

An Investigation of Machine Learning Methods for Delta-radiomic Feature Analysis

by

Yushi Chang

Graduate Program in Medical Physics
Duke Kunshan University and Duke University

Date: _____

Approved:

Fang-Fang Yin, Supervisor

James Bowsher

Xiangpeng Zheng

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the
Graduate Program in Medical Physics in the Graduate School
of Duke Kunshan University and Duke University

2018

ABSTRACT

An Investigation of Machine Learning Methods for Delta-radiomic Feature Analysis

by

Yushi Chang

Graduate Program in Medical Physics
Duke Kunshan University and Duke University

Date: _____

Approved:

Fang-Fang Yin, Supervisor

James Bowsher

Xiangpeng Zheng

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the
Graduate Program in Medical Physics in the Graduate School of
Duke Kunshan University and Duke University
2018

Copyright by
Yushi Chang
2018

Abstract

Background: Radiomics is a process of converting medical images into high-dimensional quantitative features and the subsequent mining these features for providing decision support. It is conducted as a potential noninvasive, low-cost, and patient-specific routine clinical tool. Building a predictive model which is reliable, efficient, and accurate is a vital part for the success of radiomics. Machine learning method is a powerful tool to achieve this. Feature extraction strongly affects the performance. Delta-feature is one way of feature extraction methods to reflect the temporal variation in tumor phenotype, hence it could provide better treatment-specific assessment.

Purpose: To compare the performance of using pre-treatment features and delta-features for assessing the brain radiosurgery treatment response, and to investigate the performance of different combinations of machine learning methods for feature selection and for feature classification.

Materials and Methods: A cohort of 12 patients with brain treated by radiosurgery was included in this research. The pre-treatment, one-week post-treatment, and two-month post-treatment T1 and T2 FLAIR MR images were acquired. 61 radiomic features were extracted from the gross tumor volume (GTV) of each image. The delta-features from pre-treatment to two post-treatment time points were calculated. With leave-one-out sampling, pre-treatment features and the two sets of delta-features were separately

input into a univariate Cox regression model and a machine learning model (L1-regularized logistic regression [L1-LR], random forest [RF] or neural network [NN]) for feature selection. Then a machine learning method (L1-LR, L2-regularized logistic regression [L2-LR], RF, NN, kernel support vector machine [Kernel-SVM], linear-SVM, or naïve bayes [NB]) was used to build a classification model to predict overall survival. The performance of each model combination and feature type was estimated by the area under receiver operating characteristic (ROC) curve (AUC).

Results: The AUC of one-week delta-features was significantly higher than that of pre-treatment features (p-values < 0.0001) and two-month delta-features (p-value= 0.000). The model combinations of L1-LR for feature selection and RF for classification as well as RF for feature selection and NB for classification based on one-week delta-features presented the highest AUC values (both AUC=0.944).

Conclusions: This work potentially implied that the delta-features could be better in predicting treatment response than pre-treatment features, and the time point of computing the delta-features was a vital factor in assessment performance. Analyzing delta-features using a suitable machine learning approach is potentially a powerful tool for assessing treatment response.

Keywords: Radiomics, delta-feature, machine learning, brain tumor, radiosurgery

Contents

Abstract	iv
List of Tables.....	ix
List of Figures	x
Acknowledgements	xii
1. Introduction	1
1.1 Background	1
1.1.1 Medical Imaging and Radiomics	1
1.1.2 General Research Framework of Radiomics Studies	3
1.1.2.1 Image Acquisition.....	4
1.1.2.2 Image Segmentation.....	4
1.1.2.3 Feature Extraction.....	5
1.1.2.4 Feature Analysis	6
1.1.3 Machine Learning (ML) in Radiomic Feature Analysis.....	8
1.2 Review of Previous Work.....	9
1.2.1 Previous Work on Delta-radiomic features.....	9
1.2.2 Previous Work on Machine Learning (ML) in Radiomics.....	11
1.3 Research Aims.....	14
2. Materials and Methods.....	15
2.1 Materials	15
2.2 Methods	16
2.2.1 Research Workflow	16

2.2.2	Image Acquisition	17
2.2.3	Image Segmentation	17
2.2.4	Feature Extraction	17
2.2.4.1	Preliminary Feature Extraction	17
2.2.4.2	Delta-feature Computation.....	20
2.2.5	Feature Selection.....	20
2.2.5.1	Step 1: Univariate Feature Selection: Cox Regression Model.....	20
2.2.5.2	Step 2: Multivariate Feature Selection: Machine Learning Methods	21
2.2.6	Classification.....	23
2.2.7	Model Evaluation: ROC Analysis	26
3.	Results.....	29
3.1	Prognostic Performance.....	29
3.2	Model Performance with Varied Parameters	30
3.2.1	Models Built with Pre-treatment Features.....	31
3.2.2	Models Built with Delta1- Features	32
3.2.3	Models Built with Delta2- Features	34
4.	Discussions.....	36
4.1	Innovation and Impact	36
4.2	Machine Learning Feature Selection—Parameter Adjustment.....	37
4.3	Feature Redundancy after Cox Regression Model	38
4.3.1	Pre-treatment Features selected by Cox Regression Model	38
4.3.2	Delta1-Features selected by Cox Regression Model.....	39
4.3.3	Delta2-Features selected by Cox Regression Model.....	40

4.4 Discussion on ROC Analysis	44
5. Conclusion and Future Work.....	44
5.1 Conclusion.....	44
5.2 Future Work	45
References	47

List of Tables

Table 1 : Clinical characteristics of the brain tumor patients.....	15
Table 2 : Radiomic features calculated by the in-house feature extraction tool.....	18
Table 3 : AUC values of model combinations with pre-treatment and delta-features. ..	29

List of Figures

Figure 1: Publication counts related to radiomics in PubMed Database (U.S. NLM at NIH) till February 2018.....	3
Figure 2: Flowchart for the general framework of radiomics.....	3
Figure 3: Research workflow of the present study.....	16
Figure 4: A diagram illustrating neural network feature selection structure.....	23
Figure 5: Best plane maximizes the margin (Bennett & Campbell, 2000).	25
Figure 6: Categories of prediction or diagnosis results in ROC analysis.....	27
Figure 7: AUC values of model combinations with L1-LR feature selection built by pre-treatment features.	31
Figure 8: AUC values of model combinations with RF feature selection built by pre-treatment features.	31
Figure 9: AUC values of model combinations with NN feature selection built by pre-treatment features.	32
Figure 10: AUC values of model combinations with L1-LR feature selection built by delta1- features.	32
Figure 11: AUC values of model combinations with NN feature selection built by delta1- features.	33
Figure 12: AUC values of model combinations with RF feature selection built by delta1- features.	33
Figure 13: AUC values of model combinations with L1-LR feature selection built by delta2- features.	34
Figure 14: AUC values of model combinations with NN feature selection built by delta2- features.	34
Figure 15: AUC values of model combinations with RF feature selection built by delta2- features.	35
Figure 16: Pre-treatment features selected by Cox model for more than once.	39

Figure 17: Delta1-features selected by Cox model for more than once.	39
Figure 18: Delta2-features selected by Cox model for more than once.	40
Figure 19: Spearman’s correlations among the pre-treatment feature selected by Cox regression model.	41
Figure 20: Spearman’s correlations among the delta1- feature selected by Cox regression model.	42
Figure 21: Spearman’s correlations among the delta2- feature selected by Cox regression model.	43

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Fang-Fang Yin for the continuous support of my master's thesis and relative studying. You always supported me and kept me motivated. Your guidance helped me in all the time of research and writing of this thesis. Your inspiring questions and insightful comments incited me to explore various perspectives relative to my research. Thank you.

I also want to sincerely thank all other professors at DKU, including Prof. James Bowsher, Prof. David Huang, Prof. Deedra McClearn, Prof. Edith Allen and her husband Chris Allen, Prof. Linda Daniel, and Prof. Jing Cai. You helped me building a solid knowledge foundation in Medical Physics. I not only learned great knowledge from you but also enjoyed the time we were together. You made me feel warm and supported both in China and in U.S. I am so lucky to meet you during my graduate study.

I thank my lab-mates, Kyle Lafata, Ruiqi Geng, and Xiaoyu Duan, for the stimulating discussions, for the support we gave each other. Also I thank all my friends in Medical Physics and Global Health Programs at DKU. I would not forget the nights we worked together for homework and exams, the great meals and events, and the wonderful talks we had. I cannot imagine my graduate study without you.

Last, I would like to thank my family. You are always my greatest support and motivation. I literally cannot achieve anything without you.

1. Introduction

1.1 Background

1.1.1 Medical Imaging and Radiomics

Precision medicine which customizes patient-specific treatment to maximize therapeutic success with minimum side effects (Parmar, Grossmann, Bussink, Lambin, & Aerts, 2015) has become a pursuit in modern healthcare. Medical imaging that can provide highly individualized biological information plays a vital role for precision medicine in routine clinic.

In terms of radiation oncology, tumor biopsy is considered as the golden standard for making precision oncology decisions. However, biopsy has inherent limitations. First, a single tumor biopsy can underestimate the tumor diagnosis, since tumors are spatially and temporally heterogeneous (Longo, 2012). Second, repeated tumor biopsies could increase the risk for patients (Parmar, Grossmann, Bussink, et al., 2015). Third, biopsies can only characterize limited area within tumors instead of the entire tumor. Therefore, a non-invasive approach that can characterize the whole-tumor phenotype is waiting to be developed.

Under this background, one novel approach named “radiomics” is being investigated. This term was formed as a combination of “radiography” and the suffix “omics”. “Omics” is used to describe medical research fields where complex high-dimensional data generated from a single object or sample, such as genomics, proteomics, and metabolomics (Gillies, Kinahan, & Hricak, 2015). Radiomics is a process

to convert medical images (CT, MR, PET images etc.) into high-dimensional quantitative features. Then the features, combined with other patient data, are analyzed to provide decision support (Gillies et al., 2015). Through quantitative analysis of the features, the relationships between medical images and pathophysiology can be revealed.

Unlike biopsies, the radiomics approach can characterize tumors or other tissues non-invasively and provide information about the entire tumor phenotype. Furthermore, it can provide information on intra-tumor heterogeneity (Cao, Liang, Shen, Miller, & Stantz, 2009; Yang & Knopp, 2011). Intra-tumor heterogeneity refers to the existence of habitats with distinct genotype or phenotype within a primary tumor and its metastases so that they may have dissimilative biological behaviors (Fisher, Puztai, & Swanton, 2013). It has been reported to be associated with pathophysiology.

Radiomics is potential to be developed as a non-invasive, low-cost, individualized decision support tool for routine clinic practice. First proposed in 2012 (Lambin et al., 2012), radiomics has been reported to be associated with tumor stage (Liang et al., 2016), histology (Wu et al., 2016), recurrence (Mattonen et al.) and patient overall survival (Aerts et al., 2014).

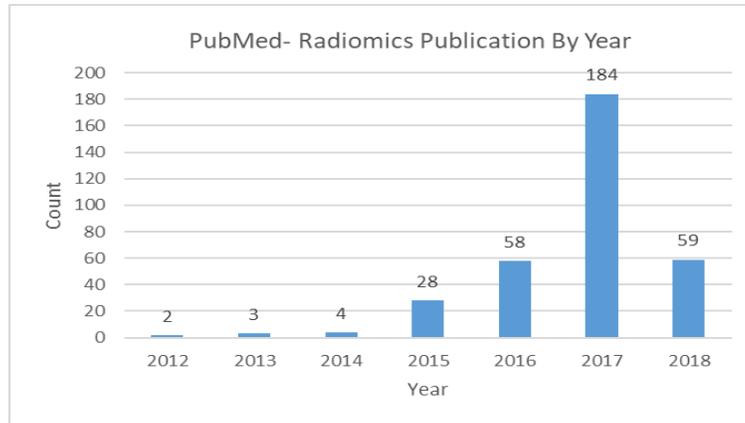


Figure 1: Publication counts related to radiomics in PubMed Database (U.S. NLM at NIH) till February 2018.

1.1.2 General Research Framework of Radiomics Studies

A general research framework of a radiomics study consists of four steps: (1) Acquire radiographic images; (2) Segment the radiographic images to obtain region of interests (ROIs) which usually contain tumors; (3) Extract features from the ROIs; (4) Analyze the features to provide diagnostic, prognostic, or predictive decision support (Lambin et al., 2012). This framework can be seen in figure 2. Each step is essential to obtain a reliable decision.

