European Radiology (2021) 31:8703–8713
https://doi.org/10.1007/s00330-021-07845-6

ONCOLOGY

# Radiomics-based differentiation between glioblastoma and primary central nervous system lymphoma: a comparison of diagnostic performance across different MRI sequences and machine learning techniques

Girish Bathla[1] · Sarv Priya[1] · Yanan Liu[2] · Caitlin Ward[3] · Nam H. Le[2] · Neetu Soni[1] · Ravishankar Pillenahalli Maheshwarappa[1] · Varun Monga[4] · Honghai Zhang[2] · Milan Sonka[2]

Received: 27 November 2020 / Revised: 3 February 2021 / Accepted: 26 February 2021 / Published online: 23 April 2021

## Abstract

**Objectives** Despite the robust diagnostic performance of MRI-based radiomic features for differentiating between glioblastoma (GBM) and primary central nervous system lymphoma (PCNSL) reported on prior studies, the best sequence or a combination of sequences and model performance across various machine learning pipelines remain undefined. Herein, we compare the diagnostic performance of multiple radiomics-based models to differentiate GBM from PCNSL.

**Methods** Our retrospective study included 94 patients (34 with PCNSL and 60 with GBM). Model performance was assessed using various MRI sequences across 45 possible model and feature selection combinations for nine different sequence permutations. Predictive performance was assessed using fivefold repeated cross-validation with five repeats. The best and worst performing models were compared to assess differences in performance.

**Results** The predictive performance, both using individual and a combination of sequences, was fairly robust across multiple top performing models (AUC: 0.961–0.977) but did show considerable variation between the best and worst performing models. The top performing individual sequences had comparable performance to multiparametric models. The best prediction model in our study used a combination of ADC, FLAIR, and T1-CE achieving the highest AUC of 0.977, while the second ranked model used T1-CE and ADC, achieving a cross-validated AUC of 0.975.

**Conclusion** Radiomics-based predictive accuracy can vary considerably, based on the model and feature selection methods as well as the combination of sequences used. Also, models derived from limited sequences show performance comparable to those derived from all five sequences.

**Key Points**

- *Radiomics-based diagnostic performance of various machine learning models for differentiating glioblastoma and PCNSL varies considerably.*
- *ML models using limited or multiple MRI sequences can provide comparable performance, based on the chosen model.*
- *Embedded feature selection models perform better than models using a priori feature reduction.*

**Keywords** Glioblastoma, lymphoma · Machine learning · Central nervous system neoplasms · Magnetic resonance imaging

**Abbreviations**

| | |
|---|---|
| ADC | Apparent diffusion coefficient |
| AFNI | Analysis of Functional NeuroImages |
| ANT | Advanced normalization tools |
| AUC | Area under the curve |
| DSC | Dynamic susceptibility contrast |
| DWI | Diffusion weighted imaging |

Girish Bathla and Sarv Priya have contributed equally to this work and are both co-first authors.

✉ Sarv Priya
sarv-priya@uiowa.edu

[1] Department of Radiology, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA

[2] College of Engineering, University of Iowa, Iowa City, IA, USA

[3] Department of Biostatistics, University of Iowa, Iowa City, IA, USA

[4] Department of Medicine, University of Iowa Hospitals and Clinics, Iowa City, IA, USA

| | |
|---|---|
| FLAIR | Fluid-attenuated inversion recovery |
| GBM | Glioblastoma |
| GBRM | Generalized boosted regression model |
| LASSO | Least absolute shrinkage and selection operator |
| LOGISMOS | Layered Optimal Graph Image Segmentation for Multiple Objects and Surfaces |
| ML | Machine learning |
| MLP | Multi-layer perceptron |
| MRI | Magnetic resonance imaging |
| PCA | Principal component analysis |
| PCNSL | Primary central nervous system lymphoma |
| r-CBV | Relative cerebral blood volume |
| ROC | Receiver operating characteristic |
| SIFT | Scale invariant feature transform |
| SVM | Support vector machine |
| T1-CE | T1 contrast enhanced |

## Introduction

Glioblastoma (GBM) and primary central nervous system lymphoma (PCNSL) together comprise the two most common primary malignant brain tumors [1]. Whereas GBM accounts for 14.6% of all brain neoplasms, PCNSL accounts for about 1.9% [2]. Even though the treatment strategies are vastly different, they both share overlapping clinical and imaging characteristics, which makes accurate pre-operative identification critical but challenging [3–5]. The utility of conventional and more advance imaging sequences (including diffusion and perfusion studies) has previously been assessed with modest success [6–10]. However, these may not be widely available.

More recently, a number of studies (Table 1) have attempted a radiomic-based differentiation between GBM and PCNSL with good success [1, 3, 5, 11–22]. A number of these were performed using machine learning (ML), which includes a wide variety of statistical analysis algorithms [23]. The success of a ML technique depends considerably on the amount, type and completeness of data, type of feature selection/reduction technique, and the problem to be addressed. As such, the predictive performance of various ML models for a specific problem can vary and remains largely unaddressed when differentiating GBM from PCNSL. This is, therefore, a need to compare the predictive performance of various models to determine the best performing models for this two-class problem. Similarly, given the heterogeneity of prior studies, it is important to determine if individual sequences or a combination of sequences have equivalent or superior performance when compared to all sequences combined. This will help guide the selection of best performing models for future studies and facilitate model selection for larger studies using multi-institutional datasets.

In this study, we compared the predictive performance of various ML techniques for differentiating between PCNSL and GBM using a combination of various feature selection and ML algorithms. The aims were to identify the best and worst performing models, as well as to determine if accurate distinction between these entities could be achieved using a single sequence or required a combination of sequences for best results.

