ORIGINAL RESEARCH

BRAIN TUMOR

Impact of SUSAN Denoising and ComBat Harmonization on Machine Learning Model Performance for Malignant Brain Neoplasms

Girish Bathla, Neetu Soni, Ian T Mark, Yanan Liu, Nicholas B Larson, Blake A Kassmeyer, Suyash Mohan, John C Benson, Saima Rathore, Amit Agarwal

ABSTRACT

BACKGROUND AND PURPOSE: Feature variability in radiomics studies due to technical and magnet strength parameters is well known and may be addressed through various pre-processing methods. However, very few studies have evaluated downstream impact of variable pre-processing on model classification performance in a multi-class setting. We sought to evaluate the impact of SUSAN denoising and ComBat harmonization on model classification performance.

MATERIALS AND METHODS: A total of 493 cases (410 internal and 83 external dataset) of glioblastoma (GB), intracranial metastatic disease (IMD) and primary CNS lymphoma (PCNSL) underwent semi-automated 3D-segmentation post baseline image processing (BIP) consisting of resampling, realignment, co-registration, skull stripping and image normalization. Post BIP, two sets were generated, one with and another without SUSAN denoising (SD). Radiomics features were extracted from both datasets and batch corrected to produce four datasets: (a) BIP, (b) BIP with SD, (c) BIP with ComBat and (d) BIP with both SD and ComBat harmonization. Performance was then summarized for models using a combination of six feature selection techniques and six machine learning models across four mask-sequence combinations with features derived from one-three (multi-parametric) MRI sequences.

RESULTS: Most top performing models on the external test set used BIP+SD derived features. Overall, use of SD and ComBat harmonization led to a slight but generally consistent improvement in model performance on the external test set.

CONCLUSIONS: The use of image pre-processing steps such as SD and ComBat harmonization may be more useful in a multiinstitutional setting and improve model generalizability. Models derived from only T1-CE images showed comparable performance to models derived from multiparametric MRI.

Received month day, year; accepted after revision month day, year.

From the Departments of Radiology, (G.B, I.T.M, J.C.B), Department of Quantitative Health Sciences (N.B.L,B.A.K), Mayo Clinic, Rochester, Minnesota; Department of Radiology (N.S, A.A), Mayo Clinic, Jacksonville, Florida; Advanced Pulmonary Physiomic Imaging Laboratory (Y.L), University of Iowa Hospitals and Clinics, Iowa City, IA; Department of Radiology (S.M), University of Pennsylvania, Philadelphia, PA 19104 USA; Avid Radiopharmaceuticals (S.R), 3711 Market Street, Philadelphia, PA 19104, USA

Conflict of interests: None for all authors. Please address correspondence to Girish Bathla, MD. Department of Radiology, Mayo Clinic, 200 1st St SW, Rochester, MN, 55902, USA, bathla.girish@mayo.edu

SUMMARY SECTION

PREVIOUS LITERATURE: Studies have previously addressed the impact of various image acquisition and processing parameters at the radiomics feature repeatability and reproducibility, primarily utilizing healthy volunteers or phantoms

KEY FINDINGS: Use of Smallest Univalue Segment Assimilating Nucleus (SUSAN) denoising and ComBat harmonization trended towards a slight but generally consistent improvement in model performance on the external test set.

KNOWLEDGE ADVANCEMENT: Imaging pre-processing steps such as SUSAN denoising and ComBat harmonization may help achieve marginally improved classification performance in a multi-institutional setting

INTRODUCTION

Glioblastoma (GB), intracranial-metastatic disease (IMD) and primary central nervous system lymphomas (PCNSL) are the three most common malignant intra-axial brain tumors. As the treatment strategies are different, their accurate non-invasive diagnosis would be ideal but is difficult due to overlapping imaging appearances, which are well described in the neuroradiology literature.[1-4] Several prior studies have addressed non-invasive image-based differentiation between GB, IMD and PCNSL using machine learning (ML), either as a binary or a three-class problem.[1-3, 5] Many studies have shown encouraging results, often better than human readers. However, one of the potential drawbacks with these studies is the variability in pre-processing steps that were followed prior to model training, either at image level or at radiomics level.[4]

Even though some studies have previously addressed the impact of various image acquisition and processing parameters at the radiomics feature repeatability and reproducibility, many studies used healthy volunteers or phantoms.[6-10] These studies have shown baseline variability in radiomic features based on acquisition parameters, scanner strength, acquisition protocols, slice thickness etc., which may be improved with pre-processing steps such as resampling, intensity normalization, denoising, bias-field correction and harmonization.[7, 8, 11, 12] These pre-processing methods can potentially improve the repeatability and reliability of the radiomics results.[13] However, the impact of these processing parameters on final model classification has seldom been comprehensively evaluated on a large dataset.

Smallest Univalue Segment Assimilating Nucleus (SUSAN) denoising is often used to help reduce image noise and improve signal to noise ratio, given its ability to simultaneously detect and preserve edges in an image.[13, 14] The technique works on a pixel-by-pixel basis and smoothens out pixel intensities based on a thresholding method. Combining Batches (ComBat), on the other hand, is a data-driven post-processing method which was initially used to correct 'batch effects' in genomic studies.[11, 15] More recently, it has been used to address scanner effects to improve downstream analysis in radiomics studies. Unlike other pre-processing methods, ComBat is applied to already extracted features at the radiomic level rather than the image level. Between 2017 and 2022, at least 51 papers reported the use of ComBat in radiomic studies on MRI (36%), CT (34%) and PET imaging (28%) with 41% reporting higher performance while 18% not reporting any additional benefit with ComBat.[15, 16] We aimed to investigate if the application of SUSAN denoising, working at image level to reduce noise, and ComBat harmonization, working at feature level to harmonize radiomic features, either alone or in combination, would improve classification performance for a three-class problem (GB vs IMD vs PCNSL) involving malignant brain neoplasms as compared to models not using either of these methodologies. Similar to the prior seminal work by *Moradmand* et. al., image resampling, co-registration, skull stripping and intensity normalization were considered as baseline image processing and were common to all feature sets.[13] Herein, we present our findings on multiple ML models derived from single or multi-parametric conventional MRI images (derived from a combination of T2WI [T2], FLAIR [F]), ADC [A] and T1-CE [CE] sequences) with (a) baseline image processing (BIP), (b) BIP with SUSAN Denoising (SD), (c) BIP with ComBat harmonization and (d) BIP with both SD and ComBat harmonization.

MATERIALS AND METHODS

Data Collection

The dual-institutional study was approved by the respective institutional review boards and informed consent was waived given the retrospective nature of the study. For the training data, institutional cancer registries from the first hospital were searched for patients with GB, PCNSL and IMD (from a lung, breast, or melanoma primary) between 2010 and 2020 who underwent a contrast-enhanced brain MRI. Inclusion criteria included (a) At least one enhancing lesion > 1 cm, (b) availability of index pre-therapy MRI imaging, (c) availability of required sequences, (d) histological confirmation (for GB and PCNSL cases) or either histological confirmation or known systemic malignancy with imaging appearance consistent with metastatic disease (for IMD cases) and (e) absence of significant motion artifacts. This yielded a total of 547 cases (GB: 231; IMD: 247; PCNSL: 69). Cases were excluded if there were one or more missing sequences (axial T1, T2, FLAIR, ADC and T1-CE) (n=30), failure of any of the below described image pre-processing steps (n=15), or any of the masks for the lesions were not available (n=92). The latter was done to avoid imputing values which could confound the impact of pre-processing steps. A total of 410 cases were eventually included in the internal dataset (GB: 171; IMD: 188; PCNSL: 51).
In addition, the external test set, obtained from another institution, included a total of 83 cases (GB: 25; IMD: 32; PCNSL: 26). Cases were again collected using the same inclusion criteria and processed using identical pipelines as detailed below. The overall study workflow is provided in Fig-1.