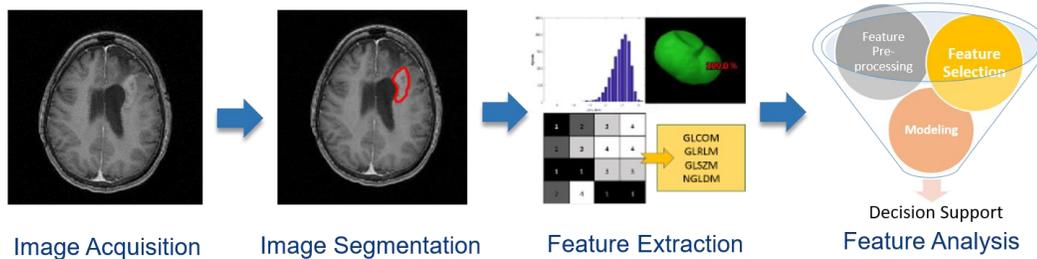


Figure 2: Flowchart for the general framework of radiomics.

1.1.2.1 Image Acquisition

Image acquisition refers to the process of acquiring medical images from imaging machines. During this process, the image acquisition and construction parameters should be reported, because variations in these parameters could introduce non-biological change to the images. Meanwhile, there has been emerging efforts to advance quantitative imaging including defining image acquisition and reconstruction standards (Clarke et al. 2008), such as the Quantitative Imaging Network (QIN) initiated by National Cancer Institute (NCI) and Quantitative Imaging Biomarkers Alliance (QIBA) initiated by Radiological Society of North America (RSNA).

A variety of medical images are involved in radiomics research, including computed tomography (CT), cone-beam computed tomography (CBCT), magnetic resonance imaging (MRI), positron emission tomography (PET), PET/CT, ultrasound, and mammography imaging.

1.1.2.2 Image Segmentation

Image segmentation is the process of identifying and delineating regions of interest (ROIs) in images (Udupa et al., 2006). First, identification is the process of determining where the ROI is in the image. Second, delineation is the process of contouring the ROI point-by-point at precise spatial position (Udupa et al., 2006). Image segmentation is a crucial procedure in radiomics, because all the quantitative features are generated from the segmented ROIs. Despite decades of research (Pal & Pal, 1993), it is still a challenging and controversial field, since many ROIs have indistinct borders

and the ground truth or reproducibility of the segmentation is hard to be authenticated (Pal & Pal, 1993).

Image segmentation can be accomplished manually (completely by human), automatically (completely by computer), or semi-automatically (human-correction based on computer segmentation).

1.1.2.3 Feature Extraction

Radiomics features are quantitative descriptors of the phenotypic characteristics of the ROI. Feature extraction is the process of mathematically computing the features from the ROI.

Features can be divided into several categories based on their information type: (1) Morphological features, which consider the geometric aspects of the ROI, such as shape and volume. (2) First-order features, which are computed based on the distribution of individual voxel intensity within the ROI, not considering the spatial relationship. This category of features contains local intensity features, statistical features, and intensity histogram-based features (Zwanenburg, Leger, Vallières, & Löck, 2016). (3) Second-order features, or texture features, which are designed to quantify the perceived texture of an image and provide spatial information of intensities in an ROI (Shapiro & Stockman, 2000). Hence, these features provide intra-tumor heterogeneity information. The texture features are subcategorized by matrices where they obtained from, including grey level co-occurrence matrix (GLCOM) (Haralick & Shanmugam, 1973), grey level run-length matrix (GLRLM) (Galloway, 1975), grey level size-zone

matrix (GLSZM) (Thibault, Angulo, & Meyer, 2011), neighborhood gray tone difference matrix (NGTDM) (Amadasun & King, 1989), and neighboring grey level dependence matrix (NGLDM) (Sun & Wee, 1983). The computation methods of the matrices and features can be referred to Zwanenburg's publication in 2016 (Zwanenburg et al., 2016).

(4) Higher-order features. They are extracted by imposing filters on the image and extracting repetitive or non-repetitive patterns (Gillies et al., 2015). This category includes wavelet transforms, Laplacian transforms of Gaussian bandpass filters, fractal analyses, Minkowski functionals etc.

For many cases, the features are extracted from images at a single time point, which is usually pre-treatment. However, sometimes features of multiple time points are available, such as pre- and post- treatments. Under this circumstance, features extracted from images at different time points could either be analyzed separately or their differences can be taken for feature analysis. The differences which reflect changes between the radiomic features at different time points are defined as "delta-features". Delta-features reflect the temporal variations in tumor phenotype, and could potentially provide better patient-specific treatment assessment (Xenia Fave et al., 2017).

1.1.2.4 Feature Analysis

Feature analysis is the process of turning the numerous quantitative features into useful information and hence to provide decision support. This process could consist of several parts: (1) feature pre-processing; (2) feature selection; (3) modeling; (4) decision

support. Feature analysis is the core of radiomics research. However, it is not necessary to accomplish all the parts in one radiomics study.

Feature pre-processing means to process the features before other data analysis procedures. From a bigger picture, features are essentially data, and data could be noisy, missing and inconsistent. Low-quality data lead to low-quality results. Pre-processing the features can potentially improve the data quality and hence the results of following data mining (Han, Pei, & Kamber, 2011). Data pre-processing includes but not limited to data cleaning, data integration, data transform. Each approach is unique and easily affects the following feature analyses. For example, data cleaning is applied for filling missing data, removing outliers and resolving inconsistency. Data integration merges data from multiple sources into a consistent warehouse. Data transform contains noise smoothing, data normalization and reconstruct features (Han et al., 2011) .

Feature selection refers to the process of keeping only the useful features out of the feature set, yet closely maintain the integrity of the original feature set (Han et al., 2011). Feature selection is applied to reduce the potential model overfitting. Overfitting, which means the model fits the training samples well but fits the test samples badly, originates from the large complexity (too many features) with insufficient data samples. While strategies like cross-validation and early stopping learning process can relieve overfitting, reducing the number of features is a fundamental way of reducing overfitting. Furthermore, feature selection is beneficial to shorten modeling time and find the meaningful features. Feature selection methods can be divided into two types:

univariate feature selection and multivariate feature selection. Features are prioritized according to a criterion. Univariate methods select features only based on the feature's relevancy to the target variable, not considering the interactions within the features. Multivariate methods, on the other hand, consider the interactions within the features during weighting their relevancy to the target variable.

Modeling is the process of discovering the patterns in the large dataset that can be valuable for decision making. The aim of modeling is to generate predictive, diagnostic, or prognosis hypotheses based on experience. With the built model, the hypotheses of new-coming cases are provided to physicians for their reference.

1.1.3 Machine Learning (ML) in Radiomic Feature Analysis

Machine learning can be used for feature selection and modeling in radiomics. Machine learning is the current application of artificial intelligence, which is a scientific discipline that focuses on how computer can automatically detect complex relationships or patterns from empirical data and make accurate decisions (Bishop, 2006).

Machine learning is a suitable tool for radiomics feature analysis. First of all, machine learning is preferred than traditional programming when tasks are too complex to program (Shalev-Shwartz & Ben-David, 2014). These tasks are often related to human behaviors like image understanding. Secondly, machine learning is good for analysis of very large and complex data, such as radiomic features. Traditional, well-specified algorithms cannot detect the information buried in the complex data. In the past decades, machine learning has been widely used in many tasks that require extracting

information from large datasets, such as searching engines, antispam software for emails, video surveillance and financial data analysis (Wang & Summers, 2012). Third, unlike traditional programming algorithms which would react to all the input data in the same way once the constructions and parameters have been set, machine learning would adjust its performance with respect to different input data (Shalev-Shwartz & Ben-David, 2014).

Machine learning models are built by “learning” from experience and converting it into expertise or knowledge (Shalev-Shwartz & Ben-David, 2014). Experience refers to the training data input to the learning algorithm, and the output of the algorithm is expertise or knowledge which usually takes the form of a model that the test data can be executed with to make predictions. During the training and testing process, machine learning considers all input data as pure numbers without meanings. This learning mode enables machine learning to detect relationships in the large and complex radiomics data that human cannot make sense of.

1.2 Review of Previous Work

1.2.1 Previous Work on Delta-radiomic features

By introducing a time component, delta- features could provide better treatment-specific assessment. However, research on delta-radiomics is very limited and still at its early stage.

Many delta-radiomics were built with CT and/ or CBCT images for lung cancers. For example, Fave et. al. (X. Fave et al., 2017) evaluated the predictive value of delta-

features from pre-radiation therapy (RT) and weekly intra-RT CT images of non-small cell lung cancer (NSCLC). They reported that delta-features were significant in predicting local-regional recurrence, but not significant in predicting overall survival and distant metastases. However, this conclusion stayed controversial because the method they used for feature selection could be with problem. They used only the pre-treatment features to build univariate Cox regression model for feature selection, aiming to find the corresponding delta-features to be predictive. However, a more direct and making-sense way could have been using delta-features to build the Cox regression model.

CT images were also used for other tumor sites in delta-radiomics studies. Cunliffe et. al. (Cunliffe et al., 2015) assessed the relationship between CT delta-features and RT dose as well as radiation pneumonitis (RP) development for esophageal cancer. They found that 20 features changed significantly with increasing dose, and 12 features changed significantly for patients with RP. Rao et. al. (Rao et al., 2016) compared CT delta-features and RECIST (Response Evaluation Criteria in Solid Tumors)-based parameters in assessing colorectal liver metastases to chemotherapy. They reported that delta-uniformity and delta-entropy were with higher predictive value than RECIST parameters.

A few delta-radiomics studies included other imaging modalities. Calvalho. et. al. (Carvalho et al.) identified delta-features from pre-RT and second week during RT were predictive to the overall survival of NSCLC. Zhang. et. al. (Z. Zhang et al., 2017) used

delta-features to classify necrosis and progression lesions after gamma knife radiosurgery. They reported that T1-contrast MRI were with greater distinguishing ability than T2 and FLAIR MRI. However, the predictive value of the delta-features was not investigated.

There has been no study on comparing the predictive value difference of pre-RT and delta-features computed at different time points based on MRI. Grossmann et. al. assessed Bevacizumab (Avastin) treatment in recurrent GBM. They identified that 6-week delta-features showed higher prognostic value than 12-week delta-features. However, this study did not include radiation therapy. Timmeren et al. investigated at which time point the CBCT delta-features was well capable for a feature selection method developed based on the longitudinal information, but they did not investigate the predictive value of the selected delta-features (van Timmeren, Leijenaar, van Elmpt, Reymen, & Lambin, 2017).

1.2.2 Previous Work on Machine Learning (ML) in Radiomics

Machine learning algorithms have been prized in radiomics because it can build models to learn complicated relationships between the features and clinical endpoints. They have been used to predicting overall survival (Parmar, Grossmann, Bussink, et al., 2015; Parmar, Grossmann, Rietveld, et al., 2015), tumor histology (Zhu et al., 2018), recurrence (Saha, Harowicz, Wang, & Mazurowski, 2018) and other clinical endpoints for lung (Parmar, Grossmann, Bussink, et al., 2015; Zhu et al., 2018), breast (Saha et al., 2018), head and neck (Parmar, Grossmann, Rietveld, et al., 2015), prostate (Algohary et

al., 2018), pulmonary nodule (Tu, Wang, Pan, Wu, & Wu, 2018), thyroid (Sollini et al., 2018), bladder (Garapati et al., 2017), and brain (Leger et al., 2017) tumors. Commonly used machine learning algorithms include support vector machine (SVM)-based methods, decision tree-based methods, random forest (RF)-based methods, neural network, Bayesian methods, boosting methods, logistic regression (LR), k-nearest neighbor (KNN) and so on.

