally, Zhang et al., performed a similar study on lung CT with unsupervised dimensionality reduction methods and proposed dimensionality reduction methods have the potential to be superior to filter methods [24]. In this study, we further demonstrate the variability of machine learning models constructed from different classifiers and dimensionality reduction techniques in a different cancer type (glioma) and imaging modality (MRI). We demonstrate that for both tasks, dimensionality reduction techniques are often lower than, or comparable to filtering methods. Specifically, we show that FA can be an improvement over PCA, which was suggested by Zhang et al. to be the best method for dimensionality reduction in radiomic studies.

The presence of ET in T1ce MRI scans are often used as a distinctive marker when attempting to distinguish LGG from HGG [4]. However, since we have only used LGG samples that contain ET components, we suggest radiomics provides novel information about underlying phenotype normally not possible in the radiological setting. Glioma grade is histopatholgically diagnosed, i.e. a biopsy must be taken for classification [28]. With our radiomics approach, we suggest that imaging data may be a useful supplement to histological data. In this study, we have only classified LGG from HGG, but more grade subclasses can be assessed using these radiomics methods, e.g. grade 1 vs grade 2 vs grade 3 vs grade 4. Previous studies have attempted to build machine learning models for glioma grade classification with dimensionality reduction techniques [29] or other feature selection methods [30], but our results show higher predictive performance, possibly due to a larger training set and class balancing with SMOTE.

Predictive performance for grade classification is much higher when compared to survival classification. This is not surprising as each classification task has its own set of optimal radiomic biomarkers that are linked to underlying biological significance. For example, the combination of shape, first-order statistics, texture, and wavelet features utilized through dimensionality reduction leads to higher predictive performance than diameter features alone for the grade classification task (Fig. 2, left). However, this is not the case for the survival classification task. Moreover, using diameter features alone in survival prediction leads to higher predictive performance than using dimensionality reduction or filter techniques with all radiomic features (Fig. 2, right). Tables S3 and S4 demonstrate the distinction in using different combinations of feature subsets (shape, statistics/texture, wavelet) for these tasks built with models using all features. Clearly, there is useful predictive information in each of the subsets for the grade classification task (Table S3). However, it becomes apparent that survival data only gleams major predictive power from shape features (Table S4). In previous studies it has been shown that texture features are difficult to gain predictive power from in GBM, with AUC values routinely falling < 0.6 [12,31]. It may be the case that current intensity based features are simply not strongly linked to survival outcome in GBM, but further studies are necessary before coming to these conclusions. This study has taken a coarse approach to building machine learning models, so it may very well be the case that more refined models for survival prediction can create useful texture based radiomics signatures for GBM survival prediction with high AUC values.

Our results demonstrate that for both classification tasks, among dimensionality reduction methods, FA yielded the highest predictive performance. Similarly, among classifier methods, MLP, LR, and KNN yielded the highest predictive performance and stability. In addition, DT tended to perform poorly for both classification tasks. This possibly points to an underlying radiomic structure in the BraTS Dataset that is preferentially fit by certain machine learning models. Where results start to significantly diverge is in the implementation of the SVM classifier. For the grade classification task, SVMs tend to perform relatively well with all feature selection methods with the exception of ICA. For the survival classification task, SVMs tend to perform poorly with all feature selection methods with the exception of FA. Interestingly, previous studies in different cancer types have suggested RF to be the best classifier method for radiomics studies [8,9,17,24], but it does not score among the best classifier methods for either task in our study.

For both classification tasks, the classifier method was the most significant contribution to variability in predictive performance (Fig. 4). This is a trend that has commonly been observed in radiomic studies investigating machine learning models using different classifiers and feature selection methods [8,9]. Oppositely, Zhang et al. observed that the dimensionality reduction method plays a larger role in predictive performance variability [24]. In our study, we have also investigated the role the number of dimensions has on variability, and found it has a larger role then the dimensionality reduction method used. To our knowledge no other studies have investigated this factors effect on predictive performance.

Some limitations of our study are as follows. Regarding image pre-processing, we have only utilized a simple method of intensity normalization (Z-score) due to its availability in Pyradiomics. Unlike CT imaging, MRI intensity is expressed in arbitrary units, necessitating intensity standardization before radiomic analysis. More refined methods of intensity normalization, such as histogram-based methods [32,33], should be explored in future studies. In addition, we have not taken advantage of classifier hyperparameter tuning, and instead relied on default hyperparameter settings to save on computational costs. Future studies should employ hyperparameter tuning to increase predictive performance and stability. While our study explores stability of our classifiers, it should be noted that RSD is only an empirical method that should not be directly compared with other studies but only as a relative reference between classifiers in a given study. Additionally, our definition of a top classifier is relative to other classifiers studied, so should not be taken as all encompassing.