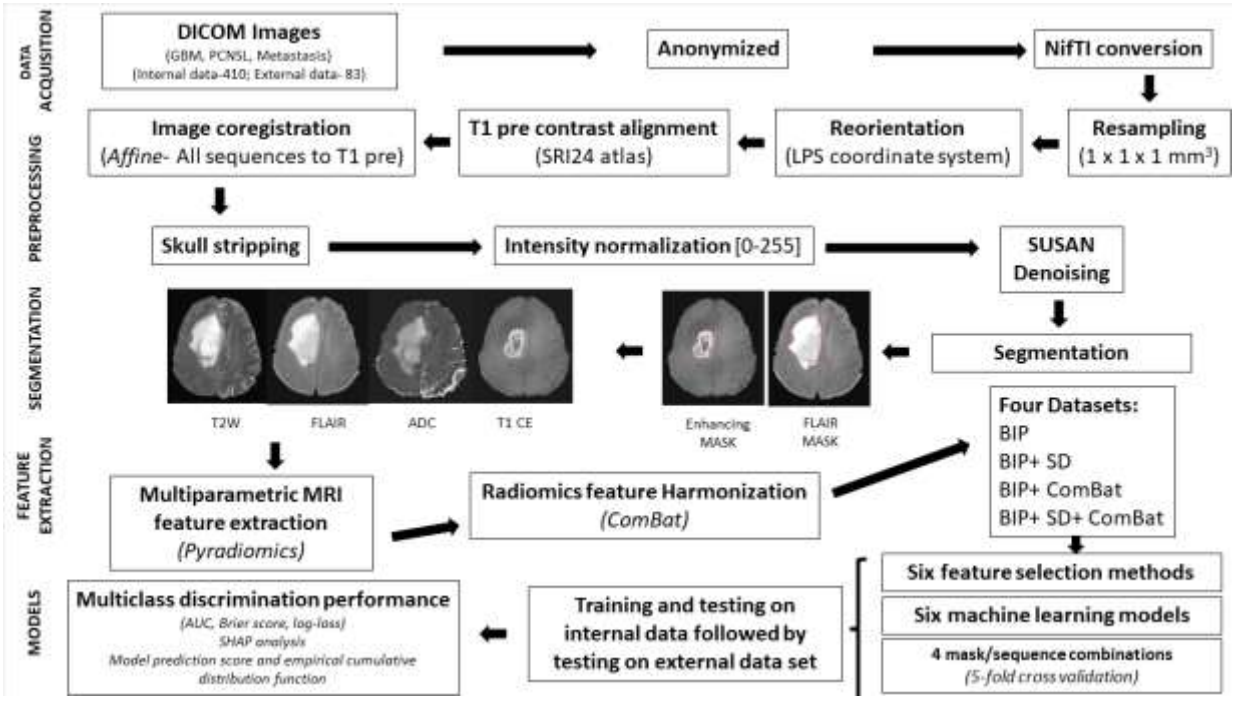


FIG 1. Schematic diagram depicting the overall study workflow.

Image Acquisition

Images were acquired on either a 1.5T (Siemens Aera, Avanto, Siemens Healthcare, Erlangen, Germany) or 3T (Skyra, Trio, Siemens Healthcare, Erlangen, Germany) systems. The typical scanner parameters of the sequences

used at both institutions are provided in the supplementary data. T1W-CE images were acquired 3–5 minutes after administration of gadobenate dimeglumine (Multihance; Bayer Healthcare Pharma) or gadobutrol (Gadavist; Bayer Healthcare Pharma, Berlin, Germany) injected at the rate of 0.1 mL/kg body weight.

Image Preprocessing

Following image anonymization and conversion of the DICOM images to NIfTI format, baseline image processing (BIP) was performed on all images as follows: (a) resampling (1x1x1 mm3); (b) reorientation to the left-posterior-superior (LPS) coordinate system; (c) alignment of T1 pre-contrast images to the SRI24 atlas (d) co-registration; (d) skull stripping; (e) intensity normalization to [0,255] (details in supplementary data). SUSAN denoising was also performed, thereby creating two sets of cases, one with and other without SUSAN denoising (BIP±SD).

Tumor Segmentation

Semi-automated three-dimensional (3D) volumetric tumor segmentation was performed on axial T1-CE and FLAIR images by two board-certified radiologists (N.S. and G.B.) in consensus using LOGISMOS, as detailed previously.[17] In patients with multiple lesions, only the largest lesion was segmented. Two regions of interest (masks) were segmented using T1-CE and FLAIR images: (i) Tumor (ET, enhancing plus necrotic/ hemorrhagic intra-tumoral components on T1-CE images) and (ii) Region of FLAIR abnormality, including tumor and peritumoral region [PTR]. The PTR mask for each lesion was generated by subtracting the ET from the corresponding FLAIR mask. Besides the segmentation of the internal and external cohorts, approximately 17% of the internal patient cohort was randomly re-segmented (n=69, GB:30; IMD: 28; PCNSL:11). This was used for the ML pipelines as described below.

Model Development:

Feature Extraction and Harmonization

For each tumor, radiomic features were extracted from the ET mask and PTR mask using PyRadiomics v3.0.[18] This was done for both datasets, with and without denoising. Radiomic features were harmonized by implementing neuroCombat package in R version 4.2.2, under default settings for both datasets, thereby resulting in four datasets for model training (BIP, BIP+SD, BIP+ComBat, BIP+SD+ComBat).[19] Details about extracted features and harmonization are provided in supplementary data. Since there were several possible mask and sequence combinations, a few select mask-sequence combinations were chosen based on prior literature to assess the classification performance. The following abbreviations follow ‘sequence_mask’ nomenclature throughout the text unless stated otherwise. ([i] CE_ET and F_PTR; [ii] CE_ET and T2_PTR; [iii] CE_ET, A_ET and F_PTR; [iv] CE_ET only).[1, 2, 20]

Feature Selection

Feature selection and reduction methods included: (a) linear combination filter (LCF), (b) correlation-based filtering, (c) principal components analysis (PCA), (d) supervised LASSO variable selection, and (e) Intra-class correlation (ICC) filtering based on the re-segmentation analysis (detailed description in supplementary data). In addition, the entire feature set, without a priori feature selection was also used.

Model Training

ML algorithms employed included kernel support vector machines using the polynomial (SVM-P) and Gaussian (SVM-RBF) kernels, multinomial elastic-net (ENET), extreme gradient boosting (XGB), generalized boosted regression models (GBRM) and random forest (RF). These ML models were chosen given their diverse nature and common use in neuro-oncology ML literature. Model training was performed using masks derived from 1-3 different sequences, either alone or in combination, across a total of four sequence permutations as mentioned earlier. To assess performance on the internal dataset, 5-fold nested cross-validation was performed, and performance was summarized using the mean of the leave-out outer-fold discrimination statistics. The top performing algorithm for each of the four sequence permutations were then trained on the complete internal dataset before performance evaluation on external test set.

Statistical analysis:

Multi-class discrimination performance measures in leave-out test data were computed using the mlr3measures R package, including multinomial log-loss, multi-class Brier score, and the multi-class area under the ROC curve (mAUC) defined by Hand and Till.[21, 22] Brier scores were calculated using the originally outlined definition which is extensible to multi-class problems and has a range of (0, 2). A purely 'non-informative' model that always assigns uniform probabilities to all classes under a three-class problem will correspond to a Brier score of 0.666. All statistical analyses and ML model fitting were performed using R 4.2.2 (R Core Team, Vienna, Austria).[22]

RESULTS

The patient demographic details, scanner and class distributions for the internal data, re-segmented data and external data are provided in Table 1**.** Figure 2 (violin plots) depicts the range of mAUC for all four image processing pipelines for the different feature sets derived from the four mask-sequence combinations for the external data. A similar representation of model performance on the internal dataset is provided in supplemental Figure 1.

Table 1: Patient demographic details, scanner, and class distributions in the internal and external datasets.

	Internal (N=410)	External (N=83)
Scanner		
1.5T	371 (90.4%)	51 (61.4%)
3T	39 (9.5%)	32 (38.6%)
Age		
Mean (SD)	62.2 (12.3)	62.6 (12.2)
Range	11.0 - 90.00	26.0 - 83.0
Sex		
Female	196 (47.8%)	40 (48.2%)
Male	214 (52.1%)	43 (51.8%)
class		
GBM	171 (41.7%)	25 (30.1%)
IMD	188 (45.8%)	32 (38.6%)
PCNSL	51 (12.4%)	26 (31.3%)

