



OPEN

Machine learning based differentiation of glioblastoma from brain metastasis using MRI derived radiomics

Sarv Priya¹✉, Yanan Liu², Caitlin Ward³, Nam H. Le², Neetu Soni¹, Ravishankar Pillenahalli Maheshwarappa¹, Varun Monga⁴, Honghai Zhang², Milan Sonka² & Girish Bathla¹

Few studies have addressed radiomics based differentiation of Glioblastoma (GBM) and intracranial metastatic disease (IMD). However, the effect of different tumor masks, comparison of single versus multiparametric MRI (mp-MRI) or select combination of sequences remains undefined. We cross-compared multiple radiomics based machine learning (ML) models using mp-MRI to determine optimized configurations. Our retrospective study included 60 GBM and 60 IMD patients. Forty-five combinations of ML models and feature reduction strategies were assessed for features extracted from whole tumor and edema masks using mp-MRI [T1W, T2W, T1-contrast enhanced (T1-CE), ADC, FLAIR], individual MRI sequences and combined T1-CE and FLAIR sequences. Model performance was assessed using receiver operating characteristic curve. For mp-MRI, the best model was LASSO model fit using full feature set (AUC 0.953). FLAIR was the best individual sequence (LASSO-full feature set, AUC 0.951). For combined T1-CE/FLAIR sequence, adaBoost-full feature set was the best performer (AUC 0.951). No significant difference was seen between top models across all scenarios, including models using FLAIR only, mp-MRI and combined T1-CE/FLAIR sequence. Top features were extracted from both the whole tumor and edema masks. Shape sphericity is an important discriminating feature.

Glioblastoma (GBM) and intracranial metastatic disease (IMD) together constitute the vast majority of malignant brain neoplasms^{1,2}. Gliomas account for about 25.5% of all primary brain and other CNS tumors and approximately 80.8% of primary malignant brain tumors. Of these, GBM is the most common, accounting for over half of the gliomas (57.3%) with an annual age-adjusted incidence rate of 3.22 per 100,000 population in the United States³. IMD on the other hand has an incidence rate of approximately 10 per 100,000 population and are more common than GBM¹.

The distinction between GBM and IMD is important since it has diagnostic, therapeutic and prognostic implications^{2,4,5}. Histopathological tissue confirmation is considered the gold standard for diagnosis, but is not always optimal, with misdiagnosis and under grading of tumors reported in 9.2 and 28% of lesions respectively⁶. The reported biopsy complication rate varies between 6 and 12% with mortality rate of 0–1.7%⁷.

On conventional imaging, factors such as multiplicity of lesions, morphology, cerebellar localization and known history of underlying primary cancer can be helpful to differentiate IMD from GBM^{1,2}. However, brain metastases may present as a solitary lesion in approximately half of the patients or be associated with undiagnosed systemic malignancy in about 15–30%^{8,9}. Thus, conventional imaging alone may be insufficient for accurate classification. Prior studies using advanced MRI imaging techniques such as perfusion imaging^{10–12}, spectroscopy, diffusion-weighted and tensor imaging¹³, new diffusion weighted techniques like neurite orientation dispersion and density imaging¹⁴, and more recently other advanced sequences like non-contrast inflow-based vascular-space-occupancy MR imaging¹⁵ have been used to distinguish amongst these entities with variable success^{16–20}. However, these advanced imaging sequences are not performed universally, and conventional imaging is still the mainstay in clinical practice. Radiomics is a technique applied on medical images to extract quantitative

¹Department of Radiology, University of Iowa Hospital and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA. ²College of Engineering, University of Iowa, Iowa City, IA, USA. ³Department of Biostatistics, University of Iowa, Iowa City, IA, USA. ⁴Department of Medicine, University of Iowa Hospitals and Clinics, Iowa City, IA, USA. ✉email: sarv-priya@uiowa.edu

	GBM	METASTASES
Patients (120)	60	60; Breast (20); Lung (40)
Age years (mean \pm SD)	62 \pm 11	62 \pm 10
Gender		
Male	36	27
Female	24	33
Localization		
Supratentorial	58	Breast (10); Lung (25)
Infratentorial	2	Breast (6); Lung (8)
Both	0	Breast (4); Lung (7)
Multiplicity		
Single	53	Breast (12); Lung (24)
Two	5	Breast (2); Lung (7)
\geq Two (Multiple)	2	Breast (6); Lung (9)
Necrosis		
Yes	59	Breast (12); Lung (24)
No	1	Breast (8); Lung (16)

Table 1. Patient demographics and tumor characteristics. GBM-Glioblastoma.

features invisible to human eye²¹. These features may provide a complimentary tool for the expert human reader. These radiomic features have been employed in multiple prior studies for tumor grading, classification and prognosis^{21–24}. The advantage of radiomics is that it can be applied to routinely acquired conventional clinical images²⁵.

The application of radiomics based machine learning techniques (MLT) to differentiate GBM from IMD has only been explored in a few prior studies, mostly using limited MRI sequences and MLT^{1,2,4,26–29}. The superiority of having one, a few, or all conventional MRI sequences (T1 WI, T2 WI, ADC, FLAIR and T1-CE) as well as the impact of feature reduction and type of machine learning models remain largely unexplored. In this study, we aimed to determine the optimal radiomics based MLT for this specific two-class problem using routinely available conventional MRI sequences.

Results

Patient characteristics. There were 120 patients (males 63, females 57) in the study population (GBM 60, metastases 60). The majority of metastatic tumors were from lung cancer (40) followed by breast cancer (20). The demographic and tumor characteristics are provided in Table 1.

Model performance on mp-MRI. Using mp-MRI, the two best performing models were the LASSO (least absolute shrinkage and selection operator) and elastic net fit to the full feature set. The LASSO classifier had mean cross-validated area under the curve (AUC) of 0.953 and the elastic net classifier had a mean cross-validated AUC of 0.952. Figure 1 displays the mean cross-validated AUC for all 45 MLT combinations fit using all sequences.