As our study demonstrates, the actual decision of which classifiers and feature reduction techniques to utilize requires user experimentation with their dataset. There appears to be no "best" classifier or feature reduction technique for radiomics studies in general, but instead different models that should be preferentially implemented depending on the types of datasets involved.

## References

1 Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **278**, 563-577, doi:10.1148/radiol.2015151169 (2016).

2 Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **5**, 4006, doi:10.1038/ncomms5006 https://www.nature.com/articles/ncomms5006 - supplementary-information (2014).

3 Network, T. C. G. A. R. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine* **372**, 2481-2498, doi:10.1056/NEJMoa1402121 (2015).

4 Upadhyay, N. & Waldman, A. D. Conventional MRI evaluation of gliomas. *The British Journal of Radiology* **84**, S107-S111, doi:10.1259/bjr/65711810 (2011).

5 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* **24**, 719-729, doi:https://doi.org/10.1016/j.mric.2016.06.006 (2016).

6 Narang, S., Lehrer, M., Yang, D., Lee, J. & Rao, A. Radiomics in glioblastoma: current status, challenges and potential opportunities. *Translational Cancer Research* **5**, 383-397 (2016).

7 Chaddad, A., Zinn, P. O. & Colen, R. R. in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI).* 84-87.

8 Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports* **5**, 13087, doi:10.1038/srep13087 https://www.nature.com/articles/srep13087 - supplementary-information (2015).

9 Parmar, C. *et al.* Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Frontiers in Oncology* **5**, 272, doi:10.3389/fonc.2015.00272 (2015).

10 Wang, J. *et al.* 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*, 1-9, doi:10.1007/s00330-017-4800-5 (2017).

11 Ingrisch, M. *et al.* Radiomic Analysis Reveals Prognostic Information in T1-Weighted Baseline Magnetic Resonance Imaging in Patients With Glioblastoma. *Investigative Radiology* **52**, 360-366, doi:10.1097/rli.0000000000000349 (2017).

12 Grossmann, P. *et al.* Quantitative Imaging Biomarkers for Risk Stratification of Patients with Recurrent Glioblastoma Treated with Bevacizumab. *Neuro-Oncology*, nox092 (2017).

13 Zacharaki, E. I. *et al.* Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magnetic resonance in medicine* **62**, 1609-1618 (2009).

14 Zacharaki, E. I. *et al.* Survival Analysis of Patients with High-Grade Gliomas Based on Data Mining of Imaging Variables. *American Journal of Neuroradiology* **33**, 1065-1071, doi:10.3174/ajnr.A2939 (2012).

15 Wu, W. *et al.* Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Frontiers in Oncology* **6**, 71, doi:10.3389/fonc.2016.00071 (2016).

16 Parekh, V. & Jacobs, M. A. Radiomics: a new application from established techniques. *Expert Review of Precision Medicine and Drug Development* **1**, 207-226, doi:10.1080/23808993.2016.1164013 (2016).

17 Zhang, B. *et al.* Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Letters* **403**, 21-27, doi:http://dx.doi.org/10.1016/j.canlet.2017.06.004 (2017).

18 Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34**, 1993-2024, doi:10.1109/TMI.2014.2377694 (2015).

19 Walid, M. S. Prognostic Factors for Long-Term Survival after Glioblastoma. *The Permanente Journal* **12**, 45-48 (2008).

20 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. *Cancer Research* (2017).

21 Kotsiantis, S. Feature selection for machine learning classification problems: a recent overview. (2011).

22 James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*. Vol. 112 (Springer, 2013).

23 Ghodsi, A. Dimensionality reduction a short tutorial.

24 Zhang, Y., Oikonomou, A., Wong, A., Haider, M. A. & Khalvati, F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Scientific Reports* **7** (2017).

25 Thornton, C., Hutter, F., Hoos, H. H. & Leyton-Brown, K. in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* 847-855 (ACM).

26 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

27 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321-357 (2002).

28 Grier, J. T. & Batchelor, T. Low-Grade Gliomas in Adults. *The Oncologist* **11**, 681-693, doi:10.1634/theoncologist.11-6-681 (2006).

29 Zacharaki, E. I., Kanas, V. G. & Davatzikos, C. Investigating machine learning techniques for MRI-based classification of brain neoplasms. *International Journal of Computer Assisted Radiology and Surgery* **6**, 821-828, doi:10.1007/s11548-011-0559-3 (2011).

30 Qin, J.-b. *et al.* Grading of Gliomas by Using Radiomic Features on Multiple Magnetic Resonance Imaging (MRI) Sequences. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research* **23**, 2168-2178, doi:10.12659/MSM.901270 (2017).

31 Yang, D., Rao, G., Martinez, J., Veeraraghavan, A. & Rao, A. Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Medical Physics* **42**, 6725-6735, doi:10.1118/1.4934373 (2015).

32 Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6**, 9-19 (2014).

33 Sun, X. *et al.* Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomedical engineering online* **14**, 73 (2015).