While machine learning can be used for both feature selection and building predictive models, in most of these studies, the use of ML methods was limited in modeling. For instance, Parmar et al. pioneered in investigating the performance of different machine learning algorithms for building classification model for NSCLC (Parmar, Grossmann, Bussink, et al., 2015) and head and neck squamous cell carcinoma (HNSCC) patients (Parmar, Grossmann, Rietveld, et al., 2015). They used 14 non-ML feature selection methods like joint mutual information and Wilcoxon with 12 ML classifiers such as random forest (RF) and support vector machine (SVM). They identified RF classifier displayed the highest predictive performance for NSCLC and generalized linear model (GLM) for HNSCC. Tu et al. (Tu et al., 2018) researched on classifying early-detected pulmonary nodules from lung cancer screening by machine learning methods. They selected features based on two-tailed t-test followed by machine learning classifiers logistic, sequential minimal optimization (SMO), J48, random forest, and instance-based K-nearest neighbors (IBK). The use of machine learning methods for feature selection was not explored in their studies.

To date only few studies investigated the combinations of machine learning feature selection and machine learning predictive modeling. Leger et al. (Leger et al., 2017) investigated the performance of different combinations of feature selection and learning algorithms for predicting loco-regional tumor control (LRC) and overall survival for head and neck squamous cell carcinoma (HNSCC). However, the machine learning algorithms they used for feature selection were limited to RF-based methods (RF-MD (minimal depth), RF-VI (variable importance), PVI (permutation variable importance)-RF, MSR (maximally selected rank)-RFVI). The modeling methods were either boosting-based or RF-based methods (boosting tree (BT) – Cox, BT-CIndex, boosting gradient linear models (BGLM)- Cox, BGLM- CIndex, random survival forest (RSF), MSR-RF, BT- and BGLM-Weibull survival regression). A wider range of machine learning methods should be evaluated. Zhang et al. (B. Zhang et al., 2017) used machine learning model combinations to predict local failure and distant failure in advanced nasopharyngeal carcinoma (NPC). They researched on more machine learning algorithms, including 5 machine learning feature selection and 9 machine learning classification methods. RF+RF was reported with the highest prognostic performance for their research purpose. The limitation of this study was that no other feature selection steps were applied before machine learning feature selection, which could cause severe overfitting of machine learning feature selection.

So far no such evaluation has been applied for brain tumor radiomics.

1.3 Research Aims

Specific Aim 1:

Investigate the effectiveness of using delta-features as an assessment tool for brain radiosurgery.

1.A Build predictive models based on pre-treatment features and delta-features;

1.B Compare the prognostic values of pre-treatment features and delta-features;

1.C Investigate the difference of using delta-features computed at different time points for assessing brain radiosurgery.

Specific Aim 2:

Investigate the effectiveness of machine learning feature selection methods with machine learning classifiers.

2.A Build both feature selection models and classification models utilizing machine learning methods;

2.B Investigate the performance of different combinations of feature selection and classification models.

2. Materials and Methods

2.1 Materials

For this study, we retrospectively reviewed the images and medical records for 12 patients with brain tumor (6 glioblastoma, 3 anaplastic oligodendroglioma, and 3 anaplastic astrocytoma) treated by SRS. The total prescription dose was either 18 or 24 Gy with single fraction SRS when the maximum tumor diameter was smaller than 3 cm or 25 Gy with 5-fraction SRS when the maximum diameter was equal or larger than 3 cm. The endpoint for this retrospective analysis was overall survival, which was between 5.3 months and 29.4 months. More information about this cohort is summarized in table 1.

Table 1 : Clinical characteristics of the brain tumor patients

Clinical Factors		Number of Patients (n=12)
Sex	Female	3
	Male	9
Age	25- 50	5
	50- 66	7
Tumor Histology	GBM	6
	AO	3
	AA	3
KPS at enrollment	80%	3
	90%	5
	100%	4
WHO grade at enrollment	3	6
	4	6
Total SRS Dose in Gy	18 (1 fraction)	6
	24 (1 fraction)	3
	25 (5 fraction)	3
Max Diameter (cm)	1-3	9
	3-3.7	3
Overall Survival	≤ 1 year	6
	≥ 1 year	6

2.2 Methods

2.2.1 Research Workflow

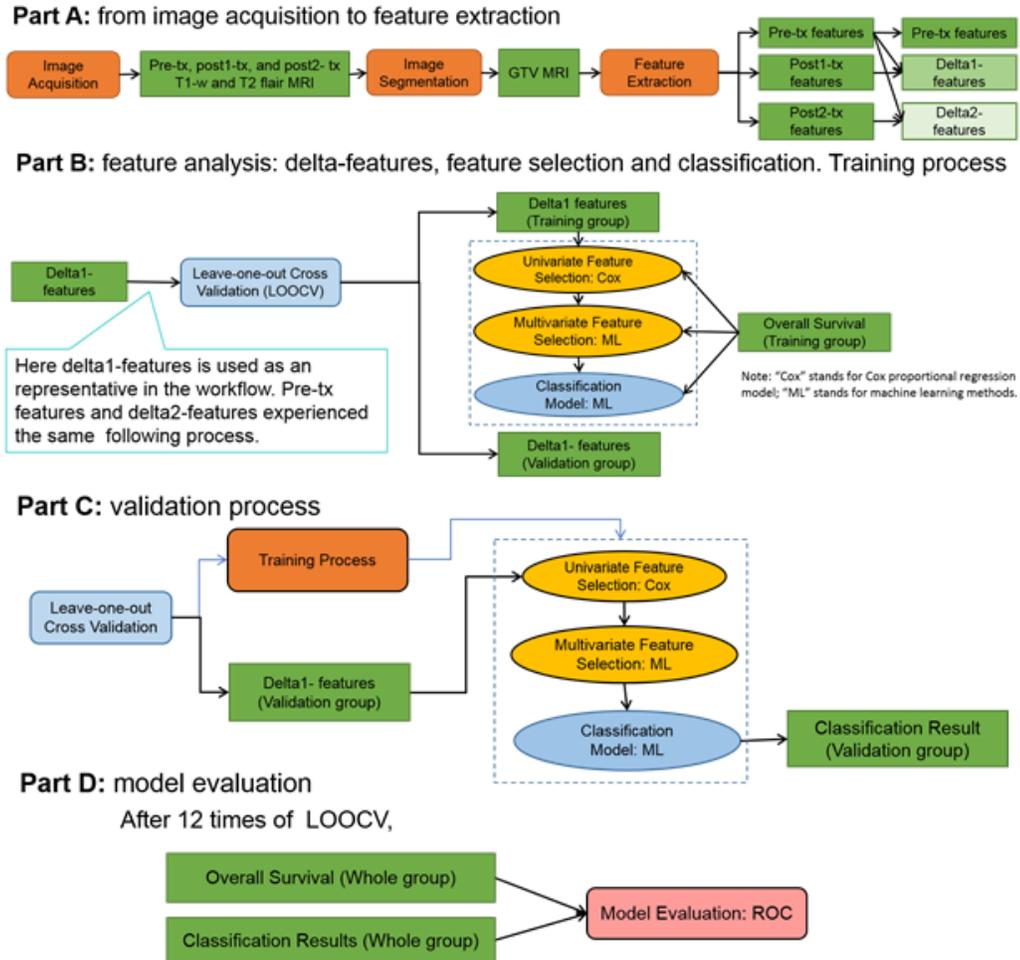


Figure 3: Research workflow of the present study.

In this study, overall survival (OS) was the prediction endpoints and was dichotomized: $OS < 1$ year was labeled as class 0, and $OS \geq 1$ year was labeled as class 1. Since the dataset contained 12 patients, LOOCV would process 12 times. Since leave-one-out cross validation (LOOCV) was used for a cohort with 12 patients, the LOOCV would be executed for 12 times before evaluating the model. ROC analysis was used to

evaluate the performance of each feature selection and classification model combination after 12 times of LOOCV.

Finally, the performances of different model combinations built based on pre-treatment features, delta1-features and delta2-features were compared.

2.2.2 Image Acquisition

For each patient, the pre-treatment, one-week post-treatment, and two-month post-treatment T1-weighted and T2-weighted FLAIR (Fluid Attenuated Inversion Recovery) MR images were acquired from a 1.5 T MR unit (GE Medical System).

2.2.3 Image Segmentation

The three-dimensional gross tumor volume (GTV) was contoured by experienced physicians on the pre-treatment MR images in Eclipse treatment plan system (Varian Ltd., the U.S.). The one-week post-treatment and two-month post-treatment images were registered to the pre-treatment MRIs, and the GTV delineations were transferred to the post-treatment MRIs.

2.2.4 Feature Extraction

2.2.4.1 Preliminary Feature Extraction

First of all, the MR images were resampled into 64 grey-level bins. Resampling the images is more likely to reflect the actual grey level changes in neighboring pixels, since the image noise was largely diminished in this way. Furthermore, it can save computing time with sufficient information amount. Then, 61 features were extracted from the GTV from each type of MRI at each time point, including 22 grey level co-

occurrence matrix (GLCOM) texture features, 11 grey level run length matrix (GLRLM) texture features, 13 grey level size zone matrix (GLSZM) texture features, 5 neighboring grey level difference matrix (NGLDM) features, 6 morphological features and 4 intensity histogram-based features. We used an in-house feature extraction tool developed based on MATLAB 2017a (MATLAB Co. Ltd) to extract the features. The mathematical formulas for computing the features can be referred in reference (Zwanenburg et al, 2016). The radiomic features calculated by the in-house feature extraction tool is listed in Table 2.

Table 2 : Radiomic features calculated by the in-house feature extraction tool.

Feature Type	Feature Name
Intensity histogram-based features (4)	Energy
	Entropy
	Kurtosis
	Skewness
GLCOM texture features (22)	Auto Correlation
	Cluster Prominence
	Cluster Shade
	Cluster Tendency
	Contrast
	Correlation
	Differential Entropy
	Dissimilarity
	Energy
	Entropy
	Homogeneity1
	Homogeneity2
	Info Measure Correlation1
	Info Measure Correlation2
	Inverse Difference Moment Normalized
	Inverse Difference Normalized
Inverse Variance	

Feature Type	Feature Name
	Maximum Probability
	Sum Average
	Sum Entropy
	Sum Variance
	Variance
GLRLM texture features (11)	Short Run Emphasis
	Long Run Emphasis
	Gray Level Non-uniformity
	Run Length Non-uniformity
	Run Percentage
	Low Gray Level Run Emphasis
	High Gray Level Run Emphasis
	Short Run Low Gray Level Emphasis
	Short Run High Gray Level Emphasis
	Long Run Low Gray Level Emphasis
	Long Run High Gray Level Emphasis
GLSZM texture features (13)	Small Zone Emphasis
	Large Zone Emphasis
	Gray Level Non-uniformity
	Size Zone Non-uniformity
	Size Percentage
	Low Gray Level Size Emphasis
	High Gray Level Size Emphasis
	Small Size Low Gray Level Emphasis
	Small Size High Gray Level Emphasis
	Large Size Low Gray Level Emphasis
	Large Size High Gray Level Emphasis
	Variation Of Intensity
Variation Of Area	
NGLDM texture features (5)	Coarseness
	Contrast
	Busyness
	Complexity
	Texture Strength
Morphological features (7)	Compactness1
	Compactness2
	Max 3D Diameter
	Sphericity
	Spherical Disproportion

Feature Type	Feature Name
	Surface Area
	Volume

2.2.4.2 Delta-feature Computation

Delta-features were computed as the relative changes of post-treatment features to pre-treatment features.

$$\Delta_{1 - features} = (Features_{Post1} - Features_{Pre}) / Features_{Pre}$$

Equation 1: Equation for computing delta1-features.

$$\Delta_{2 - features} = (Features_{Post2} - Features_{Pre}) / Features_{Pre}$$

Equation 2: Equation for computing delta2-features.

where $Features_{Pre}$ are the features extracted from pre-treatment MRI, $Features_{Post1}$ and $Features_{Post2}$ are features computed from one-week post-treatment MRI and two-month post-treatment MRI, respectively.

2.2.5 Feature Selection

A two-stage feature selection method was applied in this study.

2.2.5.1 Step 1: Univariate Feature Selection: Cox Regression Model

A univariate Cox regression model for overall survival was fitted for each pre-processed feature. Cox regression model (Cox, 1972) quantifies the effect of one or several variables upon the time when a specified event happens. The probability of happening event, which was death in this study, at time t given that individual is alive at that time is depicted by the concept of hazard. Cox regression model uses a semi-parametric logarithmic linear model

$$H(t|X_i) = H_0(t)\exp(\beta X_i)$$

Equation 3: Hazard function

where $H(t|X_i)$ is the hazard at time t for the given patient. $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$ is the feature set of the i th patient; t is overall survival; β measures the impact of the feature; $H_0(t)$ is the baseline hazard.