Model performance on individual sequences. For the models fit to each sequence separately, the LASSO and elastic net fit to the full feature set again were the top performing models, with both being fit to the FLAIR sequence. Interestingly, seven of the top 10 best performing sequence-specific models were derived from the FLAIR sequence. The LASSO classifier on the FLAIR sequence had mean cross-validated AUC of 0.951 while the elastic net classifier had a mean cross-validated AUC of 0.948. Figure 2 shows the mean AUC for all models fit using the FLAIR sequence as many of the top performing individual sequence models came from this sequence. Table 2 displays the mean and standard deviation of AUC for the 10 best performing models for mp-MRI and individual sequences.

Model performance from combined T1-CE and FLAIR sequences. For the models fit to the T1-CE and FLAIR sequences in combination, the adaBoost and LASSO models fit to the full feature set were the top performing models. The adaBoost classifier had mean cross-validated AUC of 0.951. The LASSO classifier had mean cross-validated AUC of 0.950. Figure 3 shows the mean AUC for all models fit using the combined T1-CE and FLAIR sequences. Table 3 displays the mean and standard deviation of AUC for the 10 best performing models for T1-CE and FLAIR combination.

Comparison of predictive performance between mp-MRI, individual sequence, and combination of T1-CE with FLAIR. Overall, the best performing model using mp-MRI (LASSO fit to the full feature set), FLAIR sequence (LASSO full) and combined T1-CE and FLAIR sequences (adaBoost full) had similar predictive performance (p-value > 0.05 for all) (Table 4). These results indicate no statistically significant differences in predictive performance between the top models in each of the three scenarios.

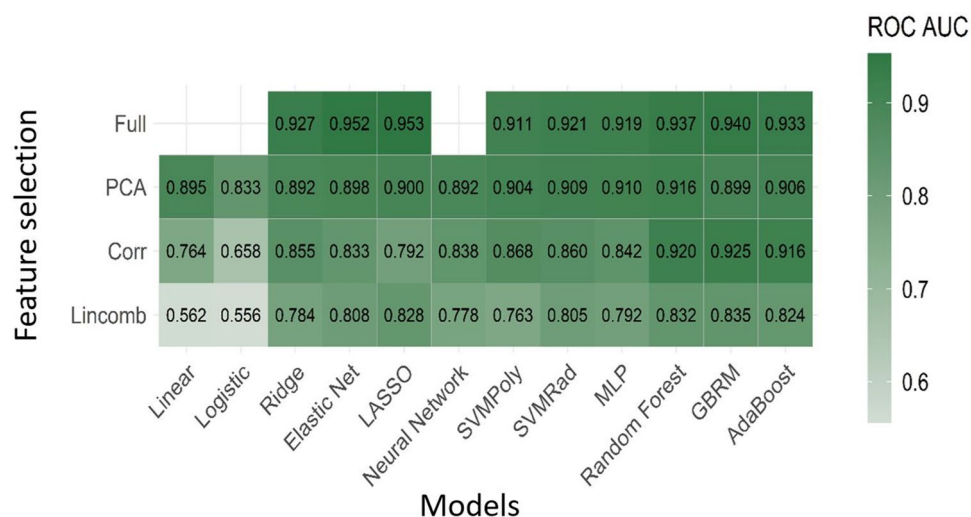


Figure 1. Diagnostic performance using multiparametric MRI. Mean cross-validated ROC AUC for all 45 machine learning and feature reduction combinations using all sequences.

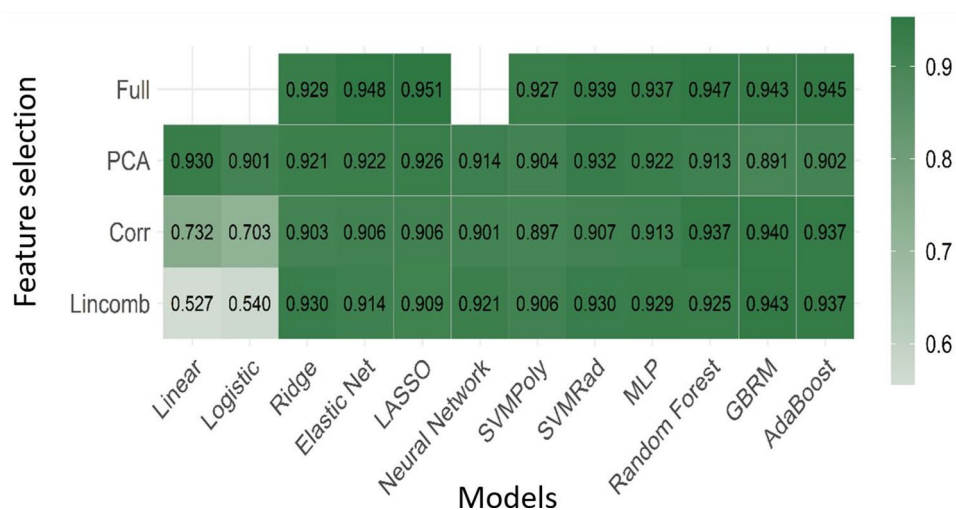


Figure 2. Diagnostic performance using FLAIR sequence. Mean AUC for all models fit using the FLAIR sequence as many of the top performing models came from this sequence.

Feature importance for the models. Features with higher relative importance were derived from both the whole tumor and edema masks. The shape sphericity was the most important feature in all sequence combinations. A boxplot showing the distribution of this feature for the two tumor types on FLAIR sequence is shown in Fig. 4. Supplementary tables S1–S3 (and supplementary Fig. 1–3) display the ranking by variable importance for the ten most important features for the two best models for mp-MRI, FLAIR and combination of T1-CE and FLAIR sequence. Supplementary Fig. 4 shows the heatmap of feature importance for mp-MRI models (features with relative importance greater than 40 were included).