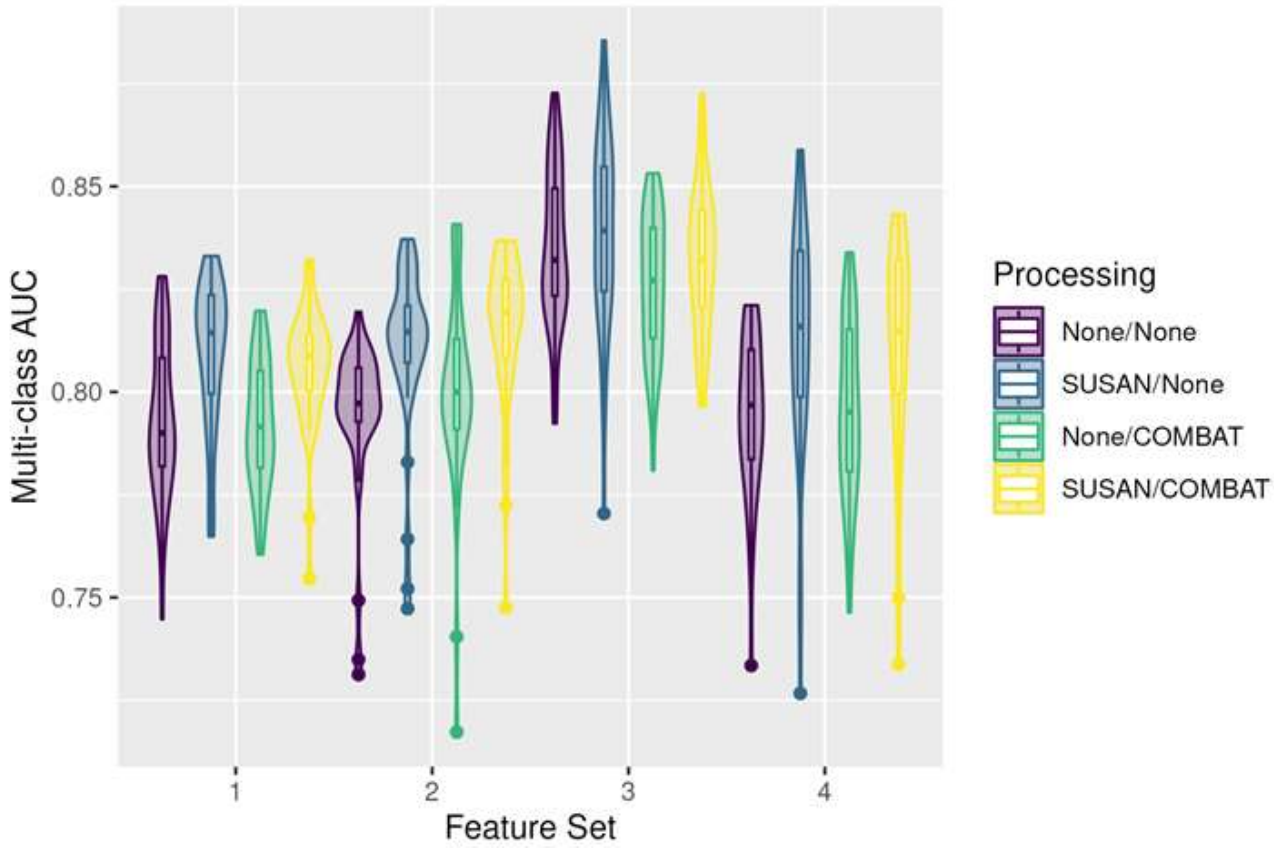


Fig 2: Violin plots for all four feature sets using the external data show the range of multi-class AUC across different pipelines. Feature set 1: CE_ET and F_PTR; 2: CE_ET and T2_PTR; 3: CE_ET, A_ET and F_PTR; 4: CE_ET only.

In general, models using three masks (CE_ET, A_ET and F_PTR), showed slightly better performance (maximal mAUC: 0.873-0.886), which was nevertheless comparable to the other models, including those using only the CE_ET masks (maximal mAUC: 0.856-0.859). Table 2 shows the top three models for each mask-sequence combination for the external data. The top three models for the internal data are provided in Supplementary Table 1. Figure 3 shows the maximum mAUC for the internal and external data for the various models based on the ML algorithm. A histogram plot showing differences between internal and external data model performance for the different pipelines is presented in Figure 4, and showed minimal mean drop in mAUC between internal and external validation dataset across all four processing pipelines with a mean drop of <0.1 in mAUC across all pipelines. In general, the addition of SUSAN denoising to the pre-processing led to slightly improved performance over BIP, with improvement in mAUC ranging between 0.009-0.040 on the external dataset (Table 3). Most of the top three performing models for each mask-sequence combination used SUSAN denoising, while only one of the models in the same list was derived from only BIP data (Table 2). Interestingly, none of the top three models from all four mask-sequence combinations used ComBat on the internal dataset, while four of the models among the top performers for each mask-sequence combination on the external test set used ComBat, which may suggest that ComBat may be helpful when testing models on data derived in a multi-institutional setting. The change in mAUC on the external dataset when comparing BIP only with BIP+ ComBat derived models for the four mask-sequence combinations ranged between -0.037 to +0.033 (table 3). Of note, all three top performing models using the CE_ET and T2_PTR used ComBat (Table 2).

Table 2: Summary of top 3 performing models in the external dataset for each feature set.

Feature Set	Processing	Algorithm	Feature Selection	mAUC	LogLoss	Brier Score
CE_ET and F_PTR	SD/None	SVM-P	ICC	0.833	0.871	0.521
	SD/COMBAT	GBRM	linearComb	0.832	0.860	0.507
	SD/None	SVM-RBF	corr	0.831	0.835	0.519
CE_ET and T2_PTR	None/COMBAT	ENET	none	0.841	0.922	0.492
	None/COMBAT	SVM-P	LASSO	0.840	0.896	0.505
	None/COMBAT	SVM-P	linearComb	0.839	0.915	0.509
CE_ET, A_ET and F_PTR	SD/None	SVM-P	ICC	0.886	0.712	0.414
	SD/None	SVM-P	PCA	0.874	0.699	0.398

Feature Set	Processing	Algorithm	Feature Selection	mAUC	LogLoss	Brier Score
	None/None	SVM-P	ICC	0.873	0.764	0.433
CE_ET	SD/None	SVM-P	ICC	0.859	0.789	0.472
	SD/None	SVM-P	none	0.856	0.800	0.499
	SD/None	SVM-P	lasso	0.856	0.786	0.494

[ENET: Multinomial elastic net, ICC: Inter-class correlation, GBRM: generalized boosted regression model, LASSO: least absolute shrinkage and selection operator, PCA: principal component analysis, SD: SUSAN denoising, SVM-P: support vector machine-polynomial kernel, SVM-RBF: support vector machine-gaussian kernel]

Bootstrapping of the mAUC was also performed using the various pipelines under the same image pre-processing as a single cluster, and 5000 bootstrap samples were drawn for percentile based 95% CI limit calculations. These did not reveal any significant differences between the pipelines (suppl table 2). However, it is pertinent to note here that each cluster had about 144 ML pipelines (four sequence combinations, six feature selection and six ML models), and was evaluating the class difference in image processing pipelines as a whole, and not just the top performing models.

We further performed nonparametric Kruskal-Wallis test to assess the impact of image pre-processing on feature importance. This was done for top 15 radiomic features across all pipelines that corresponded to a p-value <0.01 in at least one analysis. Results are presented in suppl fig 2 and 3 for internal and external datasets respectively. For the internal dataset, these showed that ComBat tended to attenuate associations among these features, whereas the application of SUSAN had mixed results (suppl figure 2).

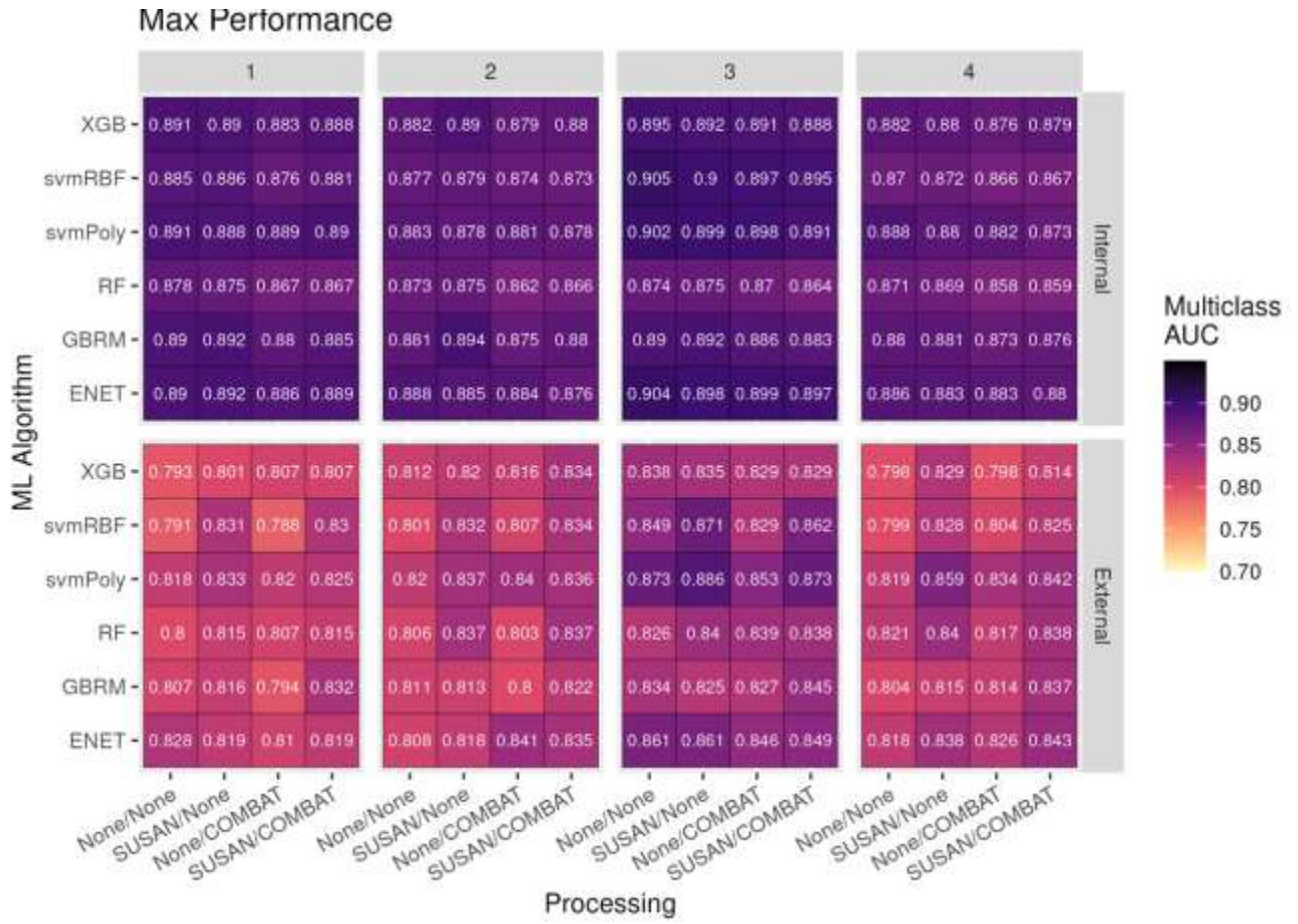


Fig 3: Maximum multi-class AUC heatmaps for the internal and external data for the various models based on the ML algorithm.