Therefore, the hazard ratio is

$$HR(X_i) = \frac{H(t|X_i)}{H_0(t)} = \exp(\beta X_i)$$

Equation 4: Hazard ratio function

As we can see from equation 4, the hazard ratio is then not related with time.

Univariate Cox regression model examines whether a single feature has significant predictive effect compared to baseline hazard. For pre-treatment features, the baseline hazard corresponded to the mean value of X_i , while it corresponded to 0 for delta-features. This criterion was set because we regarded the mean of pre-treatment features as the baseline for pre-treatment features and 0 as the baseline for delta-features. Features with significant value of $p\text{-value} < 0.1$ were selected.

2.2.5.2 Step 2: Multivariate Feature Selection: Machine Learning Methods

Features which are significant in univariate Cox regression model were further selected by 3 machine learning algorithms supervised by overall survival. The features were selected according to their importance in deciding the overall survival in each method.

1) Out-of-bag permutation random forest (Breiman, 2001). The description about random forest can be seen in the following “Machine learning classification model building” part. First, on the basis of a trained random forest model, the out-of-bag samples for building each decision tree in the random forest were defined. The out-of-bag error for each tree was calculated. Then, randomly permute each feature in building the decision tree, and the out-of-bag error again was estimated. Next, the difference between the post- and pre- permutation out-of-bag errors for each feature, normalized by the mean and standard deviation of the differences over all features, was the importance of the feature in deciding the treatment outcome. The number of features selected was varied to decide the optimal parameter for modeling.

2) L1-regularized logistic regression. The description about L1-regularized logistic regression can be seen in the following “Machine learning classification model building” part. It has been shown that using L1 regularization of the parameters, the number of training examples required to learn “well” grows only logarithmically with the number of irrelevant features (Ng, 2004). The importance of a feature was indicated by its weight in the model. The most important features with a sum of importance over a threshold were selected. The threshold was varied to decide the optimal parameter for modeling.

3) Three-layer neural network. The description about three-layered neural network can be seen in the following “Machine learning classification model building” part. For feature selection, the input features were disabled one-by-one, where “disable”

means all weights related to this feature were set to 0, as shown in figure 5. The importance of the feature was indicated by the classification error with this feature disabled. In this experiment, the number of features selected was varied to decide the optimal parameter for modeling.

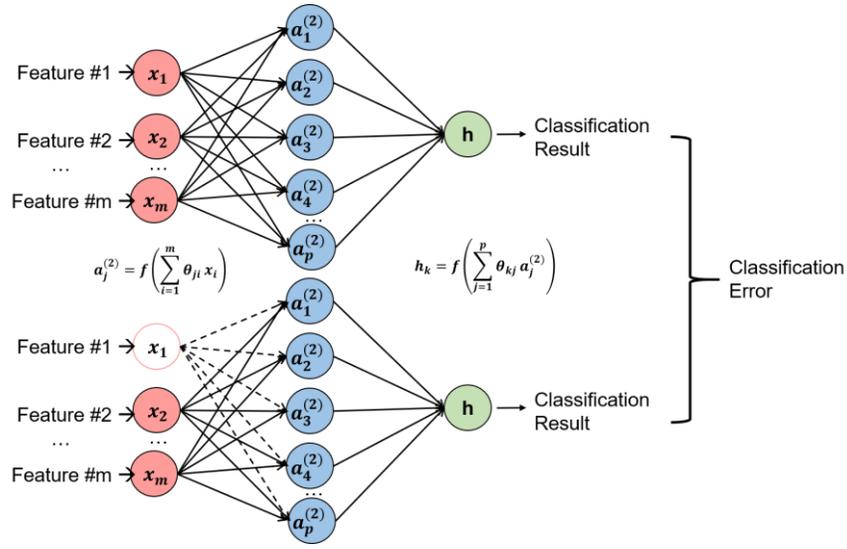


Figure 4: A diagram illustrating neural network feature selection structure.

2.2.6 Classification

Seven machine learning algorithms were used to execute binary classification: patients with overall survival shorter than 1 year were defined as class 0; patients with overall survival equal or longer than 1 year were defined as class 1.

1) Random forest (RF). Random forest classifier consists of a collection of decision tree predictors (Lior, 2014). Each tree makes a prediction, and the prediction result of RF is decided by the most popular class predicted by the decision tree models for classification. Random forest classifier is proven to be robust to noises and outliers and suitable to a small dataset (Breiman, 2001). In this study, we used the “fitcensemble”

function in MATLAB 2017a (MathWorks, Natick, MA, USA) to build the random forest binary classification model with the number of decision trees as 100.

2) L1-regularized logistic regression (L1-LR). Logistic regression can estimate the probability of a hypothesis based on the features $x = [1, x_1, x_2, \dots, x_n]$ by sigmoid function

$$h_{\theta}(x) = \text{sigmoid}(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Equation 4: Sigmoid function

where $\theta = [\theta_0, \theta_1, \theta_2, \dots, \theta_n]$ is the weight factor matrix. θ was optimized by minimizing the cost function:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \left[-y^i \log(h_{\theta}(x^i)) - (1 - y^i) \log(1 - h_{\theta}(x^i)) \right] + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j$$

Equation 5: Cost function of L1-regularized logistic regression

where m is the number of training samples, y^i is the ground true class of the i th training sample, λ is a regularization parameter which is for reducing overfitting.

3) L2-regularized logistic regression (L2-LR). L2-regularized logistic regression was similar to L1-regularized logistic regression, but the regularization term in the cost function was $\frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$.

4) Linear support vector machine (LSVM) (Cortes, Corinna, and Vladimir Vapnik, 1995). Using training data, a SVM classifier can determine a decision hyperplane that maximized the margin between the training dataset and class boundary. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points, as shown in figure 5. Linear SVM considers each training point in the form of (\vec{x}_i, y_i) , where \vec{x}_i is the feature set and y_i is the true class to which the point \vec{x}_i belongs. The hyperplane was determined by making the training points satisfy $\langle \vec{w} \cdot \vec{x} \rangle - b = 0$, where \vec{w} is the normal vector to the hyperplane.

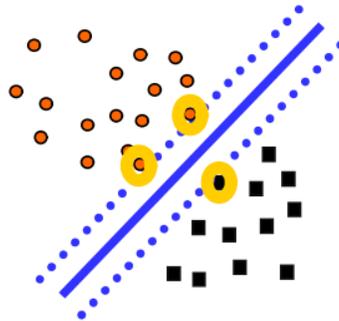


Figure 5: Best plane maximizes the margin (Bennett & Campbell, 2000).

5) Kernel support vector machine (KSVM) (Cortes, Corinna, and Vladimir Vapnik, 1995). Some binary classification problems do not have a simple hyperplane to separate the training points. For these problems, kernel SVM can transform the features into a high dimensional space and determine a hyperplane in the transformed feature space. The transformation was accomplished by replacing dot product by nonlinear

kernel function. We used Gaussian radial basis function (RBF) as the kernel function.

Kernel SVM could allow a better fit (larger-margin hyperplane) to the training dataset, however, it may also increase the generalization error of the algorithm (Bennett & Campbell, 2000).

6) Three-layer back-propagation neural network. We constructed a neural network classifier by back-propagation algorithm (Hecht-Nielsen, 1992) with one hidden layer and one output node. The number of hidden units was set as 10.

7) Naïve bayes. Naïve bayes classifier is based on Bayes's rule with a strong assumption that features are independent each other within each class, i.e.,

$P(\mathbf{X}|C) = \prod_{i=1}^n P(X_i|C)$, where $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is a feature vector and C is a class.

Despite this assumption is hard to be satisfied, naïve bayes appears to work well in practice even when that independence assumption is not valid (Manning, Raghavan, & Schütze, 2008; Rish, 2001).

The output of the classification models was the score of the predicted class being class 1. Here score refers to the posterior probability, which describes the probability of predicted class being 1 with the given feature values.

2.2.7 Model Evaluation: ROC Analysis

Receiver operating characteristics (ROC) curve is used to depict the tradeoff between hit rates (true positive fraction) and false alarm rates (false positive fraction) of classifiers. ROC analysis has the advantages of balancing the trade-offs between

diagnostic accuracy and disease prevalence (Metz, 1978) and providing foundation for numerous other assessments once ROC data are collected and the plots generated (Zweig and Campbell, 1993). Hence, it is one of the most prevalent evaluation criterion in radiomics studies. The prediction or diagnosis results are divided into four categories: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The description of these four types of results are shown in the figure 7.

	Disease is actually present	Disease is actually absent
Disease is predicted to be present	TP	FP
Disease is predicted to be absent	FN	TN

Figure 6: Categories of prediction or diagnosis results in ROC analysis.

$$\text{True positive fraction (TPF)} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Sensitivity}$$

Equation 6: Equation for computing TPF

$$\text{False positive fraction (FPF)} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{Specificity}$$

Equation 7: Equation for computing FPF

The predictive performance of the models was evaluated by area under the ROC curve (AUC). To compute AUC, first the score of the validation group during each LOOCV was obtained. Therefore, after 12 times of LOOCV, each sample in the dataset had been the validation group for once, and their prediction scores were obtained. Then, the predicted classes of the pooled data were dichotomized by a given decision threshold, and TPF and FPF were calculated based on the threshold. Next, to obtain a ROC curve, decision thresholds varied to obtain all distinct TPFs and FPFs. In practice,

the thresholds were selected as all distinct scores. Therefore, the highest threshold produced a “TPF=FPF=0” point on ROC curve, while the lowest threshold produced a “TPF=FPF=1” point on ROC curve. Finally, the AUC was calculated by trapezoidal approximation with all distinct TPFs and FPFs.

3. Results

To compare the different machine learning feature selection and classification model combinations built by pre-treatment features or delta-features, we extracted 122 features from the GTV from T1-weighted and T2-weighted FLAIR images. Leave-one-out validation was used to reduce overfitting, and the predictive performance of each model combination was assessed by AUC.

3.1 Prognostic Performance

The AUC values were recorded in table 3. The AUC values recorded in table 3 were the highest AUC value of each model combination after varying the parameters of the feature selection models.

Table 3 : AUC values of model combinations with pre-treatment and delta-features.

AUC	Feature Selection								
Classification Method	L1-LR			RF			NN		
	Pre	Delta 1	Delta 2	Pre	Delta 1	Delta 2	Pre	Delta 1	Delta 2
LR1	0.361	0.778	0.694	0.417	0.861	0.722	0.306	0.722	0.667
LR2	0.389	0.778	0.694	0.417	0.861	0.722	0.417	0.722	0.667
LSVM	0.306	0.861	0.250	0.333	0.889	0.222	0.333	0.833	0.389
KSVM	0.444	0.861	0.833	0.333	0.917	0.667	0.167	0.889	0.417
RF	0.444	0.944	0.833	0.306	0.889	0.778	0.528	0.778	0.556
NN	0.861	0.861	0.639	0.639	0.889	0.556	0.597	0.833	0.722
NB	0.194	0.889	0.556	0.361	0.944	0.472	0.361	0.833	0.500

From table 3, for each machine learning classifier, the models built with delta1-features always showed the highest AUC values compared with other two feature extractions. Using paired t-test to evaluate the significance of the difference, delta1-

features showed significant different predictive performance both from pre-treatment features (p-value = 0.000) and delta2-features (p-value= 0.000). The combinations for feature selection and classification models of both L1-LR + RF and RF + NB built with delta1-features showed the highest AUC values (AUC=0.944).

Comparing pre-treatment features and delta2-features, models with delta2-features often showed higher AUC values, except for L1-LR and RF feature selection method combined with LSVM or NN classifier. In these 4 model combinations, pre-treatment features showed higher AUC values.

3.2 *Model Performance with Varied Parameters*

The number of features selected by RF and NN feature selection method and the sum weight threshold of L1-LR feature selection method were the parameters need to be determined for optimal modeling performance. The performances of these models with varied parameters were shown in figure 8 to figure 16.