Discussion

Our study showed that radiomics based MLT can differentiate GBM and IMD with excellent performance. We found LASSO and elastic net as the top performing models. Another key observation from our study was that the diagnostic performance for best models was similar for mp-MRI, FLAIR sequence and combined T1-CE and FLAIR sequence. Finally, radiomic features with high relative importance were derived from both the whole tumor and edema masks and shape sphericity was the most important feature.

LASSO and elastic net models are both penalized regression models³⁰. LASSO model forces the coefficient estimates of the variables with limited contribution to the outcome to be exactly zero. Elastic net is an extension of LASSO and forces those estimates of minor importance variables to be zero or very close to zero. Elastic net model is also robust to extreme correlations among predictors^{30,31}. Both of these models help reduce data

All sequences			Sequence-specific			
Model	Feature reduction	ROC AUC mean (SD)	Model	Feature reduction	Sequence	ROC AUC mean (SD)
Lasso	Full	0.953 (0.041)	Lasso	Full	F	0.951 (0.049)
Enet	Full	0.952 (0.038)	Enet	Full	F	0.948 (0.049)
GBRM	Full	0.940 (0.040)	RF	Full	F	0.947 (0.042)
RF	Full	0.937 (0.046)	ada	Full	F	0.945 (0.038)
ada	Full	0.933 (0.045)	Lasso	Full	CE	0.943 (0.041)
Ridge	Full	0.928 (0.053)	GBRM	Full	F	0.943 (0.045)
GBRM	Corr	0.925 (0.056)	GBRM	Lincomb	F	0.943 (0.045)
SVMRad	Full	0.921 (0.042)	Lasso	Full	T1	0.941 (0.046)
RF	Corr	0.920 (0.045)	Enet	Full	T1	0.941 (0.041)
MLP	Full	0.919 (0.055)	GBRM	Corr	F	0.940 (0.046)

Table 2. Top ten best performing models using all sequence and sequence-specific features sets ordered by ROC AUC. Enet: elastic net; Lasso (least absolute shrinkage and selection operator; GBRM: generalized boosted regression model; RF: random forest; ada: boosting of classification trees with adaBoost; SVMRad: SVM with a radial kernel; MLP: multi-layer perceptron; full: full feature set; corr: High correlation filter; lincomb: linear combinations filter; F: FLAIR; CE: T1-CE.

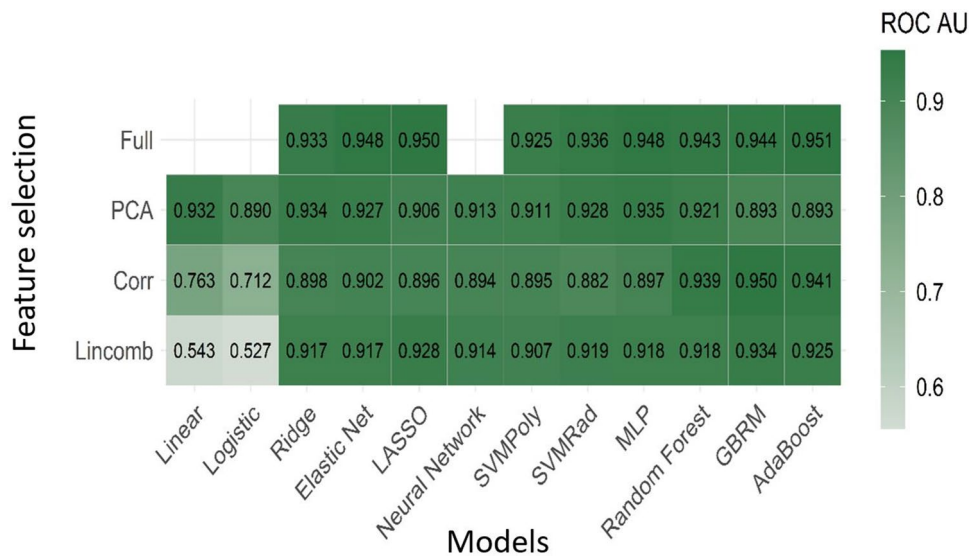


Figure 3. Diagnostic performance using combined T1- contrast enhanced (T1-CE) and FLAIR sequence. Mean AUC for all models fit using the combined T1-CE and FLAIR sequences.

overfitting. In addition, the nested cross-validation used in our study helps in model hyperparameter optimization and reduce bias in model selection, thereby improving generalizability of the model³².

The diagnostic performance for majority of the top performing models in our study for all sequence combinations ranged between AUC 0.936- 0.953. This highlights the fact that albeit small, model performance can vary depending upon the combination of chosen classifier and feature reduction method. Additionally, compared to few prior radiomics studies showing multilayer perceptron (MLP)²⁶ and support vector machine (SVM)²⁷ as high performing models in GBM and IMD classification we also found MLP (AUC 0.919) and SVM (AUC 0.921) as high performing models. In our study, performance of MLP model was similar (AUC 0.919) compared to a prior study by Ortiz et al²⁶ (AUC 0.912). Similarly, SVM model in our study performed superior (AUC 0.921) to a prior study by Qian et al²⁷ where performance of SVM model was AUC 0.900). However, both MLP and SVM were not among the highest performing models in our study and we obtained the best result using embedded LASSO model. These observations suggest that model performance depends upon the available data at hand. As such, reliance on a single classifier type approach may be suboptimal and comprehensive model evaluation should be performed.