## Methods

This is a single institution retrospective study, performed post approval of the local institutional review board. Patients were identified using a combination of institutional cancer registries and electronic medical records. Inclusion criteria were a pathologically proven diagnosis of GBM or PCNSL. Exclusion criteria included imaging studies with motion artifacts, absence of available index MRI scan, or absence of all required sequences (axial T1WI, T2WI, diffusion-weighted imaging [DWI], fluid-attenuated inversion recovery [FLAIR], and contrast-enhanced T1WI [T1-CE]). Additionally, patients were excluded where the image pre-processing (see below) or feature extraction was unsuccessful. This yielded a total of 94 patients, 34 with PCNSL and 60 with GBM, who were eventually assessed.

### Image pre-processing

Post de-identification of images, the DICOM images were converted to nifti format prior to pre-processing. Figure 1 provides an overview of the study workflow. The data were initially resampled to voxel size 1 × 1 × 5 $mm^3$ using the AFNI package (https://afni.nimh.nih.gov/) [24]. All sequences were registered to T1WI using Advanced Normalization Tools (ANTs) (http://stnava.github.io/ANTs/.) [25]. Following the resampling and co-registration, the image intensities were normalized to [0,255] using the feature scaling method available in the ANT registration suite.

Tumor segmentation was performed on axial T1W CE and FLAIR images by two radiologists (S.P., G.B.) in consensus using an in-house developed semi-automatic tool Layered Optimal Graph Image Segmentation for Multiple Objects and Surfaces (LOGISMOS) that first automatically identifies the tumor surfaces followed by an efficient “just-enough interaction” with an optional surface editing step which may be invoked if needed [26]. The T1-CE images were used to generate the masks for the enhancing disease (including internal necrosis where present). The FLAIR images were used to generate a mask for the entire lesion (tumor and surrounding edema). Figure 2 shows representative examples of ROI segmentation for both tumor types. The T1-CE-derived mask was subsequently subtracted from this mask to generate the mask

**Table 1** Summary of radiomics-based studies using MRI data to differentiate GBM from PCNSL. *NR* not reported. **Two gliomas were grade III while 69/71 were grade IV

| Author/year [Ref] | *n* (GBM/ PCNSL) | Sequence(s) used | Software/technique | Statistical/ML model | Conclusion |
|---|---|---|---|---|---|
| Yang/2017 [5] | 58/37 | T1W1, T2WI, FLAIR, T1-CE | NR | SVM | Accuracy of 96.84 |
| Alcaide-Leon/2017 [1] | 71**/35 | CE-T1WI | Matlab | SVM Classifier | AUC: 0.877 |
| Wu/2018 [19] | 70/32 | T2WI, CE-T1WI | Matlab | NR | Accuracy of 98.5%; AUC not reported |
| Kunimatsu/2018 [13] | 44/16 | CE-T1WI | ImageJ/Matlab | PCA for feature reduction. No ML | NR |
| Kim/2018 [12] | 78/65 | T2WI, DWI, CE-T1WI | 3D-Slicer/ANTs/Matlab | Logistic regression, SVM and RF | Validated AUC of 0.956 |
| Suh/2018 [17] | 23/54 | T2WI, FLAIR, CE-T1WI | PyRadiomics | RF | AUC of 0.921 |
| Kang/2018 [11] | 119/77 | ADC, CE-T1WI | MITK/Matlab | Eight ML models and twelve feature selection methods | Validated AUC: 0.944–0.984 |
| Nakagawa/2018 [16] | 45/25 | T2WI, ADC, Rcbv, CE-T1WI | LifeX/Python | Multivariate XGBoost | AUC of 0.980 |
| Xiao/2018 [21] | 60/22 | T1WI, CE-T1WI | ITK-SNAP/PyRadiomics/Python/Weka | Naïve Bayes | AUC:0.900 |
| Chen/2018 [3] | 66/30 | CE-T1WI | CNN-based segmentation | SVM | AUC: 0.991 |
| Kunimatsu/2019 [14] | 55/21 | CE-T1WI | ImageJ/Matlab | PCA. Four classification algorithms | AUC: 0.87–0.99 for training data. Only accuracy (75%) was reported for the test data |
| Yun/2019 [22] | 119/76 | DWI, CE-T1WI | MITK/Matlab | Supervised ML, MLP | AUC: 0.947–0.991 for validation cohort |
| Liu/2019 [15] | 107/60 | CE-T1WI | MRIcro/Matlab/ImageJ | Logistic regression. No ML | AUC of models based on fractal dimension and lacunarity: 0.776–0.895. AUC of model combining these with conventional imaging findings: 0.969 |
| Wang/2019 [18] | 81/28 | T2WI | ImageJ | Logistic regression. No ML | Combined conventional features and texture features had a combined AUC of 0.917 |
| Xia/2020 [20] | 129/111 | DWI/ADC, CE-T1WI, FLAIR | PyRadiomics, R. MedCalc | LASSO | Best radiomic model AUC: 0.935–0.943 using T1-CE and ADC features. |
| Current study | 60/34 | T1WI, T2WI, ADC, FLAIR, CE-T1WI | PyRadiomics, R | Four feature selection and twelve ML models | Mean cross-validated AUC of 0.977 when using T1CE, FLAIR, and ADC |