3.2.1 Models Built with Pre-treatment Features

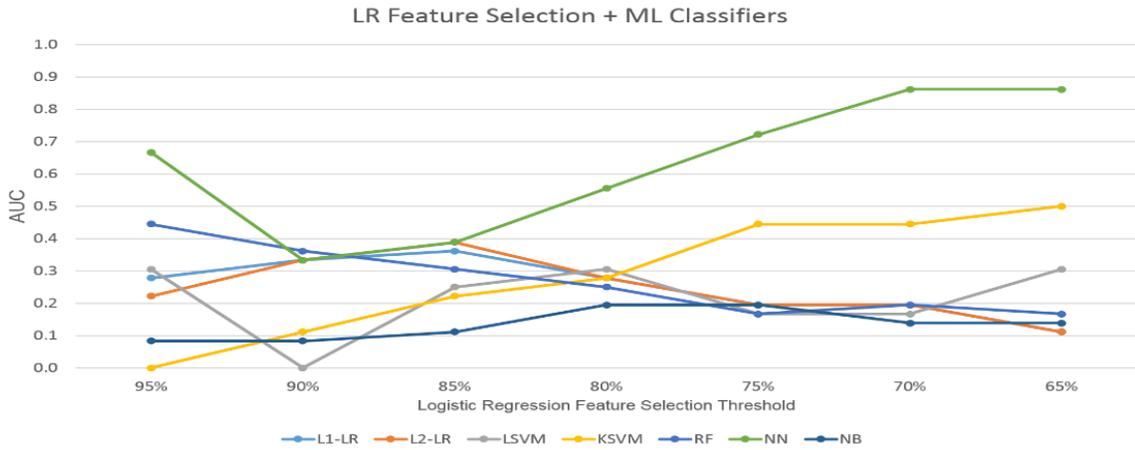


Figure 7: AUC values of model combinations with L1-LR feature selection built by pre-treatment features.

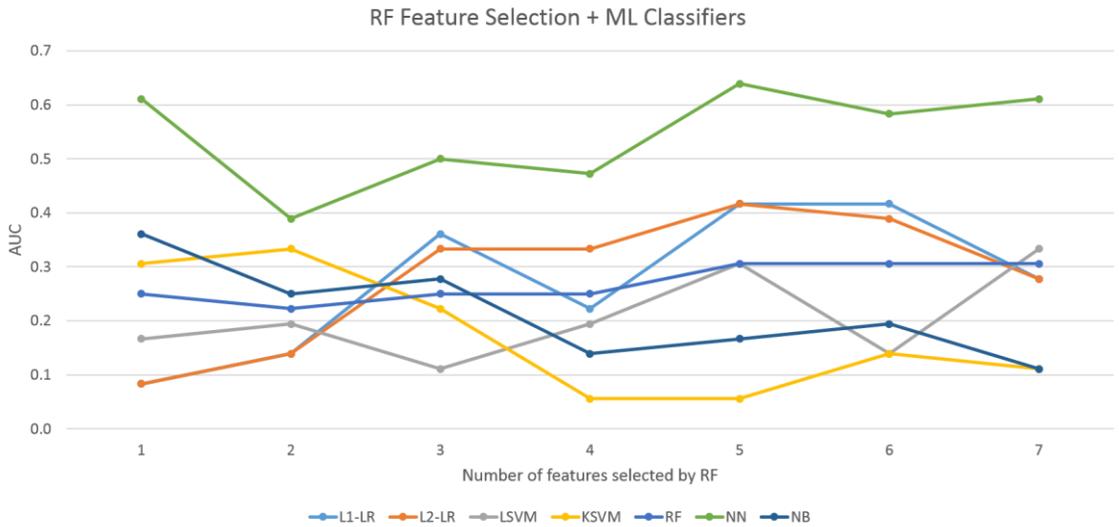


Figure 8: AUC values of model combinations with RF feature selection built by pre-treatment features.

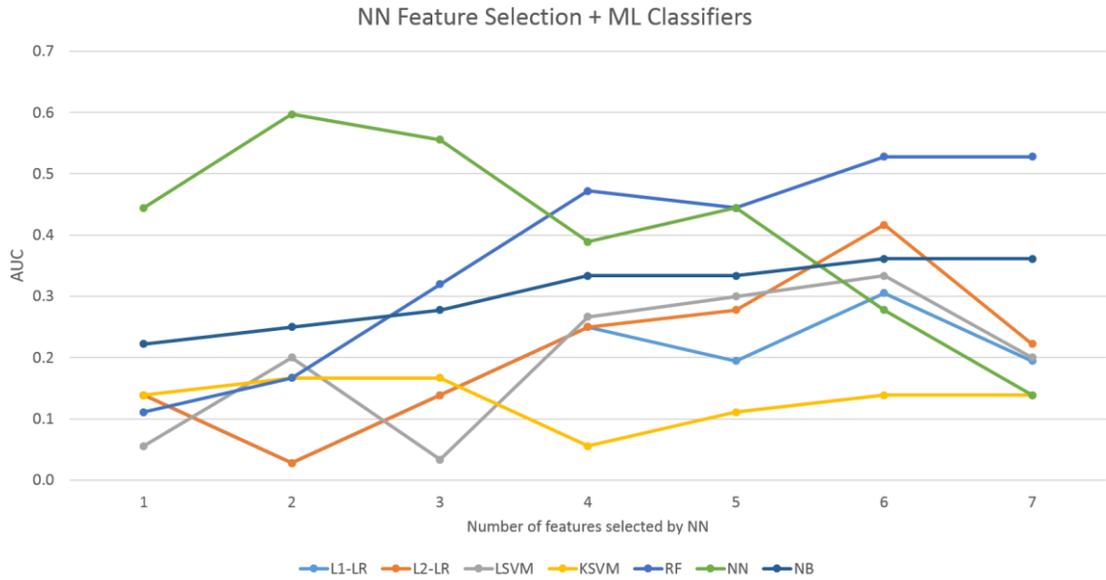


Figure 9: AUC values of model combinations with NN feature selection built by pre-treatment features.

3.2.2 Models Built with Delta1- Features

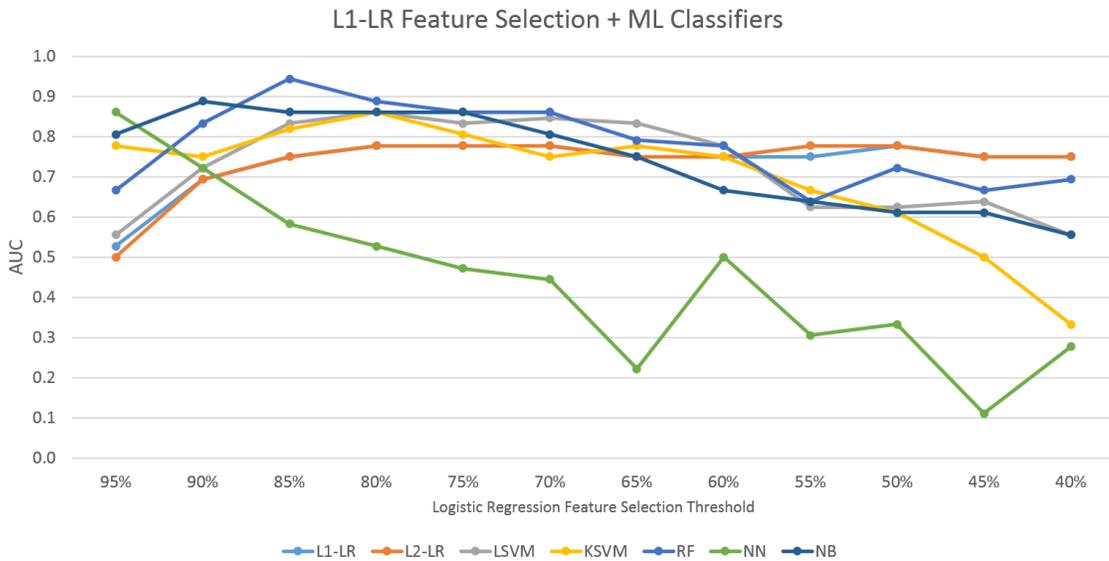


Figure 10: AUC values of model combinations with L1-LR feature selection built by delta1- features.

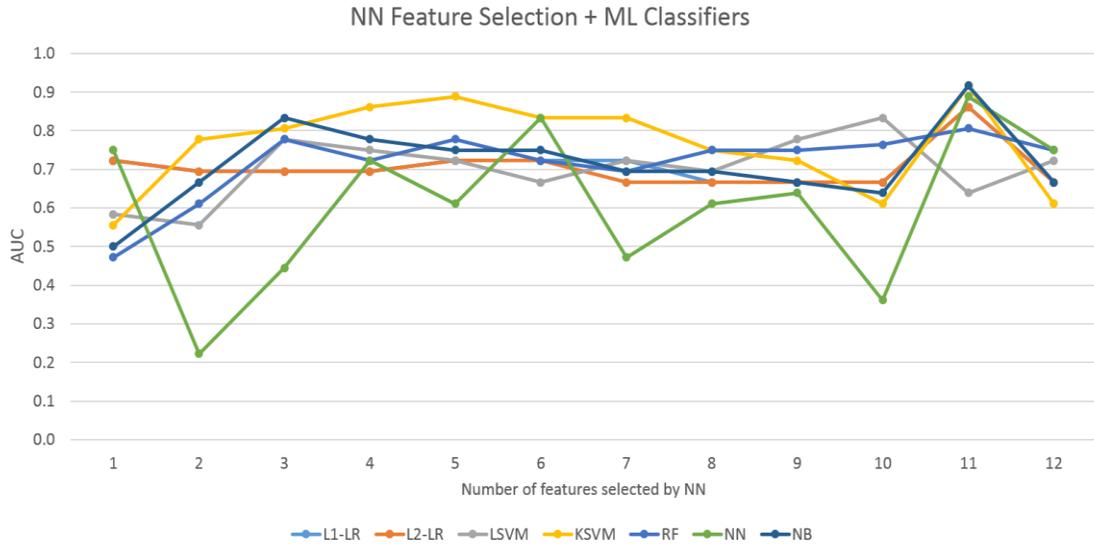


Figure 11: AUC values of model combinations with NN feature selection built by delta1- features.

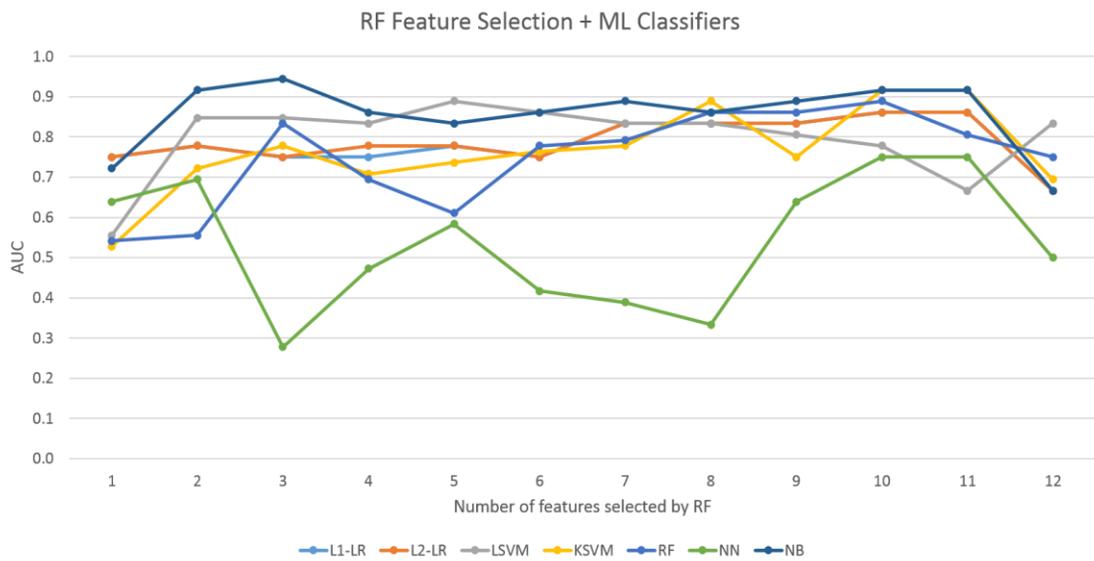


Figure 12: AUC values of model combinations with RF feature selection built by delta1- features.