Our results also compared favorably to other previously reported studies that also evaluated performance of multiple MLT and feature reduction strategies in classification of GBM versus IMD (supplementary table S4). Bae et al² evaluated seven machine learning classifiers and five feature reduction methods based on radiomics features extracted from T1-CE and T2W images and reported an AUC of 0.890 with an accuracy of 83%. Chen et al³³ evaluated thirty feature reduction and model combinations for their radiomics study and extracted features only

Combinations of CE and F sequences		
Model	Feature reduction	ROC AUC mean (SD)
ada	Full	0.951 (0.040)
Lasso	Full	0.950 (0.042)
GBRM	Corr	0.950 (0.040)
Enet	Full	0.948 (0.041)
MLP	Full	0.948 (0.042)
GBRM	Full	0.944 (0.043)
RF	Full	0.943 (0.044)
ada	Corr	0.941 (0.037)
RF	Corr	0.939 (0.039)
SVMRad	Full	0.936 (0.042)

Table 3. Top ten best performing models using combination of T1-CE and FLAIR sequences ordered by ROC AUC. ada: boosting of classification trees with adaBoost; Lasso (least absolute shrinkage and selection operator; GBRM: generalized boosted regression model; Enet: elastic net; MLP: multi-layer perceptron; rf: random forest; SVMRad: SVM with a radial kernel; full: full feature set; corr: High correlation filter; F: FLAIR; CE: T1-CE.

Performance metric	Model					
	LASSO full All Seqs	Elastic net full All Seqs	LASSO full F	Elastic net full F	adaBoost full CE + F	LASSO full CE + F
Brier	0.088 (0.036)	0.088 (0.036)	0.083 (0.042)	0.088 (0.043)	0.086 (0.040)	0.102 (0.026)
Accuracy	0.892 (0.061)	0.892 (0.063)	0.897 (0.054)	0.885 (0.054)	0.888 (0.063)	0.887 (0.070)
ROC AUC	0.953 (0.041)	0.952 (0.038)	0.951 (0.049)	0.948 (0.049)	0.951 (0.040)	0.950 (0.042)
Sensitivity	0.887 (0.086)	0.893 (0.092)	0.917 (0.064)	0.903 (0.071)	0.900 (0.080)	0.907 (0.088)
Specificity	0.897 (0.073)	0.890 (0.079)	0.877 (0.094)	0.867 (0.102)	0.877 (0.111)	0.867 (0.115)

Table 4. Mean (SD) of performance metrics for two best performing models using all sequences, individual sequences, and combined T1-CE and FLAIR sequences. LASSO (least absolute shrinkage and selection operator; Enet: elastic net; ada: boosting of classification trees with adaBoost; F: FLAIR; CE: T1-CE.

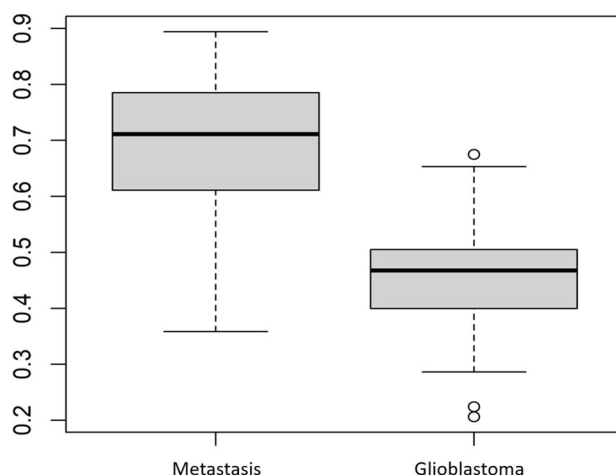


Figure 4. Feature importance. A boxplot showing the distribution of shape sphericity feature for the two tumor types on FLAIR sequence.

from T1-CE sequence with an AUC of 0.80. In contrast, we reported forty-five combinations of feature reduction and classifier MLT and reported a higher AUC of 0.953 and an accuracy of 89%. Our results were better likely due to feature extraction from mp-MRI sequences. Bae et al. also evaluated deep neural network in their study with an AUC of 0.956 and accuracy of 89%² Though, we did not evaluate deep neural network (DNN) due to its

computational complexity, our results were comparable to the study by Bae et al. and suggest that mp-MRI, or even just FLAIR/ FLAIR with T1 CE sequence derived features may provide similar results².

Feature reduction is an integral part of model building process. An important finding of our study was the higher performance of embedded classifier models (LASSO and elastic net) compared to a priori feature reduction. Embedded methods perform feature extraction at the time of model training process. Embedded models have advantage in terms of better generalizability than filter and wrapper methods. They are also quick and computationally less extensive compared to wrapper methods³⁴. A priori filter feature reduction methods may lead to loss of relevant features with a drop in model performance as seen in our study with high correlation and linear combination filter reduction methods. Although, few of the top models also included filter-based methods (Tables 2, 3), most of the top performing models were models with embedded feature selection.

In relation to the mask performance, we found that the most important features were often extracted both from the whole tumor and edema masks. This was true for all sequence combinations (mp-MRI, individual sequence, and from combined T1-CE and FLAIR sequence). Additionally, the most important radiomics feature was shape sphericity. This is likely secondary to different peritumoral environment in GBM and IMD. Since GBM is irregular with variable growth in different dimensions while IMD often has well-defined margins and a more uniform shape, it is not surprising that shape sphericity was the most important discriminating feature. Additionally, the peritumoral region in patients with IMD is invariably vasogenic edema while GBM shows neoplastic infiltration². This may explain the high performance of edema masks. Few prior studies have also shown morphometrical analysis and peritumoral radiomic features as important parameters to differentiate GBM and IMD^{35,36}.

It is pertinent to note here that the model performance did not change when using FLAIR only, as compared to FLAIR and T1-CE in our study. Also, the performance of FLAIR and combined FLAIR and T1-CE sequence was comparable to mp-MRI. Even though it may appear that FLAIR sequence alone might suffice in terms of providing robust discrimination, it is pertinent to note here that the lesion segmentation between enhancing tumor and surrounding 'edema' requires both T1 CE and FLAIR images to generate the masks. At the very minimum therefore, both these sequences would be required for differentiating between GBM and IMD and achieving performance comparable to mp-MRI. This may also be relevant for future validation studies as analysis of one or few MRI sequences is quicker and is easier to implement in terms of time and computational costs. Prior radiomics studies have only assessed single or few MRI sequences for GBM and metastasis classification. Our study provides evidence of comparable diagnostic performance of single MRI sequence versus mp-MRI.