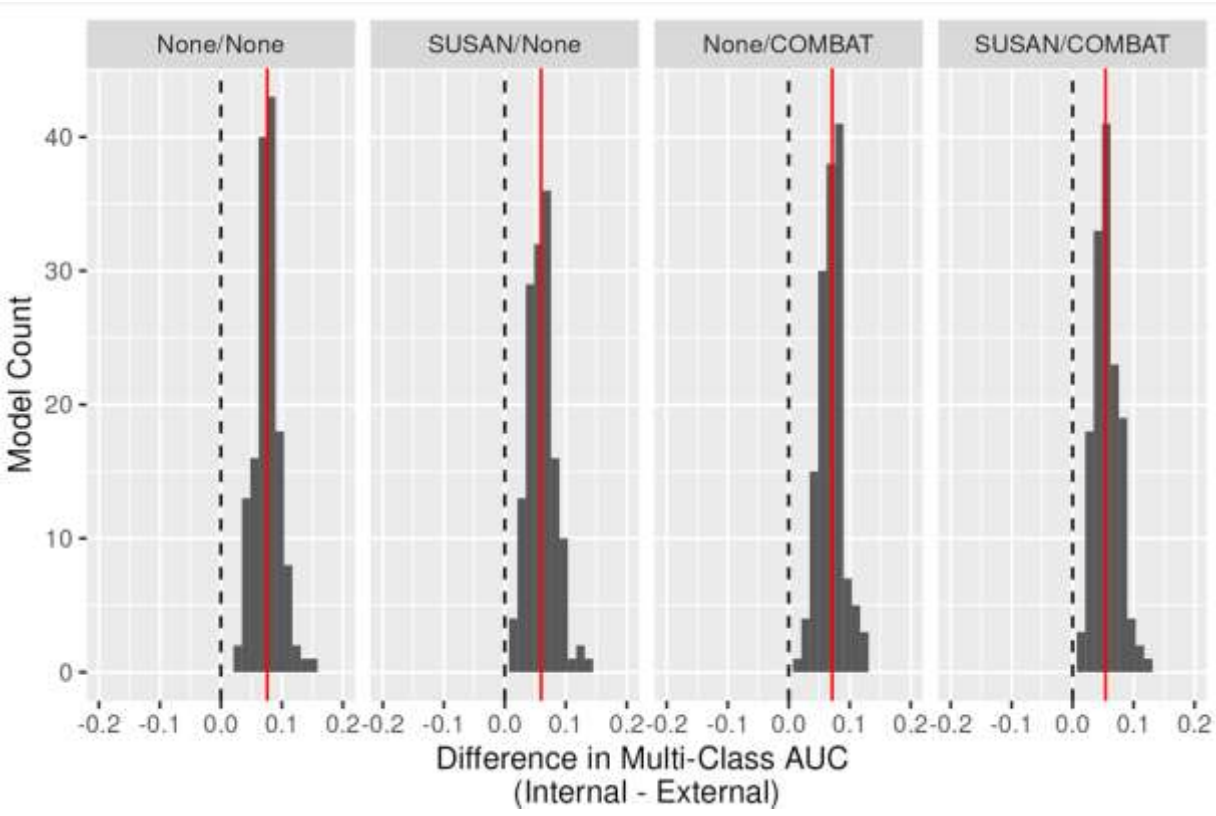


Fig 4: Histogram plot showing mean differences between internal and external data model performance for the different pipelines. The red line depicts the mean difference in model performance.

To evaluate any differences in prediction performance between disease classes (GB vs IMD vs PCNSL), we evaluated individual 1-vs-rest AUC values by class. Violin plots (suppl fig 4) suggests that there may be some evidence that SUSAN helps with PCNSL-specific discrimination performance, but this is difficult to assess systematically, given the additional confounders.

Table 3: Top performing model for each feature set, along with the three other data preprocessing results using the same modeling strategy (external data).

Feature Set	Processing	Algorithm	Feature Selection	mAUC	LogLoss	Brier Score	Best Model
CE_ET and F_PTR	*None/None*	*SVM-P*	*ICC*	*0.818*	*0.929*	*0.548*	*FALSE*
	SD/None	**SVM-P**	**ICC**	**0.833**	**0.871**	**0.521**	**TRUE**
	None/COMBAT	SVM-P	ICC	0.808	1.029	0.588	FALSE
	SD/COMBAT	SVM-P	ICC	0.817	0.949	0.564	FALSE
CE_ET and T2_PTR	*None/None*	*ENET*	*none*	*0.808*	*0.904*	*0.520*	*FALSE*
	SD/None	ENET	none	0.817	0.867	0.499	FALSE
	None/COMBAT	**ENET**	**none**	**0.841**	**0.922**	**0.492**	**TRUE**
	SD/COMBAT	ENET	none	0.835	0.891	0.487	FALSE
CE_ET, A_ET and F_PTR	*None/None*	*SVM-P*	*ICC*	*0.873*	*0.764*	*0.433*	*FALSE*
	SD/None	**SVM-P**	**ICC**	**0.886**	**0.712**	**0.414**	**TRUE**
	None/COMBAT	SVM-P	ICC	0.836	0.872	0.520	FALSE
	SD/COMBAT	SVM-P	ICC	0.873	0.749	0.444	FALSE
CE_ET	*None/None*	*SVM-P*	*ICC*	*0.819*	*0.881*	*0.499*	*FALSE*
	SD/None	**SVM-P**	**ICC**	**0.859**	**0.789**	**0.472**	**TRUE**
	None/COMBAT	SVM-P	ICC	0.821	0.962	0.531	FALSE
	SD/COMBAT	SVM-P	ICC	0.842	0.850	0.512	FALSE

[The top performing model for each feature set is highlighted in bold. The models using BIP only, but with otherwise the same modelling strategy are in italics. ENET: Multinomial elastic net, ICC: Inter-class correlation, GBRM: generalized boosted regression model, LASSO: least absolute shrinkage and selection operator, PCA: principal component analysis, SD: SUSAN denoising, SVM-P: support vector machine- polynomial kernel, SVM-RBF: support vector machine-gaussian kernel]

Finally, bar plots (suppl fig 5) were constructed to quantify the sources of variability attributable to class and batch (i.e., scanner type) on the multivariate radiomic feature distributions using principal variance component

analysis. These showed that features derived from CE_ET mask in general trended towards stronger class 'signal'.

DISCUSSION

In this study, we investigated the impact of pre-processing steps (SUSAN denoising and ComBat harmonization) on the eventual classification performance in a three-class (GB vs IMD vs PCNSL) problem involving malignant brain tumors. This was done across four mask-sequence combinations using several ML pipelines and feature reduction methods. We found that even though the mean mAUC across the various pipelines was similar (BIP, BIP+ SD, BIP+ComBat, BIP+SD+ComBat), several of the top three models across all mask-sequence combinations on the external test set used SUSAN denoising. Similarly, four of the top models across the various mask-sequence combinations used ComBat harmonization on the external dataset, including all three top performing models which used a peri-tumoral mask derived from T2WI. This is best exemplified on Table 3 which shows the model performance of the top model for each feature set, as well as other models using the same ML algorithm and feature reduction techniques but variable denoising and ComBat applications.

These findings contrast with the top performing models on the internal dataset, where most of the models used neither SUSAN denoising nor ComBat harmonization. Our findings suggest that the use of pre-processing pipelines such as SUSAN denoising and ComBat harmonization, may be more helpful in a multi-institutional setting and possibly helpful in improving model generalizability. A precise explanation of how these image pre-processing steps impacted model performance is however difficult to separate out, given the multiple confounders. This is partly because the classification performance is also considerably affected by the specific feature selection techniques and ML model used. Additionally, comparison with 'baseline' pre-processing also muddies the waters in the sense that the baseline steps such as resampling and intensity normalization by themselves can affect radiomics features and therefore impact classification performance. A few potential insights into the impact of image pre-processing and model performance may be obtained through suppl fig 2-4 which evaluate the impact of image pre-processing on important radiomic features as well as class-wise impact of image pre-processing on 1-vs-others AUC. From suppl fig 2, it is evident that pre-processing steps can variably alter the feature importance of various radiomic features, which can potentially impact how they are valued in ML-pipelines. Additionally, as shown in suppl fig 4, SUSAN denoising can potentially improve the model classification performance for PCNSL regardless of the sequences used, which likely also impacted the overall model performance.