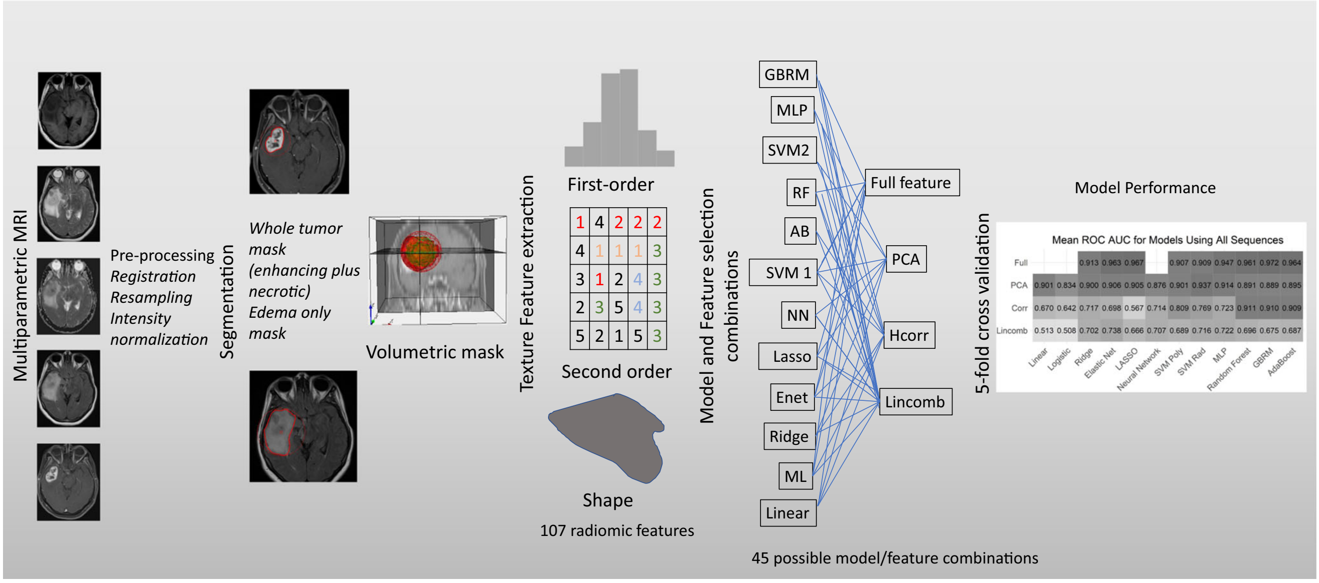


**Fig. 1** Overview of the current study workflow

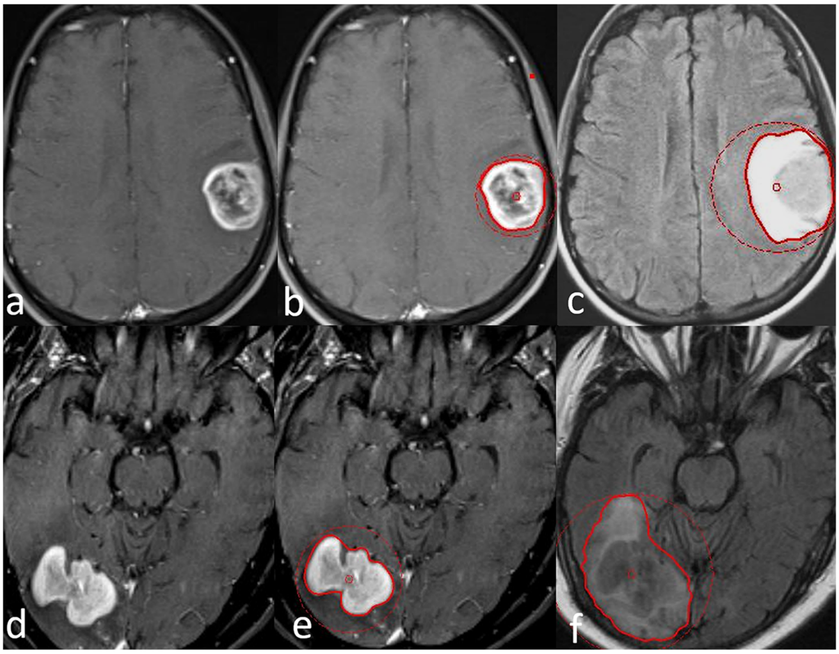
for the FLAIR signal abnormality surrounding the tumor. This way, two masks—one for the tumor and the other for the surrounding FLAIR signal abnormality—were expert-identified.

## Feature extraction

For each tumor, features were extracted using two masks, one for the tumor component (including enhancing and necrotic tumor) and the other for the surrounding FLAIR component. Features were extracted using PyRadiomics v3.0 [27]. Since there were ten possible mask and sequence combinations (five MRI sequences and two masks), on each of which 107 radiomic features were obtained, there were a total of 1070 features. For each sequence-specific model, the feature set included 214 (2 masks × 1 sequence × 107 features) radiomic features. Additionally, 3 limited sequence combinations were also evaluated and included: T1-CE/ADC, T1-CE/ADC/FLAIR, and ADC/FLAIR.

Each set of 107 features included 3D shape features ($n$ = 14), first-order features ($n$ = 18), gray level co-occurrence matrix features ($n$ = 24), gray level dependency matrix features ($n$ = 14), gray level run length matrix features ($n$ = 16), gray level size zone matrix features ($n$ = 16), and neighboring

**Fig. 2** Representative examples of the two tumor types (GBM: **a**–**c**; PCNSL: **d**–**f**) along with ROI segmentation for whole tumor and FLIAR signal abnormality. The edema mask was created through subtraction of the T1-CE mask from the FLAIR mask. The dotted red line surrounding the segmented volume (**b**, **c**, **e**, and **f**) represents the volume of interest as defined by the user



gray tone difference matrix features ($n = 5$). The default value for the number of bins was fixed by bin width of 25 gray levels. In rare cases where the edema was minimal, leading to absence of a corresponding mask, the value of the corresponding feature was set to − 9999.