There were several limitations in our study. This was a single center study with limited sample size, and we did not perform external validation of our best performing model that could further establish model generalizability. Secondly, deep neural networks were not evaluated in our study. Another limitation was a less heterogeneous group of brain metastases, since we only included two of the most common primary sites of malignancies (lung and breast). It is possible that increasing radiomic heterogeneity could negatively impact model performance. However, since shape sphericity was the most useful feature and is likely to remain consistent across metastasis as a tumor-class, the impact of additional sub-types of metastasis may not be significant.

However, our study had several strengths including radiomics assessment from both mp-MRI and individual or sequence combinations, different tumor and edema masks and comparison of forty-five different classifier model and feature reduction combinations for each sequence/ combination of sequences. Our study also has one of the best model performances reported for this specific two-class problem. Finally, our results would favor using embedded feature selection over a priori feature reduction.

Conclusion

Radiomics based machine learning can classify GBM and IMD with excellent diagnostic performance. Model performance can vary depending upon the chosen classifier and feature reduction combination and thus comprehensive model selection strategy should be chosen. The performance of mp-MRI and single FLAIR or combined T1-CE and FLAIR sequence is comparable. Radiomic features with higher relative importance came from both whole tumor and edema masks with shape sphericity being the most important feature across multiple models.

Materials and methods

This was a retrospective study approved by institutional review board (IRB) of University of Iowa Hospitals and requirement of informed consent was waived off by University of Iowa Hospitals' IRB (IRB-ID 201912239). The study was carried out in accordance with relevant guidelines and regulations. Between 2010 and 2016, patients were identified from institutional cancer registries and electronic medical records. Eligibility criteria included availability of pre-therapeutic/ pre-operative multiparametric MRI (mp-MRI) images (T1W, T2W, FLAIR, ADC, T1-contrast enhanced (T1-CE)), and the presence of a contrast-enhancing tumor measuring greater than 1 cm in at least one axial dimension. Patients with non-enhancing tumors and motion artifact were excluded. This yielded a total of 60 patients with GBM. A similar number of patients (n = 60) with known lung (n = 40) or breast (n = 20) primary tumor and intracranial metastatic disease were then consecutively selected from the same time period, again using a combination of institutional registries and electronic medical records. In patients with multiple lesions, only the largest lesion was segmented since this approach can provide reliable results by including regions containing a sufficient number of voxels and the same approach has also been utilized in prior studies^{26,37}.

Image acquisition. Preoperative imaging was performed on 1.5 T MRI system (Siemens, Erlangen, Germany). The acquisition protocol for brain tumor evaluation at our hospital includes pre-contrast axial T1W, T2W, FLAIR, DWI with ADC maps, gradient echo and tri-planar T1-CE images (details in supplementary Table S5).

Image pre-processing. DICOM images were first anonymized and converted to NIFTI format. All images were resampled to $1 \times 1 \times 5 \text{ mm}^3$ voxel size using the AFNI package (<https://afni.nimh.nih.gov/>)³⁸. All image sequences acquired during same session were mutually registered to the T1W sequence using Advanced Normalization Tools (ANTs) (<http://stnava.github.io/ANTs/>)³⁹. After image resampling and registration, intensity normalization was performed to [0,255] using the feature scaling method available in the ANTs registration suite.

Tumor segmentation. 3-D tumor segmentation was performed on axial T1-CE and FLAIR images by two radiologists (SP, GB) in consensus using an in-house developed semi-automatic tool ‘Layered Optimal Graph Image Segmentation for Multiple Objects and Surfaces’ (LOGISMOS)⁴⁰ 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. Two separate region of interests (masks) were created using T1-CE and FLAIR images: i) whole tumor (enhancing plus necrotic); and ii) peritumoral edema. First, whole tumor mask was generated from the T1-CE sequence. This was followed by generation of a mask from FLAIR images that included the entire lesion (whole tumor and surrounding edema). Finally, edema mask was extracted by subtracting the T1-CE derived whole tumor mask from FLAIR derived entire lesion mask. Both whole tumor and edema masks were expert-identified. These masks were then superimposed on all five MRI sequences.

Texture features extraction. Features were extracted using Pyradiomics 3.0⁴¹. For each tumor, features were abstracted using above two masks and on each mask five MRI sequences were used. This resulted in 10 possible masks and sequence combinations, on each of which 107 radiomic features were obtained, yielding a total of 1070 features.

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 gray tone difference matrix features ($n = 5$). The default value for the number of bins was fixed by bin width of 25 Gy levels. In rare cases where the edema was minimal (7 patients (6%)) leading to absence of a corresponding mask, the value of the corresponding feature was set to -9999.

Feature reduction. Since the number of features exceeded the sample size, feature reduction was performed to reduce noise, collinearity and dimensionality. Three feature reduction methods were considered: a linear combinations filter, a high correlation filter, and principal components analysis (PCA). The linear combinations (lincomb) filter addresses both collinearity and dimension reduction by finding linear combinations of two or more variables and removing columns to resolve the issue. This process is repeated until the feature set is full rank. The high correlation (corr) filter removes variables from the feature set which have a large absolute correlation. A user-specified threshold is given to determine the largest allowable absolute correlation (threshold 0.4 for mp-MRI, 0.6 for individual sequence, 0.5 for combined T1-CE and FLAIR). Lower thresholds were used as the number of potential features increased for consistency in the number of potential features after filtering. These thresholds led to 60–80 features remaining after high correlation filtering across all potential feature sets.