Another takeaway from the study is that even though models using three mask-sequence combinations tended to perform marginally better for GB vs IMD vs PCNSL, the performance was overall similar to models using data from the CE_ET mask only. As shown in suppl fig 5, the radiomics features of the CE_ET mask tend to more dependent on underlying disease class rather than those from other sequences, which may partly explain why T1-CE derived feature may perform comparably to other multi-parametric sequence derived models. This is in line with previous studies and may imply that using a single mask-sequence combination may yield similar results and be easier to implement logistically in the clinical setting.[1, 2]

Previous studies have evaluated the impact of pre-processing steps on the radiomics features, generally in terms

of feature robustness and reproducibility. Several of these studies have been performed on phantoms or healthy volunteers and primarily focused on identifying reproducible features.[7, 9, 23] Some of the prior studies also used patient level data and assessed the impact of image pre-processing steps on evaluating patient survival, glioma grades or impact on tumor sub-regions.[12, 13, 24, 25] However, none of the prior studies, to the best of our knowledge have extensively evaluated the impact of pre-processing steps on eventual classification performance in a multi-class problem (of GB vs IMD vs PCNSL) in neuro-oncology. Even though the variation in radiomic features with differences in sequence parameters, vendors, scanner strength and slice thickness are known, their impact on eventual classification problem in a GB vs IMD vs PCNSL scenario remains less well explored.

In the current study, models using SD or ComBat on the internal dataset did not outperform models using neither of these pre-processing steps. A potential explanation for this may be that with in the same institution, there is limited protocol and scanner heterogeneity and the effect of the additional pre-processing steps on model performance may be negligible. Additionally, the BIP in our study involved resampling and intensity normalization. Bologna et. al, previously noted that image-preprocessing steps such as normalization, resampling, gaussian filtering and bias field correction improved the stability of features on the T1 and T2WI on phantom data.[6] Similarly, Carre et. al, noted that intensity normalization considerably improved the robustness of first-order features and subsequent model classification performance for glioma grading.[24] Similarly, Li et. al. and Um et al., noted that resampling voxels to 1x1x1 mm could remove some of the scanner effects.[11, 12] Since our BIP included resampling and image normalization, it is possible that the additional benefits of further pre-processing were not apparent on internal dataset which is expected to be less heterogeneous.

On the other hand, in the external dataset, the additional post-processing steps, especially SUSAN denoising were likely useful across various mask-sequence combinations. All top three models using the T2_PTR mask used ComBat on the external dataset, which may suggest that ComBat harmonization may be more important in models using features derived from T2WI in a multi-institutional setting. As shown in suppl fig 2, ComBat does seem to disproportionately improve the feature importance of T2_PTR derived sequences which may help explain why it was useful. One difference between the two institutions is that the internal dataset was acquired before, and the external dataset was acquired after contrast injection. Though it does not visibly change the appearance of the T2 based images, it is possible that it may affect the underlying radiomics features.

Limitations of our study include the retrospective nature and a modest sample size. We also did not evaluate the effect of other pre-processing steps such as bias-field correction or various types of image and feature normalization methods, including more recently described deep learning approaches.[26, 27] Such a task would further complicate the current analysis by introducing the confounding effect of additional variables which may be better evaluated separately in future studies. Our choice of choosing only noise filtering/ SD and Combat harmonization was based on selecting a pre-processing step that works at image level and another that works at radiomic feature level. We also did not perform any bias-field correction as part of BIP on our data. However, both Um et. al., and Li et. al., noted that bias field correction had no impact on radiomics feature reproducibility in their analysis where there were no obvious bias-field effects on the MRI images. [11, 12] Next, we also did not

compare our performance to expert readers since the primary focus of the study was to assess the impact of image pre-processing on eventual model performance. Additionally, given that feature selection and ML models work in different ways and can have considerable variability by themselves, independent of the image pre-processing steps, an exhaustive assessment of each ML pipeline was beyond the scope of current work. We therefore focused on broad trends in model classification performance instead of trying to select a clear winner. Finally, not all cases of IMD in our study were pathologically proven since this is not practically feasible in a clinical setting. We therefore relied on the availability of additional imaging and follow up data, including clinical records and institutional cancer registries to identify IMD patients.

CONCLUSIONS

Imaging pre-processing steps such as SD and ComBat harmonization may help achieve marginally improved classification performance in a multi-institutional setting. Their impact is likely negligible in a single institution setting where scanner and protocol heterogeneity are likely limited. Finally, models derived from multi-parametric MRI show similar classification performance to models derived only from the T1-CE sequences.

REFERENCES

1. Bathla, G., et al., *AI-based classification of three common malignant tumors in neuro-oncology: A multi-institutional comparison of machine learning and deep learning methods.* J Neuroradiol, 2023.
2. Bathla, G., et al., *Radiomics-based differentiation between glioblastoma and primary central nervous system lymphoma: a comparison of diagnostic performance across different MRI sequences and machine learning techniques.* Eur Radiol, 2021. **31**(11): p. 8703-8713.
3. Priya, S., et al., *Radiomic Based Machine Learning Performance for a Three Class Problem in Neuro-Oncology: Time to Test the Waters?* Cancers (Basel), 2021. **13**(11).
4. Soni, N., S. Priya, and G. Bathla, *Texture Analysis in Cerebral Gliomas: A Review of the Literature.* AJNR Am J Neuroradiol, 2019. **40**(6): p. 928-934.
5. Joo, B., et al., *Fully automated radiomics-based machine learning models for multiclass classification of single brain tumors: Glioblastoma, lymphoma, and metastasis.* J Neuroradiol, 2023. **50**(4): p. 388-395.
6. Bologna, M., V. Corino, and L. Mainardi, *Technical Note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain.* Med Phys, 2019. **46**(11): p. 5116-5123.
7. Mayerhoefer, M.E., et al., *Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study.* Med Phys, 2009. **36**(4): p. 1236-43.
8. Buch, K., et al., *Quantitative variations in texture analysis features dependent on MRI scanning parameters: A phantom model.* J Appl Clin Med Phys, 2018. **19**(6): p. 253-264.
9. Lee, J., et al., *Radiomics feature robustness as measured using an MRI phantom.* Sci Rep, 2021. **11**(1): p. 3973.
10. Orlhac, F., et al., *How can we combat multicenter variability in MR radiomics? Validation of a correction procedure.* Eur Radiol, 2021. **31**(4): p. 2272-2280.
11. Li, Y., et al., *Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features.* Cancers (Basel), 2021. **13**(12).
12. Um, H., et al., *Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets.* Phys Med Biol, 2019. **64**(16): p. 165011.
13. Moradmand, H., S.M.R. Aghamiri, and R. Ghaderi, *Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma.* J Appl Clin Med Phys, 2020. **21**(1): p. 179-190.
14. Smith, S.M. and J.M. Brady, *SUSAN—a new approach to low level image processing.* International journal of computer vision, 1997. **23**(1): p. 45-78.
15. Stamoulou, E., et al. *ComBat harmonization for multicenter MRI based radiomics features*. in *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*. 2021. IEEE.
16. Orlhac, F., et al., *A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies.* J Nucl Med, 2022. **63**(2): p. 172-179.
17. 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, 2010. **29**(12): p. 2023-37.
18. van Griethuysen, J.J.M., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.
19. Fortin, J., *Harmonization of Multi-Site Imaging Data with ComBat, R Package Version 1.0. 9. neuroCombat. 2021*.

20. Bathla, G., et al., *Differentiation Between Glioblastoma and Metastatic Disease on Conventional MRI Imaging Using 3D-Convolutional Neural Networks: Model Development and Validation.* Acad Radiol, 2023.
21. Hand, D.J. and R.J. Till, *A simple generalisation of the area under the ROC curve for multiple class classification problems.* Machine learning, 2001. **45**: p. 171-186.
22. Team, R.C., *R: A language and environment for statistical computing.* 2013.
23. Ford, J., et al., *Quantitative Radiomics: Impact of Pulse Sequence Parameter Selection on MRI-Based Textural Features of the Brain.* Contrast Media Mol Imaging, 2018. **2018**: p. 1729071.
24. Carre, A., et al., *Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics.* Sci Rep, 2020. **10**(1): p. 12340.
25. Salome, P., et al., *MR Intensity Normalization Methods Impact Sequence Specific Radiomics Prognostic Model Performance in Primary and Recurrent High-Grade Glioma.* Cancers (Basel), 2023. **15**(3).
26. Cackowski, S., et al., *comBat versus cycleGAN for multi-center MR images harmonization.* 2021.
27. Mali, S.A., et al., *Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods.* J Pers Med, 2021. **11**(9).

