Physics and Imaging in Radiation Oncology 30 (2024) 100585

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Magnetic resonance imaging radiomic features stability in brain metastases: Impact of image preprocessing, image-, and feature-level harmonization

Zahra Khodabakhshi *, Hubert Gabrys, Philipp Wallimann, Matthias Guckenberger, Nicolaus Andratschke, Stephanie Tanadini-Lang

Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Intensity normalization
Radiomics
Brain metastases
Harmonization
Magnetic resonance imaging

ABSTRACT

Background and purpose: Magnetic resonance imaging (MRI) scans are highly sensitive to acquisition and reconstruction parameters which affect feature stability and model generalizability in radiomic research. This work aims to investigate the effect of image pre-processing and harmonization methods on the stability of brain MRI radiomic features and the prediction performance of radiomic models in patients with brain metastases (BMs).
Materials and methods: Two T1 contrast enhanced brain MRI data-sets were used in this study. The first contained 25 BMs patients with scans at two different time points and was used for features stability analysis. The effect of gray level discretization (GLD), intensity normalization (Z-score, Nyul, WhiteStripe, and in house-developed method named N-Peaks), and ComBat harmonization on features stability was investigated and features with intraclass correlation coefficient >0.8 were considered as stable. The second data-set containing 64 BMs patients was used for a classification task to investigate the informativeness of stable features and the effects of harmonization methods on radiomic model performance.
Results: Applying fixed bin number (FBN) GLD, resulted in higher number of stable features compare to fixed bin size (FBS) discretization (10 ± 5.5 % higher). `Harmonization in feature domain improved the stability for non-normalized and normalized images with Z-score and WhiteStripe methods. For the classification task, keeping the stable features resulted in good performance only for normalized images with N-Peaks along with FBS discretization.
Conclusions: To develop a robust MRI based radiomic model we recommend using an intensity normalization method based on a reference tissue (e.g N-Peaks) and then using FBS discretization.

1. Introduction

Radiomics is defined as high-throughput extraction of quantitative features from medical images using advanced mathematical algorithms to perform classification or prediction tasks [1]. The hypothesis is that radiomic features may capture subtle but relevant information invisible to naked eyes [1]. As a growing field of research, radiomics has been applied to many clinical problems in oncology and has demonstrated its great potential for diagnosis, prognosis, and treatment response prediction [2–4]. However, model generalizability and reproducibility of results make the translation of this technique into clinical practice challenging [5]. Several studies have shown that radiomic feature values can be affected by differences in various parts of the radiomics workflow [6–11], such as centers, scanner manufacturers, imaging protocol, and reconstruction algorithms. These variabilities hamper data pooling from different sites and may result in biased and unreliable models.

Magnetic resonance imaging (MRI), specifically T1-weighted contrast-enhanced, is widely used for diagnosis and management of patients with brain metastases (BMs) due to its excellent soft tissue contrast and spatial resolution [12]. However, a major issue in quantitative MRI studies is the high sensitivity of MRI intensities to variations in acquisition and reconstruction parameters [13,14]. Moreover, in contrast to other imaging modalities, like CT and PET, MRI is measured in an arbitrary unit and image intensities do not have any clear physical interpretation. Consequently, there are large differences in MRI intensities even between the images of the same patient scanned with the same scanner and the same protocol [15]. The aforementioned reasons make the quantitative analysis of MRI images and the reproducibility of

* Corresponding author.
E-mail address: zahra.khodabakhshi@usz.ch (Z. Khodabakhshi).

https://doi.org/10.1016/j.phro.2024.100585
Received 12 December 2023; Received in revised form 23 April 2024; Accepted 2 May 2024
Available online 6 May 2024

MRI based radiomics features challenging.

In recent years, two main approaches were proposed to address the problem of incomparability among MRI images and reduction of the scanner effects, variability in the scanner, protocol, reconstruction algorithms, etc; these are a) harmonization in the image domain, which mainly consists of image pre-processing and MRI intensity normalization, and b) harmonization in the feature domain [5,15,16].