## Feature selection

Due to the large size of the possible feature sets to be used relative to the sample size and highly correlated variables, feature selection is generally considered a critical piece of the model building process. Three feature selection methods were considered: a linear combination filter, a high correlation filter, and principal component analysis (PCA). The linear combination (lincomb) filter addresses both collinearity and dimension reduction. The high correlation (corr) filter removes variables which have a large absolute correlation. For the models using all sequences, the highest allowable correlation was set to 0.4 and for the models using each sequence separately, the threshold was set to 0.6. These thresholds were chosen to sufficiently reduce the dimensionality of the feature set for model fitting while retaining many of the important variables. The number of components retained in the PCA transformation was determined by specifying the fraction of the total variance that should be covered by the components. For the models using all sequences, this threshold was set to 80% and for the models using each sequence separately, the threshold was set to 85%, again with the goal of sufficiently retaining as much information as possible with enough dimension reduction to allow model fitting. Finally, models were also run using the entire feature set without any a priori feature reduction. These feature selection methods were implemented using the recipes package in R version 4.0.2 [28, 29]. Prior to any feature selection, all variables were standardized and missing values were imputed using mean imputation.

## Data analysis

Twelve different predictive models were fit to determine the best classifier for each feature set. These models can be categorized into three broad groups: linear classifiers, non-linear classifiers, and ensemble classifiers. The linear classifiers used were linear, logistic, ridge, elastic net, and LASSO regression. The non-linear classifiers used were neural network, support vector machine (SVM) with a polynomial kernel, SVM with a radial kernel, and multi-layer perceptron (MLP). Finally, the ensemble classifiers used were random forest, generalized boosted regression model (GBRM), and boosting of classification trees with AdaBoost.

Each model was fit using the three feature selection techniques as well as the entire feature set (full), except for the linear regression, logistic regression, and the neural network which cannot be fitted with the full feature set. This is because the model parameters cannot be uniquely estimated in linear and logistic regression models when the number of features is much larger than the sample size. For neural network, however, the problem is more of excessive computational requirement.

This yielded 45 possible model/feature selection combinations to be fit to each of the possible feature sets. These combinations were evaluated for individual MRI sequences ($n = 5$), a combination of sequences (T1-CE + ADC + FLAIR, T1-CE + ADC, and ADC+ FLAIR; $n = 3$), and all sequences combined ($n = 1$). Overall, a total of 405 different models were assessed. Predictive performance of each model was evaluated using fivefold repeated cross-validation with five repeats. For models with tuning parameters, important parameters were tuned using nested cross-validation to avoid bias. The feature selection techniques were carried out within each cross-validated split of the data, so as not to bias the estimate of predictive performance. Model fitting and cross-validated predictive performance were implemented using the MachineShop and RSNNS packages in R version 4.0.2 [29–31]. Predictive performance was measured with the area under the receiver operating characteristic curve (ROC AUC) for interpretability. As models were formulated to predict GBM, ROC AUC estimates the probability that a randomly selected subject that had GBM will have a greater predicted value than a randomly selected subject that had PCNSL. Higher ROC AUC values indicate better predictive performance.

To compare predictive performance between different model/feature selection techniques and between models fit to features from all sequences, the combination of sequences, and individual sequences, the corrected resampled $t$ test was used on the resampled ROC AUC values to properly account for the correlation arising from overlapping observations in the training and test sets from the repeated cross-validation procedure [32, 33]. To assess the variability in the performance of different model/feature selection techniques, model performance was compared between the top performing model and the top and bottom five models in each sequence category (Table 2; also Supplementary Table 3). All $p$ values were adjusted for multiple comparisons using the false discovery rate adjustment [34].

# Results

A total of 94 patients were studied, 34 with PCNSL (36.2%) and 60 with GBM (63.8%). Further details on patient demographics, scanner types, and typical scanning parameters at authors' institute are provided in Supplementary Table 1. The mean cross-validated ROC AUC for the best and worst five models, using all sequences, individual sequences, and a combination of sequences is provided in Table 2. Figure 3 shows the mean ROC AUC for all models across all feature

**Table 2** Comparison of AUC values for the highest and lowest five models when using all sequences, individual sequences, and a combination of sequences. The models are ranked based on mean and standard deviation (SD) of AUC. Models with significantly lower performance ($p < 0.05$) when compared to the best performing model have been marked with (*)