SUPPLEMENTAL FILES

1. MRI Imaging parameters: Internal data

A. MRI scanning parameters for Siemens 1.5 T MRI (Siemens Aera/Avanto, Erlangen, Germany)

- T1W (TR/TE/TI: 1950/10/840, NEX: 2, slice thickness: 5 mm, matrix: 320 × 256, field of view (FOV) 240 mm, pixel size 0.75 mm)
- T2W (TR/TE: 4000/90, NEX: 2, slice thickness: 5 mm, matrix: 512 × 408, FOV 240 mm, pixel size 0.5 mm);
- FLAIR (TR/TE/TI: 9000/105/2500, NEX: 1, slice thickness: 5 mm, matrix: 384 × 308, FOV 240 mm, pixel size 0.6 mm)
- DWI (TR/TE: 4000/74, NEX: 3, slice thickness: 5 mm, matrix: 128 × 128, FOV 240 mm, pixel size 1.8 mm)
- T1W-CE (TR/TE/TI: 570/13/232, NEX: 2, slice thickness: 5 mm, matrix: 384 × 312, field of view (FOV) 240 mm, pixel size 0.62 mm)

B. MRI scanning parameters for Siemens 3 T MRI (Siemens Skyra, Erlangen, Germany)

- T1W (TR/TE/TI: 2000/11/899, NEX: 1, slice thickness: 5 mm, matrix: 384 × 312, field of view (FOV) 240 mm, pixel size 0.62 mm)
- T2W (TR/TE: 4000/105, NEX: 2, slice thickness: 5 mm, matrix: 448 × 364, FOV 240 mm, pixel size 0.5 mm)
- FLAIR (TR/TE/TI: 9000/108/2500, NEX: 1, slice thickness: 5 mm, matrix: 384 × 312, FOV 240 mm, pixel size 0.62 mm)
- DWI (TR/TE: 4250/64, NEX: 1, slice thickness: 5 mm, matrix: 160 × 160, FOV 240 mm, pixel size 1.5 mm)
- T1W-CE (TR/TE/TI: 2000/12/900, NEX: 1, slice thickness: 5 mm, matrix: 384 × 312, field of view (FOV) 240 mm, pixel size 0.62 mm)

2. MRI Imaging parameters: External data

A: **MRI scanning parameters for Siemens 1.5 T MRI**

- T1W (TR/TE/TI: 2200/3.02/900, NEX: 1, slice thickness: 1.0 mm, matrix: 263 × 350, field of view (FOV) 250 mm, pixel size 1.0 mm)
- T2W (TR/TE: 4710/90, NEX: 1, slice thickness: 3 mm, matrix: 263 × 350, FOV 230 mm, pixel size 0.6 mm);

- FLAIR (TR/TE/TI: 10060/133/2550, NEX: 1, slice thickness: 3 mm, matrix: 180 x 240, FOV 240 mm, pixel size 0.4 mm)
- DWI (TR/TE: 6700/95, NEX: 1, slice thickness: 2.5 mm, matrix: 230 × 230, FOV 230 mm, pixel size 1.8 mm)
- T1W-CE (TR/TE/TI: 2200/3.02/900, NEX: 1, slice thickness: 1.0 mm, matrix: 263 × 350, field of view (FOV) 250 mm, pixel size 1.0 mm)

B: MRI scanning parameters for Siemens 3 T MRI

- T1W (TR/TE/TI: 2200/2.46/900, NEX: 1, slice thickness: 0.9 mm, matrix: 264 × 350, field of view (FOV) 250 mm, pixel size 1.0 mm)
- T2W (TR/TE: 4710/90, NEX: 1, slice thickness: 3 mm, matrix: 263 × 350, FOV 240 mm, pixel size 0.8 mm);
- FLAIR (TR/TE/TI: 10060/133/2550, NEX: 1, slice thickness: 3 mm, matrix: 180 x 240, FOV 240 mm, pixel size 0.4 mm)
- DWI (TR/TE: 6700/95, NEX: 1, slice thickness: 3 mm, matrix: 220 × 220, FOV 220 mm, pixel size 1.7 mm)
- T1W-CE (TR/TE/TI: 2200/2.46/900, NEX: 1, slice thickness: 0.9 mm, matrix: 264 × 350, field of view (FOV) 250 mm, pixel size 1.0 mm)

2. Image preprocessing:

Following anonymization, the DICOM images were initially converted to NIfTI format. All MRI images were subsequently pre-processed using the following pipeline: (a) resampling (1x1x1 mm3); (b) reorientation to the left-posterior-superior (LPS) coordinate system; (c) alignment of T1WI (pre-contrast) to the SRI24 atlas using an affine registration technique.[1] (d) co-registration of all MRIs sequences of each patient to the corresponding T1 pre-contrast images using affine registration; (d) skull stripping; (e) intensity normalization to [0,255], and (f) denoising to reduce the effects of noisy high-frequency features using the SUSAN technique[2]. The resampling, reorientation, registration, and normalization steps were implemented using ANTsPy version 0.2.9, a python library that wraps the C++ biomedical image processing library Advanced Normalization Tools (ANTs). [3] The Nipype python package version 1.7.0 provides an interface to the FSL implementations of the skull stripping techniques.[4-7] Cancer Imaging Phenomics Toolkit (CaPTk), which is an NIH-funded open-source research software was used from GitHub repository (https:// github.com/CBICA/CaPTk) for SUSAN denoising.[8] Both denoised and non-denoised images were subsequently used for feature extraction (BIP±SD)

3. Details of extracted features per mask:

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 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.

4. Feature harmonization:

Batch correction for scanner type was performed using the R package neuroCombat under default settings using all observed values, performed separately by sequence/mask combination. These steps were first applied to the internal data to derive the ComBat parameters for harmonization. The trained harmonization models were then applied to the internal, resegmented and external datasets for batch correction.

5. **Feature selection:**

To eliminate redundancy and/or reduce feature dimensionality, the following feature selection methods were considered on the batch corrected and imputed data (in addition to no a priori feature selection):

1. Linear combinations filter. 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.
2. High correlation filter. 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 (set to 0.90).
3. Principal components analysis (PCA). The number of components retained in the PCA transformation is determined by specifying the fraction of the total variance that should be covered by the components (set to 0.90).
4. LASSO: Supervised feature selection is performed on the training data using the glmnet R package. 10-fold cross-validation is used to select the optimal tuning parameters based on minimized error (i.e., more permissive than 1se rule). A final LASSO model will then trained and all features with non-zero coefficients will be passed to the model training phase.
5. ICC: The re-segmentation data were used to calculate intra-class correlation coefficients (ICC's) for all radiomics features using the one-way agreement definition via the irr R package. Given the limited sample size of re-segmented tumors, all available data were used regardless of the identity of samples in specific training/testing datasets during (nested) cross-validation. This was justified as has having limited data leakage impact on modeling due to its unsupervised filtering nature. The selected threshold for ICC-based filtering was ICC<0.75.

These feature selection methods were implemented on the training data only, using the recipes package in R unless otherwise indicated.

6. Model training:

For each tumor, features extracted using two different masks (enhancing and edema) and four sequences (A, CE, F, and T2) were considered as features for multi-class classification model training. A total of 4 unique sequence-mask radiomics feature combinations were considered for model training, ranging from 1 to 3 combined sets, defined in the table below:

Feature Sets	Number of features
CE_ET, F_PTR	214
CE_ET, T2_PTR	214
CE_ET, A_ET, F_PTR	321
CE_ET	107

Models were fit on the largest lesion identified per patient, and in contrast to the prior analysis we will NOT use total number of identified lesions as predictive feature. This is because it is highly predictive in its own right and may heavily mask underlying effects of the experimental conditions on radiomic modeling performance via "ceiling" effects.

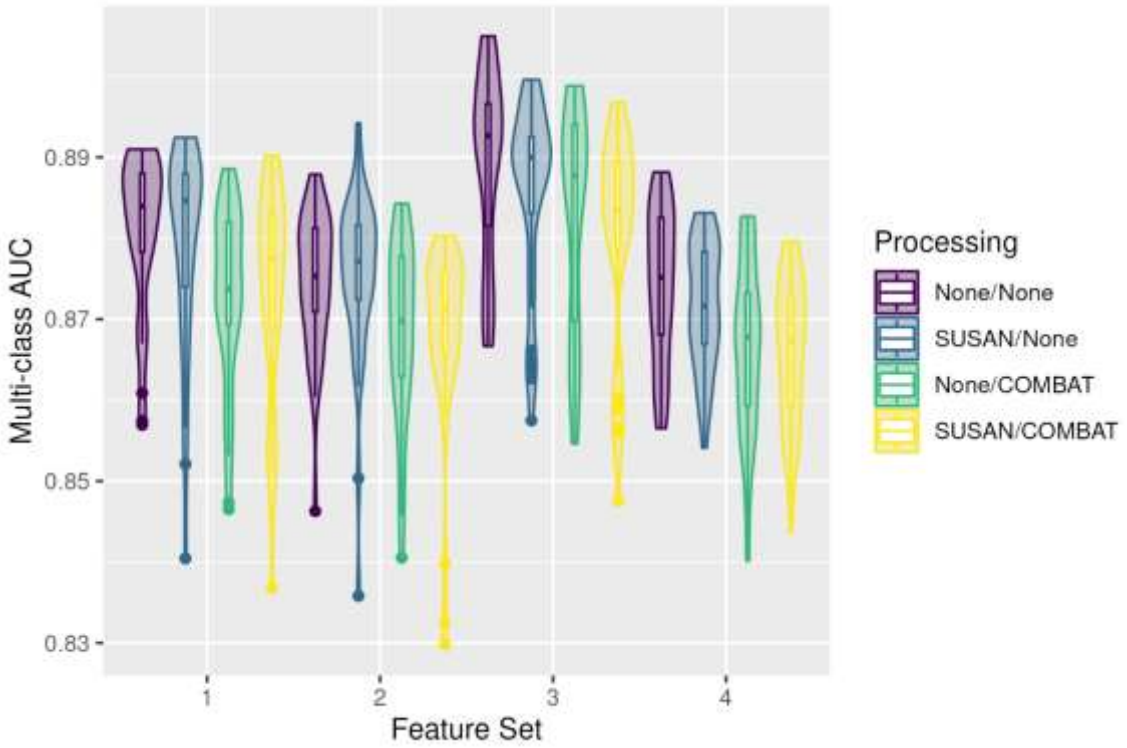
ML algorithms under consideration included:

1. Kernel support vector machines using the polynomial (SVM-P) kernel
2. Kernel support vector machines using the Gaussian (SVM-RBF) kernels
3. Multinomial elastic-net (ENET)

4. Extreme gradient boosting (XGB)
5. Generalized boosted regression models (GBRM)
6. Random forest (RF)

Training was performed using the recipes and caret R packages for non-XGB models. Bayesian model optimization was used to tune XGB hyperparameters using the mlrMBO R package.[9]

Nested cross-validation (NCV; 5-fold) was used for general model training and evaluation within the internal dataset, such that cross-validated measure of model performance will be generated from the internal dataset. The model training for hyperparameter tuning was based on 10-fold cross-validation. Final models were then trained on the full internal dataset for application to the external dataset.



Suppl Fig-1: Violin plots for all four feature sets using the internal data show the range of multi-class AUC across different pipelines. Feature set 1: CE_ET and F_PTR; 2: CE_ET and T2_PTR; 3: CE_ET, A_ET and F_PTR; 4: CE_ET only.

Suppl table 1: Summary of top 3 performing models in the internal dataset for each feature set. ENET: Multinomial elastic net, ICC: Inter-class correlation, GBRM: generalized boosted regression model, LASSO: least absolute shrinkage and selection operator, PCA: principal component analysis, SD: SUSAN denoising, SVM-P: support vector machine- polynomial kernel, SVM-RBF: support vector machine-gaussian kernel, XGB: extreme gradient boosting.

Feature Set	Processing	Algorithm	Feature Selection	mAUC	LogLoss	Brier Score
CE_ET and F_PTR	SD/None	ENET	LASSO	0.892	0.562	0.308
	SD/None	GBRM	LASSO	0.892	0.565	0.313
	SD/None	ENET	ICC	0.892	0.576	0.317
CE_ET and T2_PTR	SD/None	GBRM	Corr	0.894	0.557	0.317
	SD/None	XGB	Corr	0.890	0.558	0.310
	None/None	ENET	LASSO	0.888	0.586	0.325
CE_ET, A_ET and F_PTR	None/None	SVM-RBF	PCA	0.905	0.543	0.299
	None/None	ENET	LASSO	0.904	0.546	0.305
	None/None	ENET	None	0.903	0.551	0.307
CE_ET	None/None	SVM-P	ICC	0.888	0.582	0.313
	None/None	ENET	LASSO	0.886	0.589	0.325
	None/None	ENET	ICC	0.886	0.592	0.328

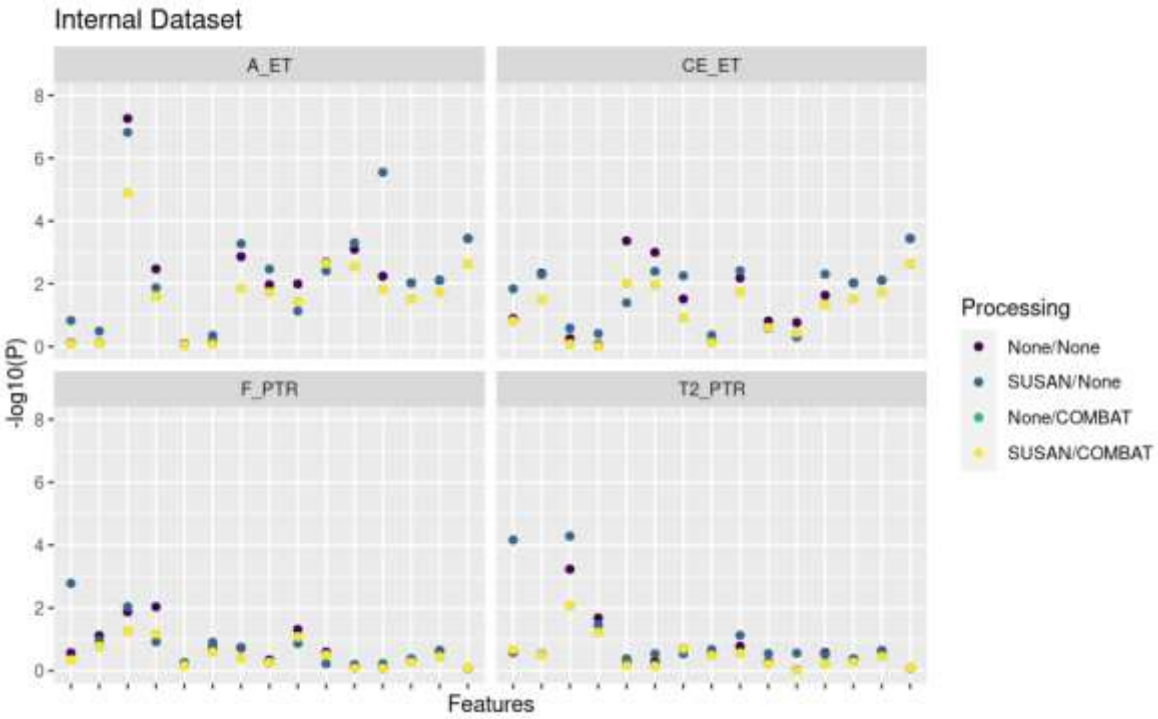
Suppl table 2: Bootstrapping of mAUC performed using the various image pre-processing pipelines. Pairwise contrasts of the pre-processing combination performances do not reveal any significant differences between pipelines.

Image pre-processing pipelines	Estimate	LL95	UL95
(SUSAN/COMBAT) - (SUSAN/None)	0.01127	-0.00505	0.03006
(SUSAN/COMBAT) - (None/COMBAT)	0.00055	-0.00353	0.00534
(SUSAN/COMBAT) - (None/None)	-0.00724	-0.02953	0.01436
(SUSAN/None) - (None/COMBAT)	-0.01071	-0.02724	0.00441
(SUSAN/None) - (None/None)	-0.01851	-0.04513	0.00342
(None/COMBAT) - (None/None)	-0.00779	-0.02953	0.01394

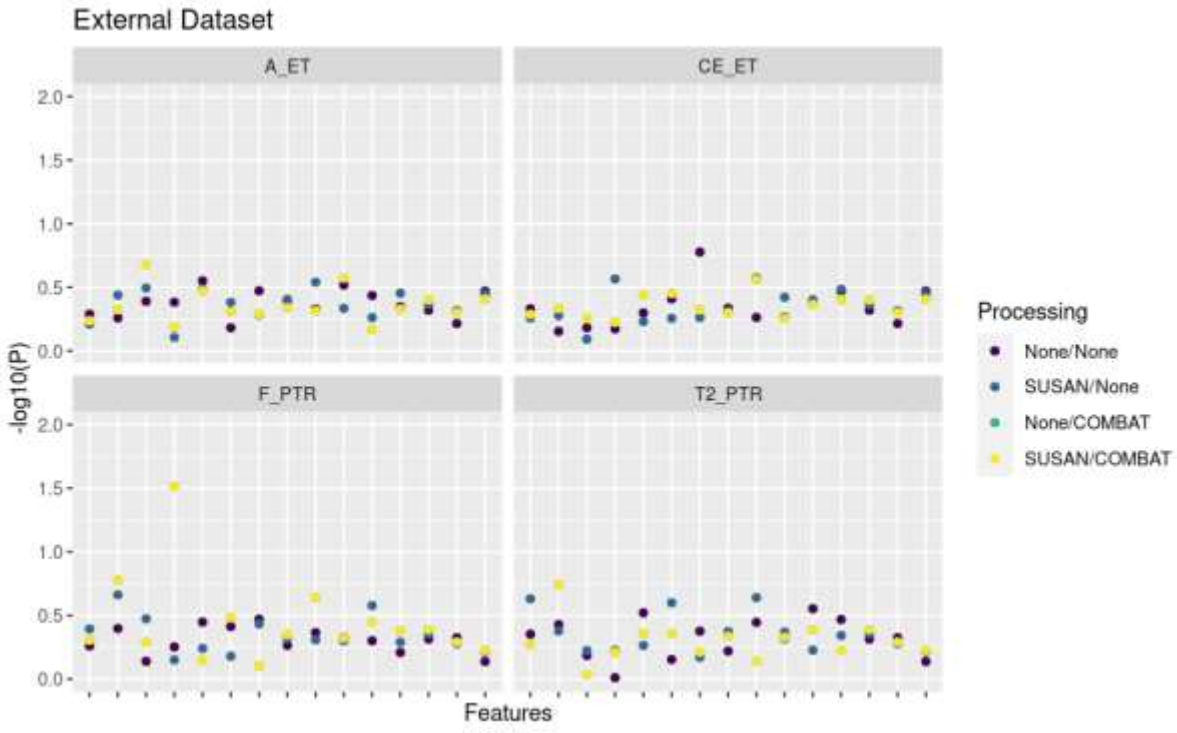
7. Impact of image pre-processing pipeline on radiomic feature importance.