Several methods of MRI intensity normalization have been proposed, however, there is no consensus among studies about which method is optimal for a specific application. Some of these methods, such as Z-score normalization, histogram matching, and Nyul normalization [17] are general, tissue-nonspecific methods. On the other hand, methods such as White-Stripe normalization, is specifically applied to brain MRI and use white matter as reference tissue for normalization [15,18].

ComBat harmonization is one of the widely used methods of feature harmonization [19]. The effectiveness of ComBat harmonization in removing the imaging parameters and scanner-related variabilities and harmonizing radiomics features in different multicentric studies including PET/CT [20,21], CT [22], and MR [16] radiomics studies has been demonstrated.

Besides the harmonization methods for removing the inconsistencies related to image acquisition and reconstruction, there are other additional critical pre-processing steps in MR radiomics studies including bias field correction to minimize the intensity inhomogeneity within a tissue [23] and gray-level discretization which clusters pixels into bins based on their intensity values in order to reduce noise and computational time [24]. Several studies have shown that MRI pre-processing has a considerable impact on the reliability and repeatability of radiomics texture features [21,24]. However, investigations of the combined effect of image pre-processing and harmonization on the whole pipeline of radiomics study including radiomics feature stability and model performance are limited.

The main objective of this study is to assess the joint effect of MRI image pre-processing and harmonization methods on the stability of radiomics features based on a test–retest analysis of MRI images of a cohort of patients with BMs. We also evaluated the performance of radiomics models for the prediction of the primary site of cancer (lung vs. melanoma) in patients with BMs.

2. Materials and methods

2.1. Patient cohort

In this study, we used two retrospective MRI datasets of BMs patients who were scanned and treated with radiosurgery for newly diagnosed BMs at the University Hospital Zurich (USZ). Dataset1 was used as a test–retest dataset for evaluating the stability of radiomics features and includes 25 patients with BMs (15 females and 10 males with an average age of 60.2 years). Each patient underwent two T1-contrast enhanced MRI scans, one standard diagnostic scan for screening/detection of BMs and another dedicated scan in treatment position for radiosurgery planning purposes [25]. These scans were performed using one of the following three different scanners: Philips Healthcare-Ingenia, GE-Discovery, and GE-Signa Premier, all of which had a field strength of 3 T. The average time interval between the two scans was 17.3 ± 5.9 days. Only patients without tumor progression and tumor morphological changes were included in this dataset.

Dataset2 was used for a proof-of-principle application for classification of the primary site of cancer after applying the different normalization/harmonization methods and consisted of pre-treatment T1-weighted contrast-enhanced MRI scans of 64 patients (31 with melanoma and 33 with non-small cell lung cancer) with a total number of 104 BMs. The patients' characteristics are provided in Supplemental data (Table S-1). The Study was approved by Local Ethic Committee (BASEC-Nr. 2018-01794) and consent for retrospective analysis was obtained for both Datase1 and Dataset2.

2.2. Image preprocessing and segmentation

We used the HD-BET brain extraction tool [26] to remove the skull from MR images; the provided brain masks were used for normalization. In order to correct the intensity non-uniformity in MR images, bias field correction was applied on skull striped images via the N4itk algorithm [27]. The tumor volumes were delineated by board certified radiation oncologists according to the institutional guidelines. In both datasets, the initial slice thickness and in-plane resolution were between 0.47–0.6 mm. The images in both datasets were resized into isotropic voxel size of 0.6 mm using linear interpolation.

2.3. Image intensity normalization

One of the aims of this study was to investigate the effect of MR intensity normalization and feature harmonization on the stability of radiomics features. To this end, we applied three commonly used normalization methods including Z-score, Nyul [17,28], White-Stripe [15], and an in-house developed normalization method named N-Peaks normalization. The detailed explanation about the normalization methods used in this study is provided in the Supplementary material.