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 (80% threshold for mp-MRI and 85% for both individual sequence and combined T1-CE and FLAIR). These thresholds were chosen to sufficiently reduce the dimensionality of the feature set for model fitting while retaining many of the important variables and as with the high correlation filter, different thresholds were used to ensure similar numbers of PC’s being used in the modeling process. These thresholds led to 25–30 PC’s used to be used in model fitting. These feature reduction methods were implemented using the recipes package in R version 4.0.2^{42,43}. All variables were standardized, and missing data was imputed using mean imputation prior to any feature reduction. Mean imputation was used for simplicity as the amount of missing data was small. The feature reduction techniques, feature standardization, and imputation were carried out within each cross-validated split of the data, so as not to bias the estimate of predictive performance.

Model fitting. Twelve different machine learning models were fit to determine the best classifier for each feature set. These models can be broadly classified as: linear classifiers, non-linear classifiers, and ensemble classifiers. The linear classifiers used were linear, logistic, ridge, elastic net (enet), and LASSO (least absolute shrinkage and selection operator) regression. The non-linear classifiers used were neural network, support vector machine with a polynomial kernel (svmPoly), SVM with a radial kernel (svmRad), 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.

Classifier model performance evaluation. Each model was fit using the three feature reduction techniques as well as the entire feature set (full), except for the linear regression, logistic regression, and the neural network which cannot be fit with the full feature set due to numerical stability (linear and logistic) and computational complexity (neural network). The entire feature set was used to determine if embedded feature selection of the machine learning model itself performed better than a priori feature reduction using either of the three feature reduction techniques. This yielded 45 possible model/feature reduction combinations to be fit to each of the possible feature sets. A fivefold repeated cross-validation with five repeats was then performed to evaluate

Model	MachineShop Model constructor	Parameters tuned	Size of tuning grid
Linear	LMMModel	N/A	–
Logistic	GLMModel	N/A	–
Ridge	GLMNetModel Alpha = 0	Lambda	5
Elastic Net	GLMNetModel	Alpha, Lambda	25
LASSO	GLMNetModel Alpha = 1	Lambda	5
Neural Network	NNetModel	Size, Decay	25
SVM Poly	SVMPolyModel	C, Degree, Scale	125
SVM Radial	SVMRadialModel	C, Sigma	25
MLP	N/A	Size, Maxit, learnFuncParams	27
Random Forest	RandomForestModel	Mtry	5
GBRM	GBMModel	n.trees, interaction.depth	25

Table 5. Summary of tuning for each model. LASSO: (least absolute shrinkage and selection operator); SVM Poly: support vector machine with a polynomial kernel; SVM Radial; MLP: multi-layer perceptron; LMMModel: linear model; GLMModel: generalized linear model; GLMNetModel: generalized linear model with penalized maximum likelihood; NNetModel: feed-forward neural networks; GBMModel: generalized boosted regression; Lambda: regularization parameter; Alpha: elastic net mixing parameter; Size: number of units in the hidden layer(s); Decay: parameter for weight decay; C: cost of constraints violation, regularization term in the Lagrange formulation; Degree: degree of the polynomial kernel function; Scale: scaling parameter of the polynomial kernel; Sigma: inverse kernel width; Maxit: maximum iterations to learn; learnFuncParams: parameters of the learning function; Mtry: number of variables randomly sampled as candidates at each split; n.trees: total number of trees to fit; interaction.depth: maximum depth of variable interactions.

the predictive performance of each model, giving 25 total estimates of prediction performance for each model/feature reduction combination.

For models with tuning parameters, the important parameters were tuned using nested cross-validation, which has been shown to provide unbiased performance estimates regardless of the sample size⁴⁶. Tuning of important parameters was handled using the functionality of the MachineShop R package⁴⁴. Tuning grids across 5 different values for each tuning parameter were used, except for the MLP model which used a grid across 3 different values for each tuning parameter due to the computational complexities of that model. Other parameters were set to the package default values, which can be found in the online documentation for each model constructor function⁴⁵. For example, for the random forest model, the number of trees was 500 (default), the number of variables randomly sampled as candidates at each split was tuned across a grid of five possible values, and the minimum size of the terminal nodes was 1 (default). Although the MachineShop package offers users the ability to choose the values in the tuning grid, we utilized the functionality of the package which automatically generates the grid of possible values for the tuning parameters due to the large number of models under consideration.

Nested Cross-validation was implemented using the “resample” function in the MachineShop package using the “TunedModel” constructor function to perform model tuning within each fold. This function automatically creates cross-validated splits using different seeds, and an overall seed is provided to ensure the splits are the same for each candidate model. Table 5 provides the MachineShop model constructor function name, the parameters that were tuned, and the total length of the tuning grid for each of the models we considered. The MLP is not implemented by default in MachineShop but was implemented using the package’s ability for users to define a custom model to ensure it was evaluated on the same cross-validated splits as all of the other models.

Feature importance. Feature importance was calculated for the top models using the final model fit to the entire data set from each of the three scenarios. Feature importance for the LASSO and elastic net models was defined as the absolute value of the regression coefficients. As the features are standardized, larger values of the coefficients correspond to higher importance. For the adaBoost model, the feature importance was determined by finding the improvement in the Gini index attributed to a split on that feature in a tree and the weight of that tree⁴⁷. Importance values were rescaled to range from 0 to 100, with the most important feature having a value of 100.

Statistical analysis. The goal of this analysis was to determine the best model for tumor classification and to determine if models fit using all five MRI sequences (mp-MRI) performed better than models fit to the individual sequences or models fit using combined T1-CE and FLAIR sequences. For the models using mp-MRI,

there were 1070 possible features (2 masks \times 5 sequences \times 107 features). For each sequence specific model, the feature set included 214 (2 masks \times 1 sequence \times 107 features) possible features, and for the T1-CE and FLAIR combination models, the feature set included 428 (2 masks \times 2 sequences \times 107 features) possible features.