| Top performing models | | | | Lowest performing models | | | |
|---|---|---|---|---|---|---|---|
| All sequences combined | | | | | | | |
| Rank | Model | Feature selection | AUC mean (SD) | Rank | Model | Feature selection | AUC mean (SD) |
| 1 | GBRM | full | 0.972 (0.033) | 41 | LASSO | lincomp | *0.666 (0.087) |
| 2 | LASSO | full | 0.967 (0.032) | 42 | logistic | corr | *0.642 (0.147) |
| 3 | AdaBoost | full | 0.964 (0.032) | 43 | LASSO | corr | *0.567 (0.102) |
| 4 | Enet | full | 0.963 (0.024) | 44 | linear | lincomp | *0.513 (0.181) |
| 5 | RF | full | 0.961 (0.043) | 45 | logistic | lincomp | *0.508 (0.151) |
| T1WI only | | | | | | | |
| 1 | RF | corr | 0.973 (0.027) | 41 | SVM poly | lincomp | *0.813 (0.084) |
| 2 | LASSO | full | 0.964 (0.037) | 42 | linear | corr | *0.682 (0.133) |
| 3 | GBRM | corr | 0.961 (0.041) | 43 | logistic | corr | *0.639 (0.129) |
| 4 | Enet | full | 0.953 (0.042) | 44 | linear | lincomp | *0.591 (0.14) |
| 5 | AdaBoost | corr | 0.947 (0.050) | 45 | logistic | lincomp | *0.571 (0.134) |
| T2WI only | | | | | | | |
| 1 | GBRM | full | 0.959 (0.037) | 41 | logistic | pca | *0.83 (0.086) |
| 2 | RF | full | 0.953 (0.043) | 42 | linear | corr | *0.687 (0.173) |
| 3 | RF | corr | 0.948 (0.046) | 43 | logistic | corr | *0.687 (0.157) |
| 4 | GBRM | corr | 0.945 (0.047) | 44 | linear | lincomp | *0.59 (0.121) |
| 5 | SVM poly | full | 0.941 (0.049) | 45 | logistic | lincomp | *0.58 (0.119) |
| FLAIR only | | | | | | | |
| 1 | LASSO | full | 0.968 (0.036) | 41 | SVM poly | lincomp | *0.807 (0.107) |
| 2 | Enet | full | 0.963 (0.038) | 42 | linear | corr | *0.64 (0.127) |
| 3 | RF | full | 0.961 (0.037) | 43 | logistic | corr | *0.631 (0.132) |
| 4 | GBRM | full | 0.957 (0.044) | 44 | logistic | lincomp | *0.532 (0.144) |
| 5 | RF | corr | 0.938 (0.05) | 45 | linear | lincomp | *0.529 (0.153) |
| ADC only | | | | | | | |
| 1 | GBRM | full | 0.963 (0.031) | 41 | GBRM | pca | *0.845 (0.101) |
| 2 | RF | full | 0.963 (0.03) | 42 | linear | corr | *0.694 (0.127) |
| 3 | RF | corr | 0.961 (0.035) | 43 | logistic | corr | *0.677 (0.126) |
| 4 | AdaBoost | full | 0.96 (0.044) | 44 | linear | lincomp | *0.522 (0.148) |
| 5 | AdaBoost | corr | 0.955 (0.042) | 45 | logistic | lincomp | *0.506 (0.123) |
| T1-CE only | | | | | | | |
| 1 | AdaBoost | full | 0.968 (0.031) | 41 | GBRM | lincomp | *0.838 (0.08) |
| 2 | GBRM | full | 0.968 (0.037) | 42 | linear | corr | *0.724 (0.128) |
| 3 | RF | full | 0.967 (0.027) | 43 | logistic | corr | *0.689 (0.119) |
| 4 | AdaBoost | corr | 0.954 (0.052) | 44 | linear | lincomp | *0.527 (0.189) |
| 5 | RF | corr | 0.954 (0.04) | 45 | logistic | lincomp | *0.508 (0.177) |
| Sequence combinations: T1-CE + ADC + FLAIR | | | | | | | |
| 1 | AdaBoost | full | 0.977 (0.03) | 41 | LASSO | corr | *0.764 (0.143) |
| 2 | GBRM | full | 0.972 (0.032) | 42 | linear | corr | *0.705 (0.147) |
| 3 | RF | full | 0.971 (0.028) | 43 | logistic | corr | *0.677 (0.146) |
| 4 | LASSO | full | 0.963 (0.037) | 44 | linear | lincomp | *0.527 (0.189) |
| 5 | AdaBoost | corr | 0.96 (0.053) | 45 | logistic | lincomp | *0.508 (0.177) |
| T1-CE + ADC | | | | | | | |
| 1 | AdaBoost | full | 0.975 (0.028) | 41 | LASSO | corr | *0.833 (0.121) |
| 2 | RF | full | 0.975 (0.029) | 42 | linear | corr | *0.815 (0.113) |

**Table 2** (continued)

| Top performing models | | | | Lowest performing models | | | |
|---|---|---|---|---|---|---|---|
| All sequences combined | | | | | | | |
| Rank | Model | Feature selection | AUC mean (SD) | Rank | Model | Feature selection | AUC mean (SD) |
| 3 | GBRM | full | 0.972 (0.031) | 43 | logistic | corr | *0.757 (0.101) |
| 4 | AdaBoost | corr | 0.966 (0.039) | 44 | linear | lincomp | *0.527 (0.189) |
| 5 | GBRM | corr | 0.965 (0.044) | 45 | logistic | lincomp | *0.508 (0.177) |
| ADC + FLAIR | | | | | | | |
| 1 | GBRM | full | 0.971 (0.036) | 41 | LASSO | corr | *0.771 (0.151) |
| 2 | RF | full | 0.969 (0.036) | 42 | logistic | corr | *0.71 (0.103) |
| 3 | AdaBoost | full | 0.958 (0.048) | 43 | linear | corr | *0.686 (0.142) |
| 4 | AdaBoost | corr | 0.957 (0.058) | 44 | linear | lincomp | *0.522 (0.148) |
| 5 | LASSO | full | 0.953 (0.032) | 45 | logistic | lincomp | *0.506 (0.123) |

combinations, when built using data from all five MRI sequences.

In general, most of the high performing models, regardless of whether they were derived from all sequences, individual sequences, or a combination of limited sequences, were the ones using full feature set without any a priori feature reduction. On the other hand, most of the worst performing models used the linear combination feature reduction strategy. The model performance among the top 5 models was comparable across categories, and the corrected resampled *t* test with false discovery rate adjustment did not indicate any differences between the top models and the top performing model (AUC 0.977). The top performing model was however significantly better when compared against the bottom five models in each category. The adjusted *p* values and the overall performance of the 45 models when using all sequences, T1-CE, and T1-CE/FLAIR/ADC-derived models are provided in Supplementary Data (Supplementary Tables 3–6).

The best performing model used only a limited combination of sequences (T1-CE/FLAIR/ADC) and achieved an AUC of 0.977. This was comparable to the model using all five sequences as input, which had the best cross-validated AUC of 0.972 (*p* value of 0.790 when compared with best performing model). Similarly, the best performing models using only T1-CE/ADC (mean AUC of 0.975) and using only FLAIR/ADC (mean AUC of 0.971) also had comparable performance to the best performing model. Table 3 lists the overall top five models of the analysis, along with the performance metrics.