3.2.3 Models Built with Delta2- Features

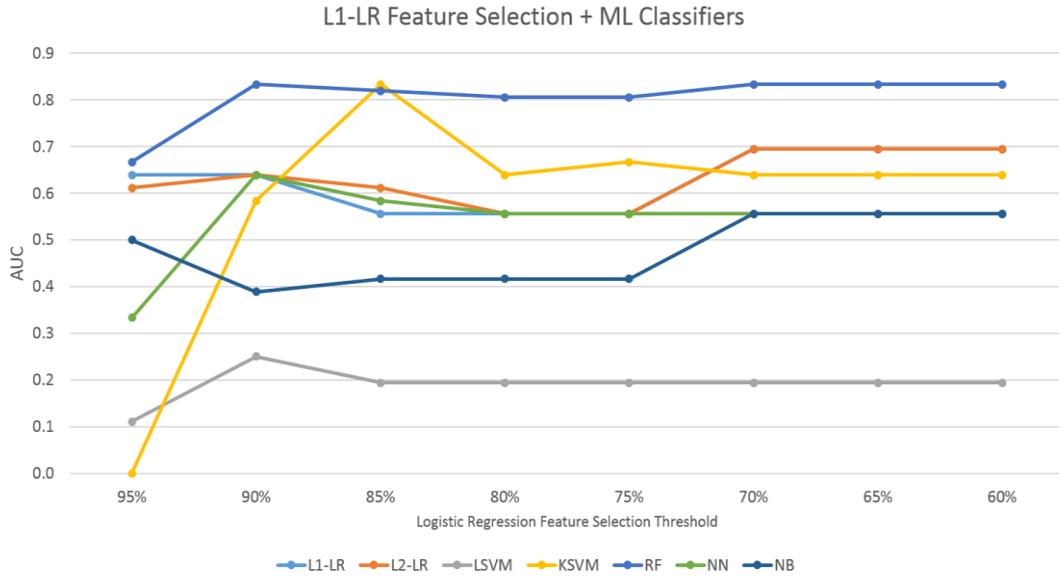


Figure 13: AUC values of model combinations with L1-LR feature selection built by delta2- features.

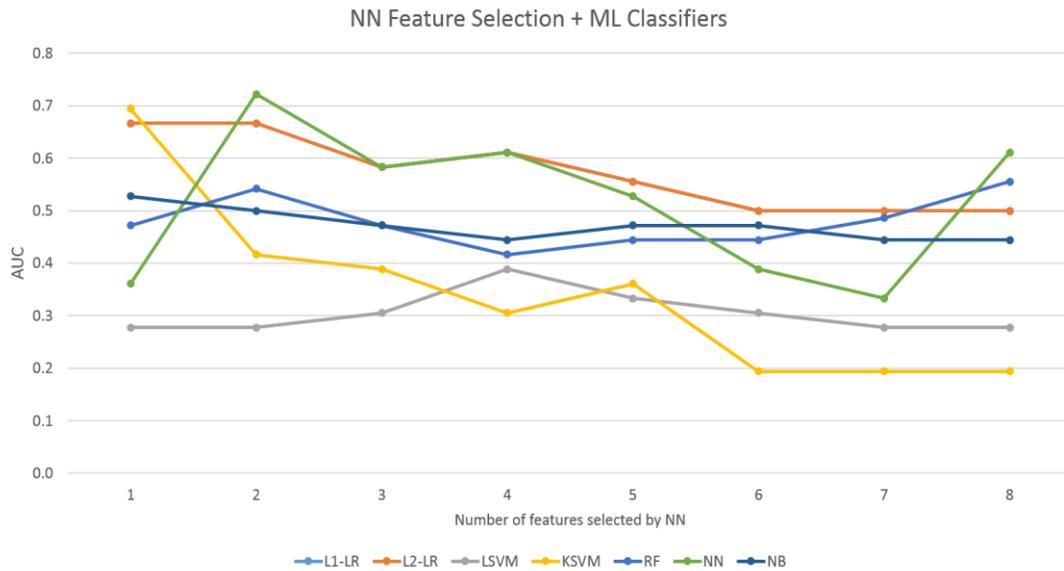


Figure 14: AUC values of model combinations with NN feature selection built by delta2- features.

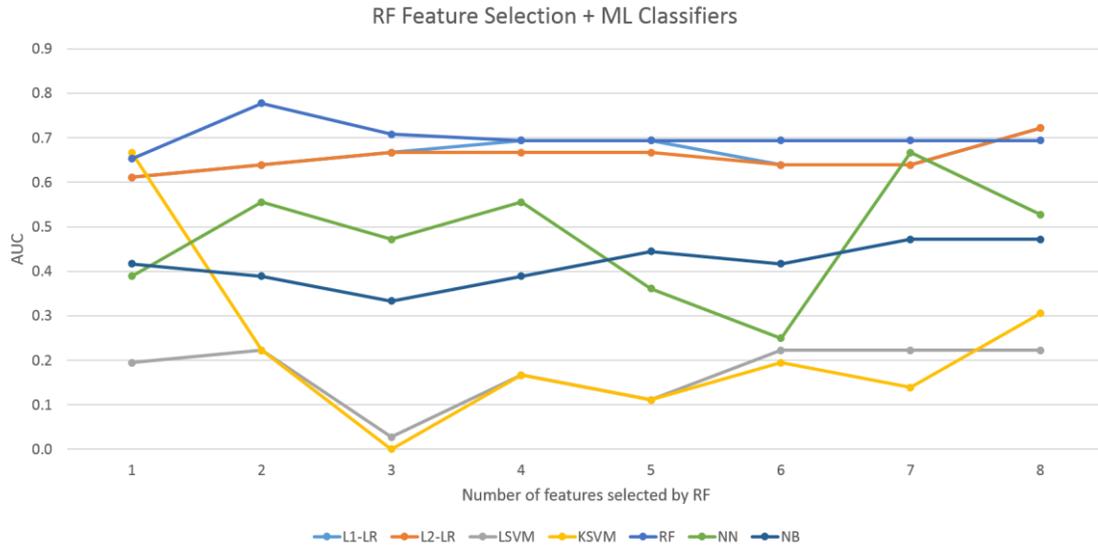


Figure 15: AUC values of model combinations with RF feature selection built by delta2- features.

From figure 8 to figure 15, we can see that changing the feature selection model parameters affected the model performance. However, no obvious trend of the performance was found by varying the parameters for this dataset so far.

Comparatively, delta-features were more stable with different feature selection model parameters, especially with LR, NB, and LSVM classifiers. In particular, delta1-features were more stable with different feature selection model parameters except for the neural network classifier.

Overall speaking the neural network classifier was the most sensitive to the number of features imported into it no matter for which type of feature extraction.

4. Discussions

4.1 Innovation and Impact

One novelty of this study is the dataset. Research on delta-radiomics has not been well promoted. This situation is partly due to the fact that the demanded data are not always available. Under this situation, although the dataset included in this research is small, it provided MRI information both pre- and post- brain radiosurgery. Hence, the predictive value of delta-features for assessing brain radiosurgery could be firstly investigated.

From table 3, we can see that the three feature extraction methods, pre-treatment features, one-week delta-features and two-month delta-features were with different predictive values with respect to assessing brain tumor radiosurgery. Compared with other two feature extractions, one-week delta-features displayed higher predictive performances.

Although it is hard to make a generalization from this small dataset, we can reasonably infer that delta-features are promising to provide better predictive decision support than pre-treatment features. Additionally, the time point for computing delta-features would affect the model performance. A generalized conclusion requires a larger dataset to validate this result.

Another novelty is using machine learning methods for both feature selection and classification and for seeking an optimal combination. Although a few studies has done this comparison, but either the machine learning methods they used were limited

(Leger et al., 2017) or they did not consider the overfitting of machine learning feature selection (Z. Zhang et al., 2017). In this study, univariate Cox regression model selected features which could be predictive to overall survival, without considering the interactions among the features. This deficiency was offset by machine learning feature selection methods, which considered both the relevancy between features and overall survival as well as the cooperation among the features. Another benefit of applying Cox regression model was to reduce the overfitting of machine learning feature selection models. This methodology could be promoted to other imaging modalities and other tumor sites for predictive purpose.

4.2 Machine Learning Feature Selection—Parameter Adjustment

As we can see from the results, the parameters in machine learning feature selection models greatly affected the model performance. The parameters seemed to be arbitrarily determined, but they were determined by a comprehensive attempt and comparison.

For pre-treatment features, at most 7 features could be selected by ML methods, since Cox model selected at least as 7 features during the 12 times of LOOCV. Similarly, for delta2-features, at most 8 features could be selected by ML methods, since Cox model selected at least as 8 features during the 12 times of LOOCV. For delta1-features, the smallest number of feature selected by Cox regression model was 23, but the number of features imported into machine learning methods should be at most comparable to

the size of dataset to avoid more overfitting. Hence, up to 12 features were selected by ML for delta1-features.

The best performance each model combination was determined after all these attempts.

4.3 Feature Redundancy after Cox Regression Model

Cox regression model selected features with significant predictive relevance to overall survival. However, the features selected by univariate Cox regression model could be redundant. Features selected by Cox regression model for more than once and their Spearman's correlation can be seen below. The feature selected for only one time out of the 12 times of LOOCV was not considered as prominent.

4.3.1 Pre-treatment Features selected by Cox Regression Model

In figure 16, the horizontal axis represents the feature names, and the vertical axis represents the frequency of the pre-treatment features selected during the 12 times of LOOCV.

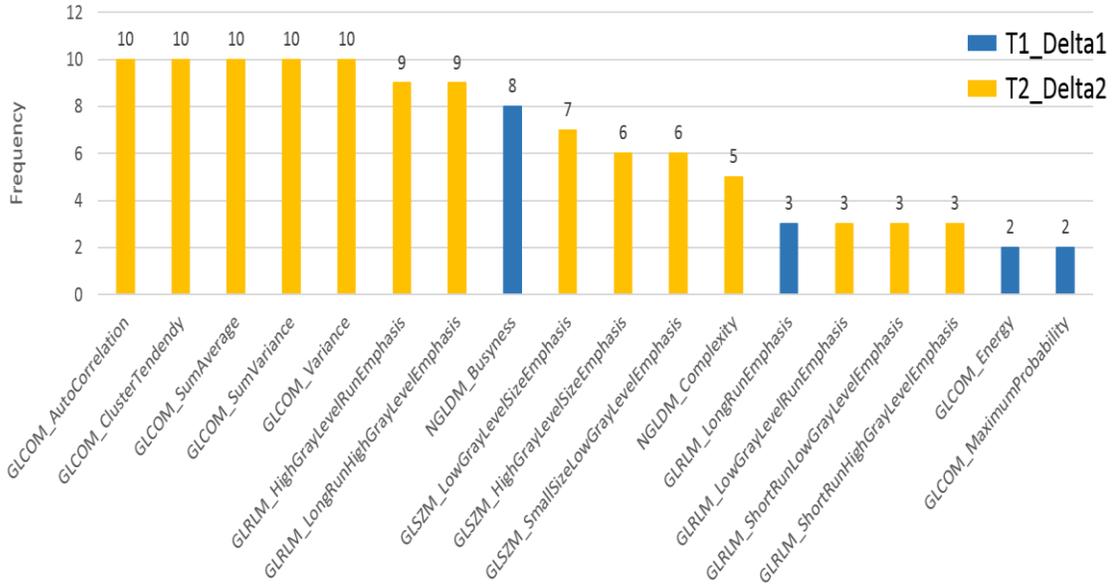


Figure 16: Pre-treatment features selected by Cox model for more than once.

4.3.2 Delta1-Features selected by Cox Regression Model

Delta1- feature selected by Cox regression model for more than once can be seen in figure 18. The proportion of each type of feature selected is illustrated in figure 17.