Model fitting and cross-validated predictive performance was implemented using the MachineShop and RSNNS packages in R version 4.0.2⁴⁴. Predictive performance was measured with the area under the receiver operating characteristic curve (AUC) for interpretability. As models were formulated to predict GBM, AUC estimates the probability that a randomly selected subject that had a GBM will have a greater predicted value than a randomly selected subject that had a metastatic tumor. Higher AUC values indicate better predictive performance. The mean AUC between each of these three models was compared using a paired t-test. We provide the mean/SD of the AUC over the 25 cross-validated estimates to describe the distribution of AUC estimates.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 10 November 2020; Accepted: 5 May 2021

Published online: 18 May 2021

References:

- Artzi, M., Bressler, I. & Ben Bashat, D. Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. *Journal of magnetic resonance imaging: JMIR* **50**, 519–528. doi:<https://doi.org/10.1002/jmri.26643> (2019).
- Bae, S. *et al.* Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. *Sci. Rep.* **10**, 12110. <https://doi.org/10.1038/s41598-020-68980-6> (2020).
- Ostrom, Q. T. *et al.* CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016. *Neuro Oncol.* **21**, v1–v100. <https://doi.org/10.1093/neuonc/noz150> (2019).
- Dong, F. *et al.* Differentiation of supratentorial single brain metastasis and glioblastoma by using peri-enhancing oedema region-derived radiomic features and multiple classifiers. *Eur. Radiol.* **30**, 3015–3022. <https://doi.org/10.1007/s00330-019-06460-w> (2020).
- Skogen, K. *et al.* Texture analysis on diffusion tensor imaging: discriminating glioblastoma from single brain metastasis. *Acta Radiol.* <https://doi.org/10.1177/0284185118780889> (2018).
- Bander, E. D. *et al.* Tubular brain tumor biopsy improves diagnostic yield for subcortical lesions. *J. Neurooncol.* **141**, 121–129. <https://doi.org/10.1007/s11060-018-03014-w> (2019).
- Callovini, G. M. *et al.* How is stereotactic brain biopsy evolving? A multicentric analysis of a series of 421 cases treated in Rome over the last sixteen years. *Clin. Neurol. Neurosurg.* **174**, 101–107. <https://doi.org/10.1016/j.clineuro.2018.09.020> (2018).
- Berghoff, A. S. *et al.* Descriptive statistical analysis of a real life cohort of 2419 patients with brain metastases of solid cancers. *ESMO Open.* <https://doi.org/10.1136/esmoopen-2015-000024> (2016).
- Fink, K. R. & Fink, J. R. Imaging of brain metastases. *Surg Neurol Int* **4**, S209–219. <https://doi.org/10.4103/2152-7806.111298> (2013).
- Mouthuy, N., Cosnard, G., Abarca-Quinones, J. & Michoux, N. Multiparametric magnetic resonance imaging to differentiate high-grade gliomas and brain metastases. *Journal of neuroradiology = Journal de neuroradiologie* **39**, 301–307. <https://doi.org/10.1016/j.neurad.2011.11.002> (2012).
- Bauer, A. H., Erly, W., Moser, F. G., Maya, M. & Nael, K. Differentiation of solitary brain metastasis from glioblastoma multiforme: a predictive multiparametric approach using combined MR diffusion and perfusion. *Neuroradiology* **57**, 697–703. <https://doi.org/10.1007/s00234-015-1524-6> (2015).
- Suh, C. H., Kim, H. S., Jung, S. C., Choi, C. G. & Kim, S. J. Perfusion MRI as a diagnostic biomarker for differentiating glioma from brain metastasis: a systematic review and meta-analysis. *Eur. Radiol.* **28**, 3819–3831. <https://doi.org/10.1007/s00330-018-5335-0> (2018).
- Wang, S. *et al.* Differentiation between glioblastomas, solitary brain metastases, and primary cerebral lymphomas using diffusion tensor and dynamic susceptibility contrast-enhanced MR imaging. *AJNR Am. J. Neuroradiol.* **32**, 507–514. <https://doi.org/10.3174/ajnr.A2333> (2011).
- Kadota, Y. *et al.* Differentiation between glioblastoma and solitary brain metastasis using neurite orientation dispersion and density imaging. *J. Neuroradiol.* **47**, 197–202. <https://doi.org/10.1016/j.neurad.2018.10.005> (2020).
- Li, X. *et al.* Discrimination between glioblastoma and solitary brain metastasis: comparison of inflow-based vascular-space-occupancy and dynamic susceptibility contrast MR imaging. *AJNR Am. J. Neuroradiol.* **41**, 583–590. <https://doi.org/10.3174/ajnr.A6466> (2020).
- Bulakbasi, N. *et al.* Assessment of diagnostic accuracy of perfusion MR imaging in primary and metastatic solitary malignant brain tumors. *AJNR Am. J. Neuroradiol.* **26**, 2187–2199 (2005).
- Calli, C. *et al.* Perfusion and diffusion MR imaging in enhancing malignant cerebral tumors. *Eur. J. Radiol.* **58**, 394–403. <https://doi.org/10.1016/j.ejrad.2005.12.032> (2006).
- Kono, K. *et al.* The role of diffusion-weighted imaging in patients with brain tumors. *AJNR Am. J. Neuroradiol.* **22**, 1081–1088 (2001).
- Server, A. *et al.* Proton magnetic resonance spectroscopy in the distinction of high-grade cerebral gliomas from single metastatic brain tumors. *Acta radiologica (Stockholm, Sweden : 1987)* **51**, 316–325. <https://doi.org/10.3109/02841850903482901> (2010).
- Tsougos, I. *et al.* Differentiation of glioblastoma multiforme from metastatic brain tumor using proton magnetic resonance spectroscopy, diffusion and perfusion metrics at 3 T. *Cancer Imaging* **12**, 423–436. <https://doi.org/10.1102/1470-7330.2012.0038> (2012).
- Soni, N., Priya, S. & Bathla, G. Texture analysis in cerebral gliomas: A review of the literature. *AJNR Am. J. Neuroradiol.* **40**, 928–934. <https://doi.org/10.3174/ajnr.A6075> (2019).
- Kandemirli, S. G. *et al.* Presurgical detection of brain invasion status in meningiomas based on first-order histogram based texture analysis of contrast enhanced imaging. *Clin. Neurol. Neurosurg.* **198**, <https://doi.org/10.1016/j.clineuro.2020.106205> (2020).
- Priya, S. *et al.* Glioblastoma and primary central nervous system lymphoma: differentiation using MRI derived first-order texture analysis - a machine learning study. *The neuroradiology journal*, 1971400921998979. <https://doi.org/10.1177/1971400921998979> (2021).
- P. Sun, D. W., V. C. Mok and L. Shi. Comparison of Feature Selection Methods and Machine Learning Classifiers for Radiomics Analysis in Glioma Grading. *IEEE Access*, 7, 102010–102020. <https://doi.org/10.1109/ACCESS.2019.2928975> (2019).
- Priya, S. *et al.* Survival prediction in glioblastoma on post-contrast magnetic resonance imaging using filtration based first-order texture analysis: Comparison of multiple machine learning models. *The neuroradiology journal*, 1971400921990766. <https://doi.org/10.1177/1971400921990766> (2021).
- Ortiz-Ramón, R., Ruiz-España, S., Mollá-Olmos, E. & Moratal, D. Glioblastomas and brain metastases differentiation following an MRI texture analysis-based radiomics approach. *Phys. Medica* **76**, 44–54. <https://doi.org/10.1016/j.ejmp.2020.06.016> (2020).