## Discussion

Our study suggests that the model performance for a radiomics-based differentiation between GBM and PCNSL can vary significantly, based on the chosen model, although several best predictive models have comparable performance.

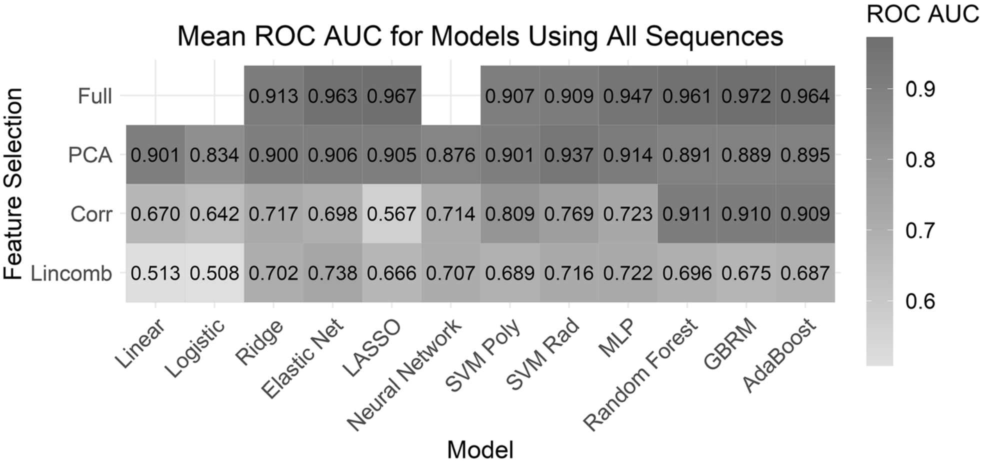


**Fig. 3** Comparison of mean cross-validated AUC for various model-feature combinations when using data from all five MRI sequences

**Table 3** Top five model-feature combinations with observed and cross-validated AUC values and cross-validated performance metrics

| Rank | Sequence(s) | Model | Feature selection | Observed AUC | CV AUC | CV Brier | CV accuracy | CV sensitivity | CV specificity |
|---|---|---|---|---|---|---|---|---|---|
| 1 | CE + A + F | AdaBoost | full | 1.000 | 0.977 | 0.082 | 0.934 | 0.970 | 0.871 |
| 2 | CE + A | AdaBoost | full | 1.000 | 0.975 | 0.090 | 0.909 | 0.947 | 0.844 |
| 3 | CE + A | RF | full | 1.000 | 0.975 | 0.068 | 0.917 | 0.957 | 0.849 |
| 4 | CE + A + F | GBRM | full | 1.000 | 0.972 | 0.054 | 0.940 | 0.977 | 0.877 |
| 5 | CE + A | GBRM | full | 1.000 | 0.972 | 0.074 | 0.909 | 0.940 | 0.855 |

Interestingly, the best performing models were those that performed embedded feature selection fit to the full texture feature set, as opposed to models using any of the a priori feature reduction strategies. Another important observation is that models derived from a limited combination of sequences may perform as well as the models using data from all five sequences. These observations are relevant not only in terms of highlighting the variability across machine learning models for the same problem but also in terms of evaluating the necessity of a priori feature reduction which is almost routinely performed. Finally, the excellent performance of a limited sequence combination may allow for a less computationally intensive workflow that may be easier to integrate in clinical practice.

In the current study, of the three feature selection/reduction strategies, models using PCA overall performed better while the lincomb-based models performed the worst. None of these however performed as well as models using full feature set with embedded feature selection. It is pertinent to note here that PCA leads to the fewest number of features included in the models and the linear combination filter leads to the largest. As the number of tumors in the data is small relative to the number of features, it makes sense that feature selection methods leading to smaller feature sets would perform better, particularly for those models which do not do any embedded feature selection, such as the linear, logistic, neural network, and SVM models. However, any feature selection filter would result in some loss of information. The impact of loss of information is evident when evaluating performance of embedded feature selection methods of the elastic net, LASSO, random forest, AdaBoost, and GBRM models which show poor predictive performance with feature reduction (AUC: 0.666–0.738) and considerably improved performance when using full feature set (AUC: 0.961–0.972). Of note, both the ensemble models and penalized regression models like GBRM or LASSO do perform feature selection. However, the feature selection in such cases is part of the model fitting process, unlike other models which require a priori feature reduction using some form of feature selection strategy.

In the analysis using all sequences, the GBRM classifier had an observed ROC AUC of 1.000 and mean cross-validated ROC AUC of 0.972, while the LASSO classifier had an observed ROC AUC of 1.000 and mean cross-validated ROC AUC of 0.967. The absence of a compelling drop in performance and the use of nested cross-validation for tuning parameters would argue against over-fitting of models in our analysis. The low Brier scores of multiple top performing models also attest to the robustness of results in our case. Similarly, the models using a combination of sequences also had excellent cross-validated performance, when using full features without any specified feature reduction strategy. This is an interesting observation since most of the prior studies have consistently used some form of feature reduction strategy, varying between principal component analysis, recursive random forest, and minimum redundancy, maximum relevance [14, 17, 22]. Our findings suggest that feature reduction, though meant to reduce redundant and highly correlated features, may not always be ideal and can negatively impact model performance. Embedded feature selection in the model, on the other hand, may perform better.