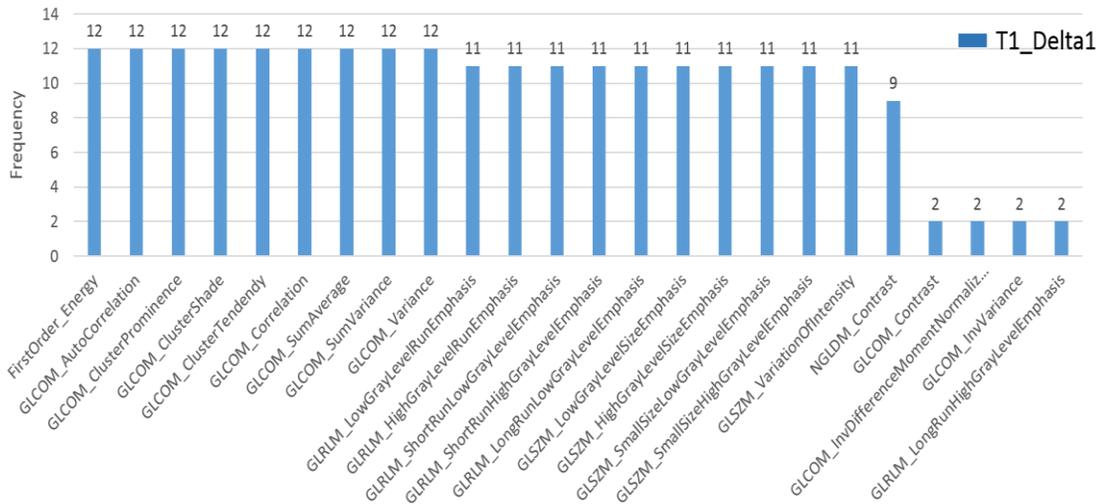


Figure 17: Delta1-features selected by Cox model for more than once.

4.3.3 Delta2-Features selected by Cox Regression Model

Delta2- feature selected by Cox regression model for more than once can be seen in figure 18.

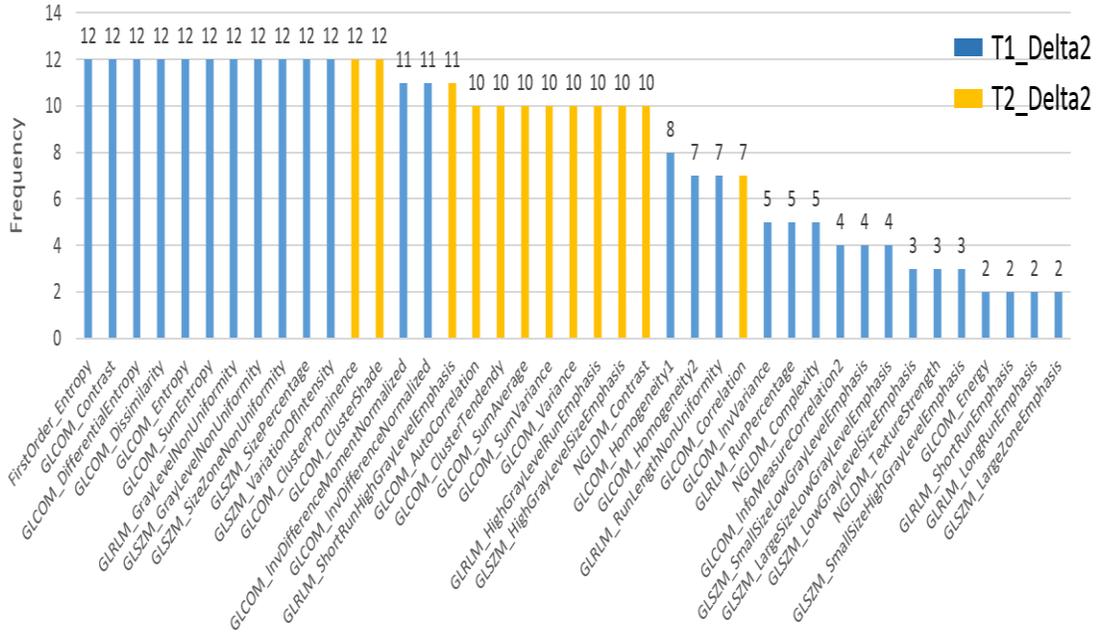


Figure 18: Delta2-features selected by Cox model for more than once.

Using Spearman's correlation, the correlations among the features selected by Cox regression model in each feature extraction are shown in figure below.

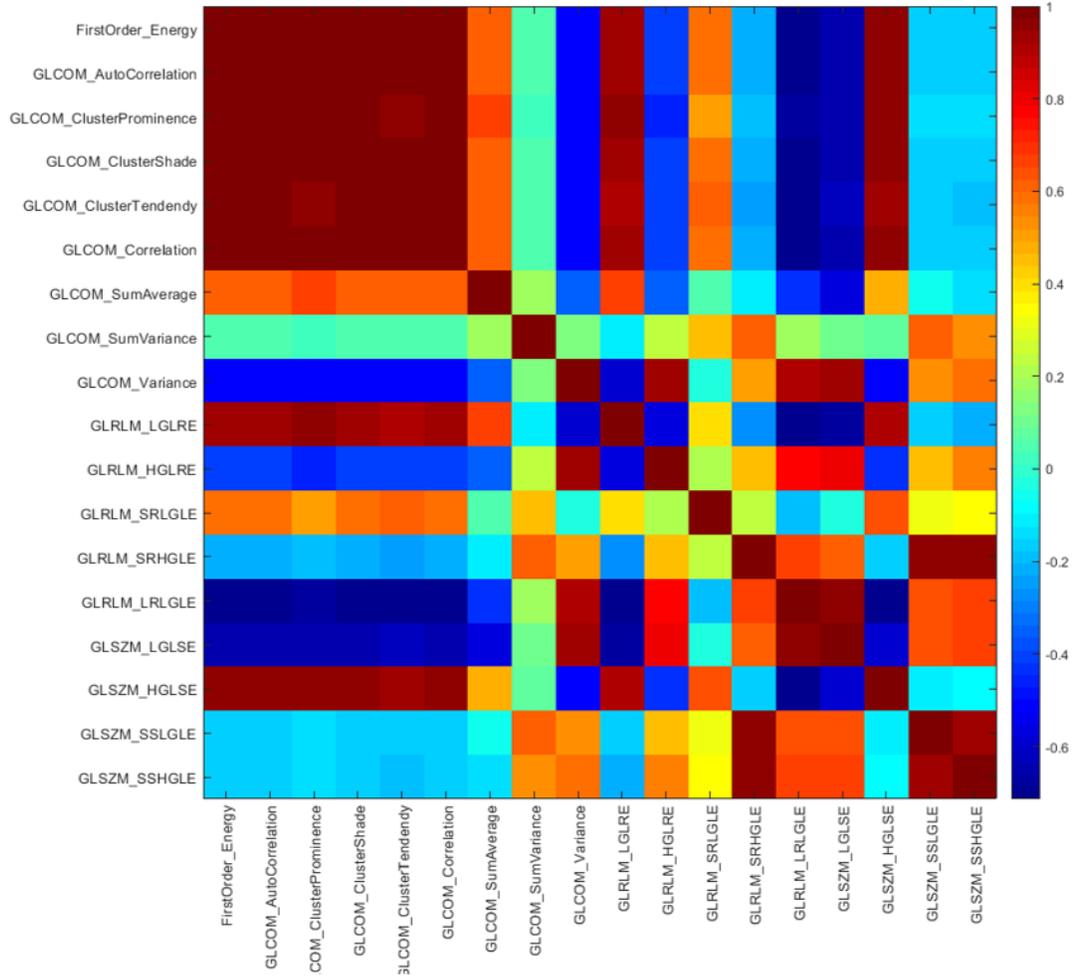


Figure 19: Spearman's correlations among the pre-treatment feature selected by Cox regression model.

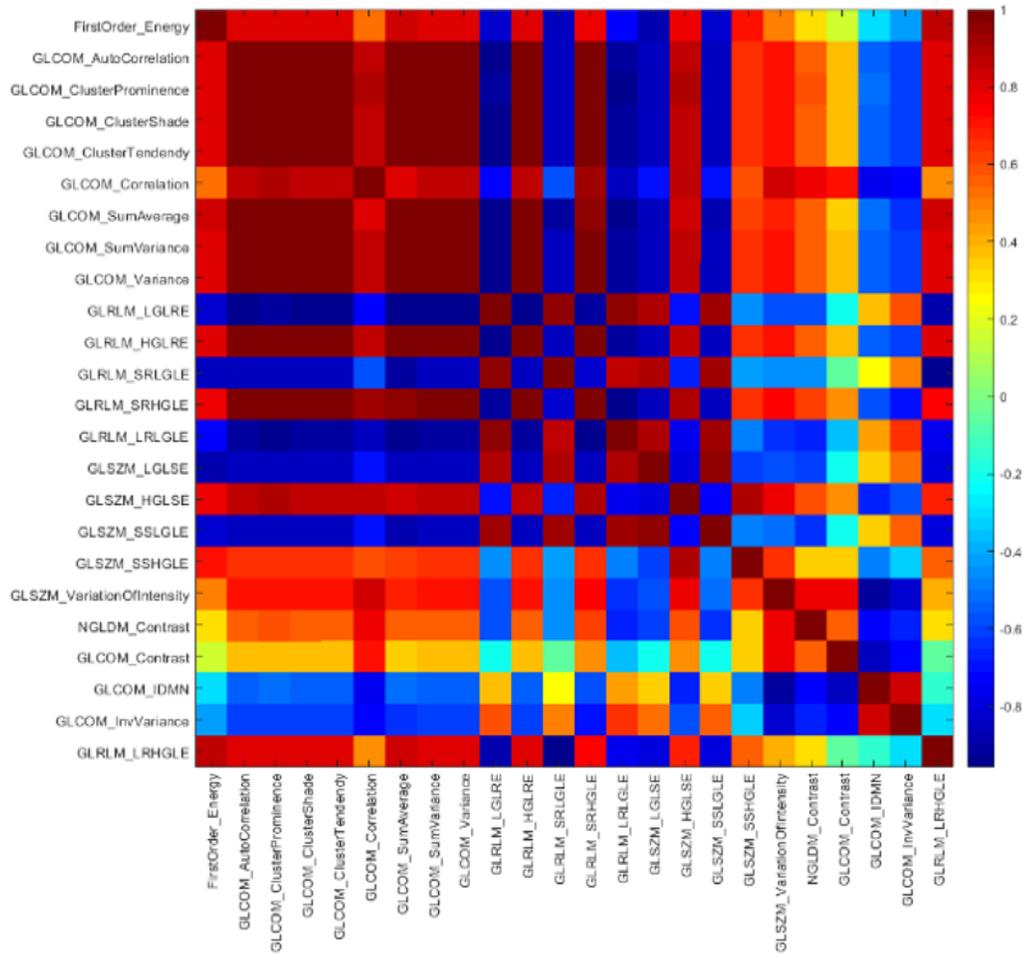


Figure 20: Spearman's correlations among the delta1- feature selected by Cox regression model.

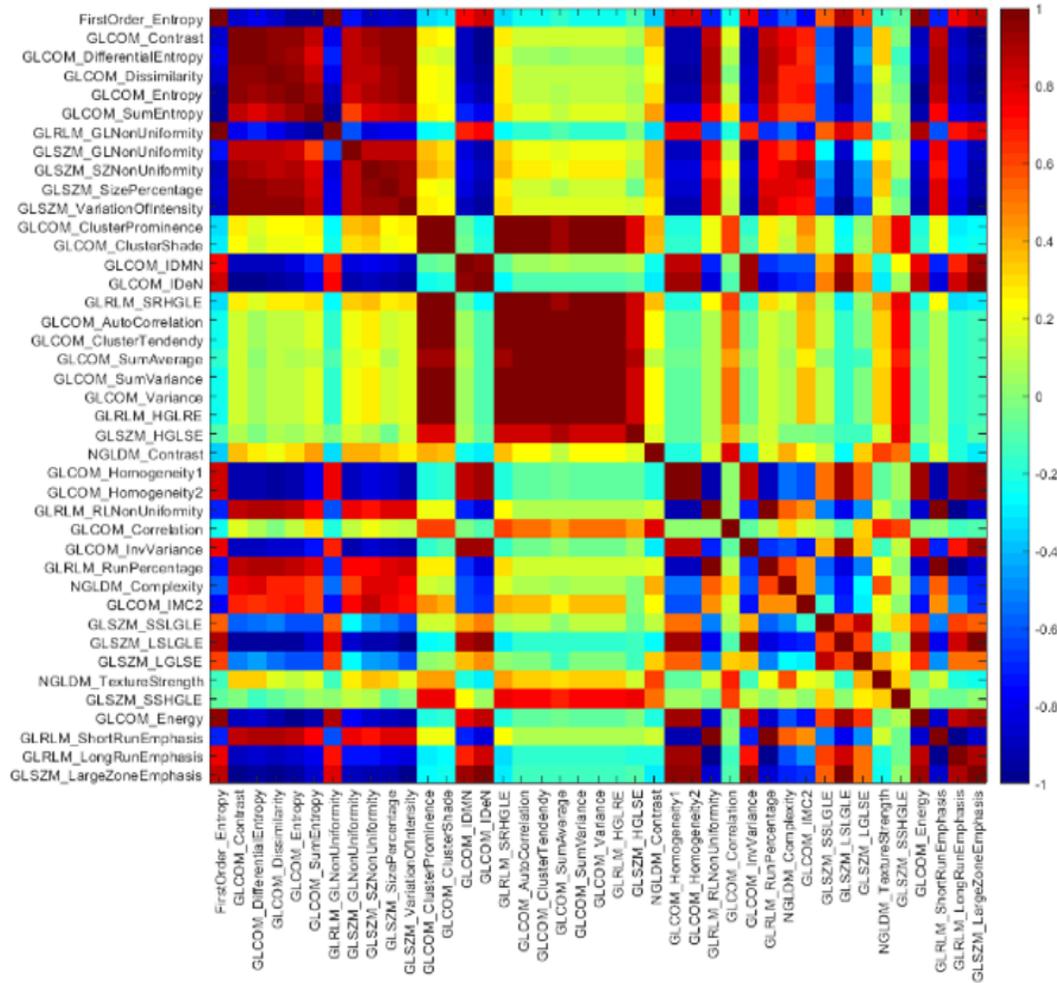


Figure 21: Spearman’s correlations among the delta2- feature selected by Cox regression model.