27. Qian, Z. *et al.* Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers. *Cancer Lett.* **451**, 128–135. <https://doi.org/10.1016/j.canlet.2019.02.054> (2019).
28. Skogen, K. *et al.* Texture analysis on diffusion tensor imaging: discriminating glioblastoma from single brain metastasis. *Acta radiologica (Stockholm, Sweden : 1987)* **60**, 356–366. <https://doi.org/10.1177/0284185118780889> (2019).
29. Zhang, G. *et al.* Discrimination between solitary brain metastasis and glioblastoma multiforme by using ADC-based texture analysis: A comparison of two different ROI placements. *Acad. Radiol.* **26**, 1466–1472. <https://doi.org/10.1016/j.acra.2019.01.010> (2019).
30. Ogutu, J. O., Schulz-Streeck, T. & Piepho, H. P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC proceedings* **6 Suppl 2**, S10. <https://doi.org/10.1186/1753-6561-6-s2-s10> (2012).
31. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1–22 (2010).
32. Parvande, S., Yeh, H.-W., Paulus, M. P. & McKinney, B. A. Consensus Features Nested Cross-Validation. *bioRxiv*, 2019.2012.2031.891895. <https://doi.org/10.1101/2019.12.31.891895> (2020).
33. Chen, C., Ou, X., Wang, J., Guo, W. & Ma, X. Radiomics-based machine learning in differentiation between glioblastoma and metastatic brain tumors. *Front. Oncol.* **9**, 806–806. <https://doi.org/10.3389/fonc.2019.00806> (2019).
34. Lal, T. N., Chapelle, O., Weston, J. & Elisseeff, A. in *Feature Extraction: Foundations and Applications* (eds Isabelle Guyon, Masoud Nikravesh, Steve Gunn, & Lotfi A. Zadeh) 137–165 (Springer Berlin Heidelberg, 2006).
35. Yang, G., Jones, T. L., Howe, F. A. & Barrick, T. R. Morphometric model for discrimination between glioblastoma multiforme and solitary metastasis using three-dimensional shape analysis. *Magn. Reson. Med.* **75**, 2505–2516. <https://doi.org/10.1002/mrm.25845> (2016).
36. Blanchet, L. *et al.* Discrimination between metastasis and glioblastoma multiforme based on morphometric analysis of MR images. *AJNR Am. J. Neuroradiol.* **32**, 67–73. <https://doi.org/10.3174/ajnr.A2269> (2011).
37. Lohmann, P. *et al.* Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. *NeuroImage: Clinical* **20**, 537–542. <https://doi.org/10.1016/j.nicl.2018.08.024> (2018).
38. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173. <https://doi.org/10.1006/cbmr.1996.0014> (1996).
39. Avants BB, T. N., Song G. Advanced normalization tools (ANTS). *Insights Journal*, 365:361–335. (2009).
40. Yin, Y. *et al.* LOGISMOS-layered optimal graph image segmentation of multiple objects and surfaces: cartilage segmentation in the knee joint. *IEEE Trans. Med. Imaging* **29**, 2023–2037. <https://doi.org/10.1109/tmi.2010.2058861> (2010).
41. van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Can. Res.* **77**, e104–e107. <https://doi.org/10.1158/0008-5472.Can-17-0339> (2017).
42. Kuhn, M. a. W., H. Preprocessing Tools to Create Design Matrices. R package version 0.1.9 (2020).
43. R Development Core Team (2006). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, V, Austria. ISBN 3–900051–07–0.
44. Smith, B. J. MachineShop: Machine Learning Models and Tools. R package version 2.4.0. (2020).
45. <https://brian-j-smith.github.io/MachineShop/reference.html>.
46. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **14**, <https://doi.org/10.1371/journal.pone.0224365> (2019).
47. Alfaró, E., Gamez, M. & García, N. adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software; Vol 1, Issue 2 (2013)* (2013).

Author contributions

Study concepts and design: GB, VM, MS; Literature search: GB, SP; NS, RPM; Data acquisition: SP, GB, YL, VM, NHL, HZ; Data analysis and interpretation: SP, GB, YL, CW, NS, RPM, NHL, VM, HZ, MS; Writing first draft: SP, GB; Writing, review, and/or revision of the manuscript: All authors.

Competing interests

Other 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. Rest of the authors report no relationships that could be construed as a conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-90032-w>.

Correspondence and requests for materials should be addressed to S.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021