Evaluating the impact of various pre-processing pipelines on feature importance is technically challenging, as different algorithms have different latent measures of importance. Moreover, different feature filtering algorithms further complicate this, as correlation-based filters will semi-arbitrarily choose a representative feature from multiple highly correlated features (e.g., the same feature from various sequence/mask combinations), and PCA-based pre-processing renders feature identity entirely moot.

One way around these complexities is to alternatively assess individual feature association with the three lesion classes via hypothesis testing, notably via the nonparametric Kruskal-Wallis test. We performed these analyses across all pre-processing conditions and identified 15 radiomics features that corresponded to a p-value <0.01 in at least one analysis. We then plotted the -log10(P) for these results, stratified by pre-processing conditions and sequence/mask combination for both the Internal and External datasets.These results are below (suppl fig 2), where higher values of -log10(P) indicate stronger evidence against the null condition of no association:



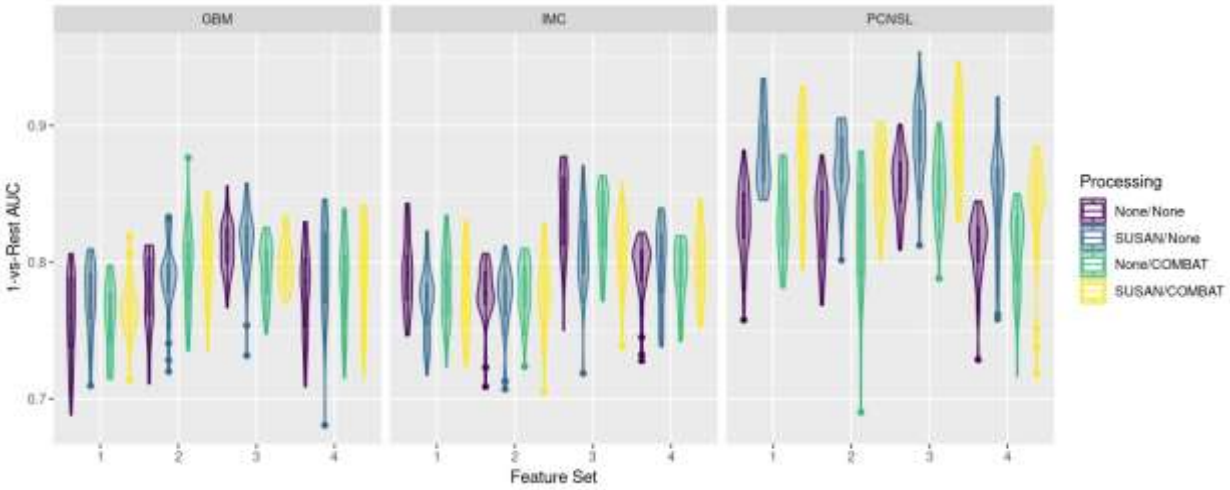
Suppl fig 2: Feature importance as a function of image pre-processing pipelines for the internal set. Here we see pre-processing conditions using ComBat tended to attenuate associations among these features, whereas the application of SUSAN had mixed results.The same type of results are plotted for the external dataset in suppl fig 3. Note that power is naturally lower given the smaller sample size, but the interpretation of higher values indicating strong association evidence remains the same. Here, we observe that the top results tend to be more heterogeneous with respect to pre-processing condition, with less of an obvious pattern than for the internal data.



Suppl fig 3: Feature importance as a function of image pre-processing pipelines for the external test set.

8. Impact of image pre-processing on lesion class.

The impact of image pre-processing on individual lesion class (GB, IMD or PCNSL) was evaluated through assessment of 1-vs-rest AUC values by class (suppl fig 4). We plotted these distributions across all fitted models as violin-plots for the external data validation. Visualization suggests there may be some evidence that SUSAN helps with PCNSL-specific discrimination performance, but this is difficult to assess systematically, given additional confounders.



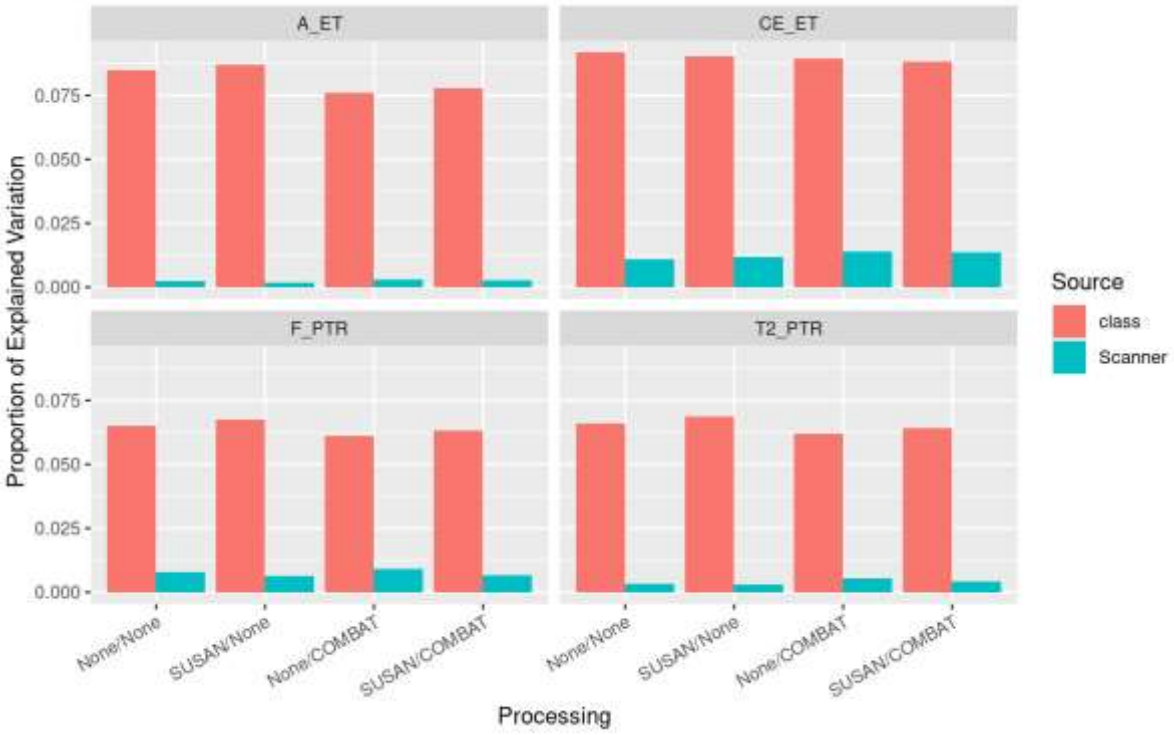
Suppl fig 4: 1-vs-rest AUC values by class, for all pre-processing pipelines and feature set combinations (([1] CE_ET and F_PTR; [2] CE_ET and T2_PTR; [3] CE_ET, A_ET and F_PTR; [4] CE_ET only).

9. Disease class and scanner-based variability in radiomic features.

To quantify the sources of variability attributable to class and batch (i.e., scanner type) on the multivariate

radiomic feature distributions, we used principal variance component analysis (PVCA). This method estimates proportions of variation corresponding to design factors along with residual error (i.e., unexplained variation or "noise"). For this analysis, we set the threshold of explained variation for number of leading PCs to be 90%.

A bar plot of the explained proportions of variation by source across sequence/mask feature-sets and processing workflows is presented in suppl fig 5. Denoising should in theory make all the bars go up (i.e., reducing residual error) while combat correction should make scanner-specific bar (i.e., batch) go down. In general, we observe fairly stable results at a broad radiomics level; moreover, CE_ET actual tends to have the strongest class "signal" among the sequence/mask combinations (albeit a relatively modest difference). Thus, attributions of SD to improvement in CE_ET-based ML modeling are likely to be more nuanced and may be due to improved resolution of particular feature(s), informative of disease class.



Suppl fig 5: Bar plots depicting sources of variability attributable to class and batch (i.e., scanner type).