2.4. Gray-level discretization and feature extraction

To assess the combined effect of GLD and MR intensity normalization, we implemented two commonly used GLD approaches. One approach is relative GLD which clusters the intensities of pixels into a fixed number of bins (FBN). In this study, we applied five different numbers of bins which are commonly used in other studies (FBN = 16, 32, 64, 128, 256).

Another approach is absolute GLD which clusters the pixels into a fixed bin size (FBS). Since ROIs have different ranges of intensities we defined the size of bins by calculating the following scaling factor:

$$BS = \frac{Mean\ ROI\ intensity\ Range}{BN}$$

BS denotes bin size and BN denotes bin number. For each ROI the intensity range is the difference between maximum and minimum intensities. Here we calculated the average intensity range of all ROIs and then divided it by the bin numbers. We tested five different bin sizes for BN = 16, 32, 64, 128, 256, denoted as FBS16, FBS32, …, FBS256.

In total, 837 radiomics features were extracted from each ROI including first order intensity based, texture and wavelet features. Feature extraction was performed using PyRadiomics [29], a standard open source Python package for radiomic feature extraction with feature definition in compliance with the Image Biomarker Standardization Initiative (IBSI) [30]. The list of all extracted features are provided in Supplemental data (Table S-2). Since we used the same ROI mask for the test retest dataset we did not calculate shape features for test–retest analysis. However, for the classification task 14 additional shape features were extracted from the ROI.

For radiomic feature harmonization we applied the well-known ComBat harmonization [20] on our data. More explanation about ComBat harmonization is provided in Supplementary materials.

2.5. Data analysis

We used the intraclass correlation coefficient (ICC) to assess the stability of radiomics features in dataset1 and features with ICC ≥ 0.8 were considered as stable features, as this is a commonly used value in published studies [31,32]. In order to compare the proportions of stable features before and after ComBat harmonization we used McNemar's test.

To evaluate the effects of MR intensity normalization and GLD on the classification task for primary cancer sites (lung or melanoma) of BMs, dataset2 was used and four different classifiers including Logistic

Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB) were implemented. A nested cross-validation scheme with two 5-fold cross-validations as inner and outer loops was implemented to train, optimize, and test the model. Each inner and outer loop was repeated 10 times.

To reduce data dimensionality and computation time, the Spearman correlation coefficient was calculated between each pair of features and for values above 0.9, one of the features was removed. Three feature selection methods including k-best (F-score), Lasso, and variance threshold were implemented to select the most informative features. The area under the receiver operating characteristic curve (AUC) was used as the evaluation metric and finally, the average test AUC for each combination of classifier and feature selection was reported.

In order to check if the effects of different normalization methods on radiomics feature stability and the predictive performance of radiomics models are significantly different, the Friedman test followed by Conover post hoc analysis were applied on ICCs and AUCs.

3. Results

3.1. Impact of intensity normalization, feature harmonization and gray-level discretization methods on the stability of radiomics features

According to the results, increasing the bin number or decreasing the size of bins resulted in a higher percentage of stable features. For all normalization methods, images discretized with FBN had a higher percentage of stable features in comparison to images discretized with FBS. When using FBS, the Nyul method provided the highest percentage of stable features, followed by Z-score, N-Peaks, no normalization and WS. On the other hand, when using FBN, the Nyul method resulted in the lowest number of stable features, whereas the other methods provided a slightly higher percentage of stable features (Fig. 1).

More detailed analysis of the effect of GLD on features stability shows that intensity features has the lowest average ICC (ICCs between 0.21–0.48 for FBN and 0.18–0.49 for FBS) (Fig. 2). On the other hand, for all configurations, texture features category based on neighborhood gray tone difference matrix (NGTDM) are the most stable and had the highest average ICC (ICCs between 0.39–0.80 for FBN and 0.31–0.77 for FBS). The results for wavelet features are shown in Supplementary Figs. 1 and 2. A similar trend is evident for wavelet features. For different groups of wavelet features, HHL group has the most unstable features. In comparison to original features, wavelet features (specifically LLH-with FBN discretization) have higher average ICCs.