Another important observation from our study is that information derived from a limited combination of sequences may suffice to differentiate GBM from PCNSL. In fact, none of the overall top five performing models used the full feature set (Table 3). It is pertinent to note here that all five models, at the very least, consistently used post-contrast imaging and ADC map–derived texture features. This is relevant since it could help reduce the required computational effort and time. We found the model using T1-CE, ADC, and FLAIR to have the highest mean cross-validated AUC of 0.977, though other combinations (all sequences, T1-CE + ADC, ADC + FLAIR) also had comparable performance. In this regard, the model performance using only the ADC and FLAIR images is worth noting since it implies that excellent accuracy could be obtained without contrast administration, a finding that may be useful for patients who are unable to get a contrast-enhanced study. Almost all prior studies dealing with the same two-class problem, except for Wang et al, have used a contrast-enhanced sequence for analysis [18]. The top radiomics features and their relative importance for the best overall performing model are provided in Supplementary Table 2. Most of the top-ranking features were second-order features and derived from the T1-CE images. A combination of sequences, in general, did better than individual sequences alone, with the maximal AUC achieved for T1-CE and FLAIR, both with mean cross-validated AUC of 0.968.

Overall, multiple machine learning–based models in our study achieved excellent accuracy which was better than multiple previously reported studies where the AUC varied between 0.877 and 0.956, and comparable to other studies by Yun et al, Nakagawa et al, and Chen et al (see Table 1). Of the multiple previously reported studies, only Yun et al reported the model performance across three different models, and using three feature reduction techniques. [22]. Their best performing machine learning model used generalized linear model boosting with backward feature elimination as a feature reduction technique, achieving an AUC of 0.943 which is inferior to the best performing model in the current study (AUC of 0.977). However, the best overall performing model in their study was MLP (AUC of 0.991). In our analysis, even though MLP-based models showed high performance (mean AUC: 0.913–0.933), they were not among the top 5 performing models in any category. This could possibly be secondary to the imaging data used. For example, all MRI studies in their cohort were done on a 3-T magnet with a slice thickness of 1 mm for the post-contrast images. The magnet strength and image resolution are known to affect texture features and could account for some of these differences.

Chen et al also achieved excellent results (AUC of 0.982–0.991) for the same two-class problem using features derived from T1-CE images [3]. However, unlike general radiomic features, which are derived from image features such as intensity, shape, wavelet, or texture parameters, they extracted radiomic features using scale invariant feature transform (SIFT), which may explain the slight variability in results compared to our current study. Nakagawa et al were also able to achieve a similar model performance (AUC of 0.980) in their study using eXtreme gradient boosting (XGBoost), and a combination of T1-CE, T2WI, ADC, and rCBV maps derived from dynamic susceptibility contrast (DSC)–enhanced perfusion MRI [16]. However, DSC imaging may not be widely available. Our results, which use conventional and routinely available sequences, are more pragmatic.

Limitations of our study include its retrospective nature and a relatively small sample size. Given this limitation, it is quite possible that the model performance may vary with additional training data. Another limitation would be the absence of external validation cohort which would help determine the generalizability of our findings. Unlike GBM, there are no freely available data repositories for PCNSL patients to our knowledge which could be readily used. We also did not compare our results to expert human readers. However, multiple prior studies have compared machine learning models with human readers and noted the AUC of the expert human readers to vary between 0.79 and 0.94, indicating that the top performing machine learning models would likely outperform expert human readers [11, 16, 20]. Future studies should evaluate the combined accuracy of human readers and best performing machine learning models. Finally, despite the number of different models that were evaluated, we did not assess deep neural networks since this technique is more computationally demanding and data hungry. Nevertheless, our study evaluates the performance of 45 different feature machine learning techniques and provides additional insights into factors that may determine final model performance. Other relative strengths of the study include a documented imaging protocol, use of feature selection techniques, discrimination and calibration statistics, pathological confirmation in all cases, nested cross-validation, and potential clinical utility, all of this contribute to the radiomic quality score as proposed previously [35].

## Conclusion

Our findings suggest that excellent discrimination between PCNSL and GBM can be achieved through machine learning. Additionally, features derived from a limited number of sequences may perform as well as features derived from multiple sequences. Another pertinent observation is that models without any a priori feature reduction strategy generally outperform models derived post-feature reduction. Finally, the model performance varies considerably, based on the feature selection technique and the model used. Future studies should focus on a more extensive evaluation of different machine learning models to determine the best possible combination of feature selection and learning model for a more optimized model performance.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-07845-6

**Funding** The authors state that this work has not received any funding.

## Declarations

**Guarantor** The scientific guarantor of this publication is Girish Bathla.

**Conflict of interest** Girish Bathla has a research grant from Siemens AG, Forchheim, Germany, as well as American Cancer Society, that is unrelated to the submitted work. The rest of the authors report no relationships that could be construed as a conflict of interest.

**Statistics and biometry** Caitlin ward (co-author) kindly provided statistical advice for this manuscript.

One of the authors has significant statistical expertise.