As we can see from figure 19-21, some of the features selected by Cox regression model are highly correlated based on Spearman’s correlation. Therefore, a feature clustering method could be developed to further reduce the feature redundancy before applying machine learning feature selection methods. Interestingly, higher feature redundancy not necessary leads to worse predictive performances. As we can see from figure 21, the delta1-features imported into machine learning feature selection models

were quite correlated. However, the predictive performance of delta1-features were better than delta2-features, which were not as high correlated from figure 22. Compared with feature redundancy, the feature values themselves could be more significant factors in machine learning algorithms.

4.4 Discussion on ROC Analysis

After feature extraction and classification, ROC analysis was used for evaluating the predictive performance of each model combination. Conventionally, ROC analysis can only be used for models with fixed parameters (weights). Nevertheless, the AUC values in this study were used to compare the predictive performance of different methods. Therefore, although the weights of machine learning methods adapted to different inputs, the predictive performances among different model combinations can still be compared by ROC analysis.

5. Conclusion and Future Work

5.1 Conclusion

This study used pre-treatment features and delta-features computed at two different time points to assess brain tumor radiosurgery by machine learning approaches. We computed the pre-treatment features, one-week delta-features and two-month delta-features from the GTV of T1-weighted and T2-weighted FLAIR MRI. During feature analysis, we used univariate Cox regression model and 3 ML methods for feature selection and 7 ML methods for classification. The model performances were evaluated by AUC analysis.

We found that one-week delta-features displayed higher predictive performance than both pre-treatment features and two-month delta-features for this cohort of patient data. Even though limited by the size of dataset, this study provides evidence that delta-features are potentially better in treatment assessment than pre-treatment features. The time point of computing the delta-features could also be a significant parameter regarding to predictive performance. In this study, change within a shorter duration after treatment (one-week post-treatment) provided better prediction than that of a longer one (two-month post-treatment). However, obtaining a generalized conclusion require a larger dataset to validate the results in this study.

With a univariate Cox regression model to select features ahead, L1-regularized logistic regression feature selection combined with random forest classifier and random forest feature selection combined with naïve bayes classifier provided the most acceptable predictions (AUC=0.944). This methodology could be promoted to a variety of tumor sites and imaging modalities.

5.2 Future Work

Based on this study, some future works are interested:

- (1) A larger dataset is warranted for validating this study and make a generalized conclusion.
- (2) Other imaging modalities, tumor sites, and clinical endpoints are warranted for the methodology in this research. For example, for brain tumor patients,

the toxicity after treatment could be a valuable clinical endpoint for physician's reference.

- (3) Further research on the influence of image segmentation on delta-radiomics is interested. In this study, the post-treatment contours were acquired from registering the post-treatment images to the pre-treatment images. In the future, we can research the influence on building models with registered contours and contours by re-segmentation.
- (4) For different diseases, the additional useful ROIs should be defined. While GTV is important, we could research on V12 for radiomics analysis.
- (5) A feature clustering method could be developed to reduce feature redundancy prior to machine learning for feature selection and improve model robustness.

References

- Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., . . . Rietveld, D. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5, 4006.
- Algothary, A., Viswanath, S., Shiradkar, R., Ghose, S., Pahwa, S., Moses, D., . . . Madabhushi, A. (2018). Radiomic features on MRI enable risk categorization of prostate cancer patients on active surveillance: Preliminary findings. *J Magn Reson Imaging*. doi:10.1002/jmri.25983
- Amadasun, M., & King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, 19(5), 1264-1274.
- Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah? *Acm Sigkdd Explorations Newsletter*, 2(2), 1-13.
- Bishop, C. M. (2006). Machine learning and pattern recognition. *Information Science and Statistics*. Springer, Heidelberg.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Cao, M., Liang, Y., Shen, C., Miller, K. D., & Stantz, K. M. (2009). Developing DCE-CT to quantify intra-tumor heterogeneity in breast tumors with differing angiogenic phenotype. *IEEE transactions on medical imaging*, 28(6), 861-871.
- Carvalho, S., Leijenaar, R. T. H., Troost, E. G. C., van Elmpt, W., Muratet, J. P., Denis, F., . . . Lambin, P. Early variation of FDG-PET radiomics features in NSCLC is related to overall survival - the "delta radiomics" concept. *Radiotherapy and Oncology*, 118, S20-S21. doi:10.1016/S0167-8140(16)30042-1
- Cunliffe, A., Armato, S. G., 3rd, Castillo, R., Pham, N., Guerrero, T., & Al-Hallaq, H. A. (2015). Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *Int J Radiat Oncol Biol Phys*, 91(5), 1048-1056. doi:10.1016/j.ijrobp.2014.11.030

- Fave, X., Zhang, L., Yang, J., Mackin, D., Balter, P., Gomez, D., . . . Liao, Z. (2017). Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Scientific Reports*, 7(1), 588.
- Fave, X., Zhang, L., Yang, J., Mackin, D., Balter, P., Gomez, D., . . . Court, L. (2017). Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Sci Rep*, 7(1), 588. doi:10.1038/s41598-017-00665-z
- Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3), 479.
- Galloway, M. (1975). Texture classification using gray level run length. *Comput. Graph. Image Process*, 4(2), 172-179.
- Gillies, R. J., Kinahan, P. E., & Hricak, H. (2015). Radiomics: images are more than pictures, they are data. *Radiology*, 278(2), 563-577.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.
- Haralick, R. M., & Shanmugam, K. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and Cybernetics*(6), 610-621.
- Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network *Neural networks for perception* (pp. 65-93): Elsevier.
- Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., . . . Aerts, H. J. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*, 48(4), 441-446. doi:10.1016/j.ejca.2011.11.036
- Leger, S., Zwanenburg, A., Pilz, K., Lohaus, F., Linge, A., Zophel, K., . . . Richter, C. (2017). A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci Rep*, 7(1), 13206. doi:10.1038/s41598-017-13448-3

- Liang, C., Huang, Y., He, L., Chen, X., Ma, Z., Dong, D., . . . Liu, Z. (2016). The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. *Oncotarget*, 7(21), 31401-31412. doi:10.18632/oncotarget.8919
- Lior, R. (2014). *Data mining with decision trees: theory and applications* (Vol. 81): World scientific.
- Longo, D. L. (2012). Tumor heterogeneity and personalized medicine. *N Engl J Med*, 366(10), 956-957.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Text classification and naive bayes. *Introduction to information retrieval*, 1, 6.
- Mattonen, S. A., Palma, D. A., Johnson, C., Louie, A. V., Landis, M., Rodrigues, G., . . . Ward, A. D. Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. *International Journal of Radiation Oncology • Biology • Physics*, 94(5), 1121-1128. doi:10.1016/j.ijrobp.2015.12.369
- Metz, C. E. (1978). *Basic principles of ROC analysis*. Paper presented at the Seminars in nuclear medicine.
- Ng, A. Y. (2004). *Feature selection, L 1 vs. L 2 regularization, and rotational invariance*. Paper presented at the Proceedings of the twenty-first international conference on Machine learning.
- Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition*, 26(9), 1277-1294. doi:[https://doi.org/10.1016/0031-3203\(93\)90135-I](https://doi.org/10.1016/0031-3203(93)90135-I)
- Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. (2015). Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep*, 5, 13087. doi:10.1038/srep13087

- Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M. M., Lambin, P., & Aerts, H. J. (2015). Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front Oncol*, 5, 272. doi:10.3389/fonc.2015.00272
- Rao, S. X., Lambregts, D. M., Schnerr, R. S., Beckers, R. C., Maas, M., Albarello, F., . . . Beets-Tan, R. G. (2016). CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy? *United European Gastroenterol J*, 4(2), 257-263. doi:10.1177/2050640615601603
- Rish, I. (2001). *An empirical study of the naive Bayes classifier*. Paper presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence.
- Saha, A., Harowicz, M. R., Wang, W., & Mazurowski, M. A. (2018). A study of association of Oncotype DX recurrence score with DCE-MRI characteristics using multivariate machine learning models. *J Cancer Res Clin Oncol*. doi:10.1007/s00432-018-2595-7
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*: Cambridge university press.
- Shapiro, L., & Stockman, G. (2000). Computer Vision Google Scholar.
- Sun, C., & Wee, W. G. (1983). Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, 23(3), 341-352.
- Thibault, G., Angulo, J., & Meyer, F. (2011). *Advanced statistical matrices for texture characterization: Application to dna chromatin and microtubule network classification*. Paper presented at the Image Processing (ICIP), 2011 18th IEEE International Conference on.
- Tu, S. J., Wang, C. W., Pan, K. T., Wu, Y. C., & Wu, C. T. (2018). Localized thin-section CT with radiomics feature extraction and machine learning to classify early-detected pulmonary nodules from lung cancer screening. *Phys Med Biol*. doi:10.1088/1361-6560/aaafab

- Udupa, J. K., LeBlanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., . . . Woodburn, J. (2006). A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 30(2), 75-87. doi:<https://doi.org/10.1016/j.compmedimag.2005.12.001>
- van Timmeren, J. E., Leijenaar, R. T. H., van Elmpt, W., Reymen, B., & Lambin, P. (2017). Feature selection methodology for longitudinal cone-beam CT radiomics. *Acta Oncologica*, 56(11), 1537-1543. doi:10.1080/0284186X.2017.1350285
- Wang, S., & Summers, R. M. (2012). Machine learning and radiology. *Medical image analysis*, 16(5), 933-951.
- Wu, W., Parmar, C., Grossmann, P., Quackenbush, J., Lambin, P., Bussink, J., . . . Aerts, H. J. (2016). Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Front Oncol*, 6, 71. doi:10.3389/fonc.2016.00071
- Yang, X., & Knopp, M. V. (2011). Quantifying tumor vascular heterogeneity with dynamic contrast-enhanced magnetic resonance imaging: a review. *BioMed Research International*, 2011.
- Zhang, B., He, X., Ouyang, F., Gu, D., Dong, Y., Zhang, L., . . . Zhang, S. (2017). Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett*, 403, 21-27. doi:10.1016/j.canlet.2017.06.004
- Zhang, Z., Yang, J., Ho, A., Jiang, W., Logan, J., Wang, X., . . . Li, J. (2017). A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. *Eur Radiol*. doi:10.1007/s00330-017-5154-8
- Zhu, X., Dong, D., Chen, Z., Fang, M., Zhang, L., Song, J., . . . Tian, J. (2018). Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. *Eur Radiol*. doi:10.1007/s00330-017-5221-1
- Zwanenburg, A., Leger, S., Vallières, M., & Löck, S. (2016). Image biomarker standardisation initiative-feature definitions. *arXiv preprint arXiv:1612.07003*.