**Informed consent** Written informed consent was waived by the Institutional Review Board.

**Ethical approval** Institutional Review Board approval was obtained.

**Methodology**
- retrospective
- diagnostic or prognostic study
- performed at one institution

## References

1. Alcaide-Leon P, Dufort P, Geraldo AF et al (2017) Differentiation of enhancing glioma and primary central nervous system lymphoma by texture-based machine learning. AJNR Am J Neuroradiol 38(6):1145–1150
2. Ostrom QT, Cioffi G, Gittleman H et al (2019) CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012-2016. Neuro Oncol 21(Suppl 5):v1–v100
3. Chen Y, Li Z, Wu G et al (2018) Primary central nervous system lymphoma and glioblastoma differentiation based on conventional magnetic resonance imaging by high-throughput SIFT features. Int J Neurosci 128(7):608–618
4. Bathla G, Hegde A (2016) Lymphomatous involvement of the central nervous system. Clin Radiol 71(6):602–609
5. Yang Z, Feng P, Wen T, Wan M, Hong X (2017) Differentiation of glioblastoma and lymphoma using feature extraction and support vector machine. CNS Neurol Disord Drug Targets 16(2):160–168
6. Choi YS, Lee HJ, Ahn SS et al (2017) Primary central nervous system lymphoma and atypical glioblastoma: differentiation using the initial area under the curve derived from dynamic contrast-enhanced MR and the apparent diffusion coefficient. Eur Radiol 27(4):1344–1351
7. Kickingereder P, Sahm F, Wiestler B et al (2014) Evaluation of microvascular permeability with dynamic contrast-enhanced MRI for the differentiation of primary CNS lymphoma and glioblastoma: radiologic-pathologic correlation. AJNR Am J Neuroradiol 35(8): 1503–1508
8. Kickingereder P, Wiestler B, Sahm F et al (2014) Primary central nervous system lymphoma and atypical glioblastoma: multiparametric differentiation by using diffusion-, perfusion-, and susceptibility-weighted MR imaging. Radiology 272(3):843–850
9. Nakajima S, Okada T, Yamamoto A et al (2015) Differentiation between primary central nervous system lymphoma and glioblastoma: a comparative study of parameters derived from dynamic susceptibility contrast-enhanced perfusion-weighted MRI. Clin Radiol 70(12):1393–1399
10. Toh CH, Castillo M, Wong AM et al (2008) Primary cerebral lymphoma and glioblastoma multiforme: differences in diffusion characteristics evaluated with diffusion tensor imaging. AJNR Am J Neuroradiol 29(3):471–475
11. Kang D, Park JE, Kim YH et al (2018) Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation. Neuro Oncol 20(9):1251–1261
12. Kim Y, Cho HH, Kim ST, Park H, Nam D, Kong DS (2018) Radiomics features to distinguish glioblastoma from primary central nervous system lymphoma on multi-parametric MRI. Neuroradiology 60(12):1297–1305
13. Kunimatsu A, Kunimatsu N, Kamiya K, Watadani T, Mori H, Abe O (2018) Comparison between glioblastoma and primary central nervous system lymphoma using MR image-based texture analysis. Magn Reson Med Sci 17(1):50–57
14. Kunimatsu A, Kunimatsu N, Yasaka K et al (2019) Machine learning-based texture analysis of contrast-enhanced MR imaging to differentiate between glioblastoma and primary central nervous system lymphoma. Magn Reson Med Sci 18(1):44–52
15. Liu S, Fan X, Zhang C et al (2019) MR imaging based fractal analysis for differentiating primary CNS lymphoma and glioblastoma. Eur Radiol 29(3):1348–1354
16. Nakagawa M, Nakaura T, Namimoto T et al (2018) Machine learning based on multi-parametric magnetic resonance imaging to differentiate glioblastoma multiforme from primary cerebral nervous system lymphoma. Eur J Radiol 108:147–154
17. Suh HB, Choi YS, Bae S et al (2018) Primary central nervous system lymphoma and atypical glioblastoma: differentiation using radiomics approach. Eur Radiol 28(9):3832–3839
18. Wang BT, Liu MX, Chen ZY (2019) Differential diagnostic value of texture feature analysis of magnetic resonance T2 weighted imaging between glioblastoma and primary central neural system lymphoma. Chin Med Sci J 34(1):10–17
19. Wu G, Chen Y, Wang Y et al (2018) Sparse representation-based radiomics for the diagnosis of brain tumors. IEEE Trans Med Imaging 37(4):893–905
20. Xia W, Hu B, Li H et al (2020) Multiparametric-MRI-based radiomics model for differentiating primary central nervous system lymphoma from glioblastoma: development and cross-vendor validation. J Magn Reson Imaging 53(1):242–50
21. Xiao DD, Yan PF, Wang YX, Osman MS, Zhao HY (2018) Glioblastoma and primary central nervous system lymphoma: preoperative differentiation by using MRI-based 3D texture analysis. Clin Neurol Neurosurg 173:84–90
22. Yun J, Park JE, Lee H, Ham S, Kim N, Kim HS (2019) Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma. Sci Rep 9(1):5746
23. Wang H, Zhou Y, Li L, Hou W, Ma X, Tian R (2020) Current status and quality of radiomics studies in lymphoma: a systematic review. Eur Radiol 30:6228–40
24. Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29(3):162–173
25. Avants BB, Tustison N, Song G (2009) Advanced normalization tools (ANTS). Insight J 2(365):1–35
26. Zhang H, Lee K, Chen Z, Kashyap S, Sonka M (2020) LOGISMOS-JEI: Segmentation using optimal graph search and just-enough interaction. In: Handbook of medical image computing and computer assisted intervention. Academic Press, Cambridge, p 249–272
27. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77(21):e104–e107
28. Kuhn M, Wickham H (2018) RStudio. recipes: Preprocessing Tools to Create Design Matrices
29. Team RC (2013) R: a language and environment for statistical computing
30. Smith BJ (2020) MachineShop: machine learning models and tools. R package version 2.5.0
31. Bergmeir CN, Benítez Sánchez JM (2012) Neural networks in R using the Stuttgart neural network simulator: RSNNS. American Statistical Association

32. Bouckaert RR, Frank E (2004) Evaluating the replicability of significance tests for comparing learning algorithms. In: Dai H, Srikant R, Zhang C (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2004. Lecture Notes in Computer Science, vol 3056. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24775-3_3
33. Nadeau C, Bengio Y (2003) Inference for the generalization error. Mach Learn 52(3):239–281
34. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol 57(1):289–300
35. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14(12):749–762