Although, none of the normalization methods could result in stable intensity based features, there is significant difference in average ICC of intensity based features between different normalization methods. Fig. 3 represents the average ICC of first-order features versus different bin numbers and bin sizes. According to this figure, for both discretization methods, N-Peaks normalization could better improve the ICC of intensity based features.

The results of the impact of ComBat harmonization on the stability of radiomic features are summarized in Table S3 and S4. Based on the results for both cases of using FBN and FBS discretization, ComBat harmonization could slightly improve the stability of radiomics features for non-normalized and normalized images with the WS and Z-score methods. In most of the cases the p-values are less than 0.05.

3.2. Impact of normalization methods on the performance of different models for classification of primary site of cancer

For the classification task, Lasso in combination with Logistic Regression, performed better than other models (Fig. 4.A). We report the results for FBN32 and FBS32 as these two discretization methods are frequently used in other studies. The Friedman test resulted in a p-value of 0.003 for the AUCs, suggesting presence of significant differences among normalization methods. Therefore, we further used the Conover post-hoc test for a deeper analysis of our data. The results of the Conover test are represented in Supplemental materials (Supplementary Fig. 3.A). According to the p-values, there are no significant differences between the normalization methods. However, the difference between no-normalized scans with FBS discretization and all other methods is significant (Pvalue < 0.05).

The critical difference plot of average score rank of different normalization methods based on AUCs of Fig. 4.A is presented in Fig. 4.B. In this plot the lower rank indicates the better performance of the normalization method and the horizontal lines connect the methods with similar performance. According to the results, WS-FBN32 normalization has better rank in comparison to other methods. In general normalization methods with FBN discretization have resulted in better ranks in comparison to FBS discretization.

All the radiomics models were trained and tested using only the stable features. Average test AUCs of different methods are presented in Fig. 4.C. The preselection of stable features significantly decreased the performance of the models for all normalization methods except for N-Peaks and Nyul normalization The Friedman test on AUCs resulted in p-value < 0.05. The further analysis with the Conover post-hoc test is represented in Supplementary Fig. 3.B. According to the critical difference plot (Fig. 4.D) N-Peaks-FBS32 achieved the best rank in comparison to other methods which indicates this method is successful in normalizing the intensities while retaining useful biological information.

4. Discussion

The robustness and stability of a radiomic feature is a key factor to it

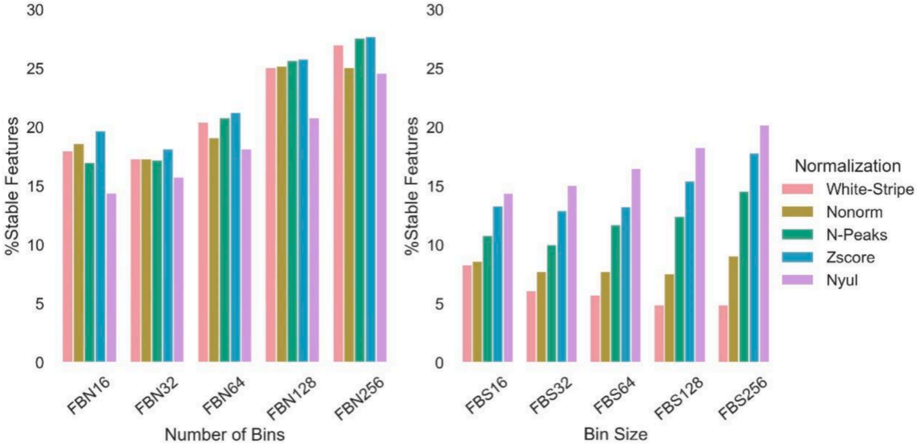

Fig. 1. Percentage of stable features for different normalization methods and gray-level discretization.

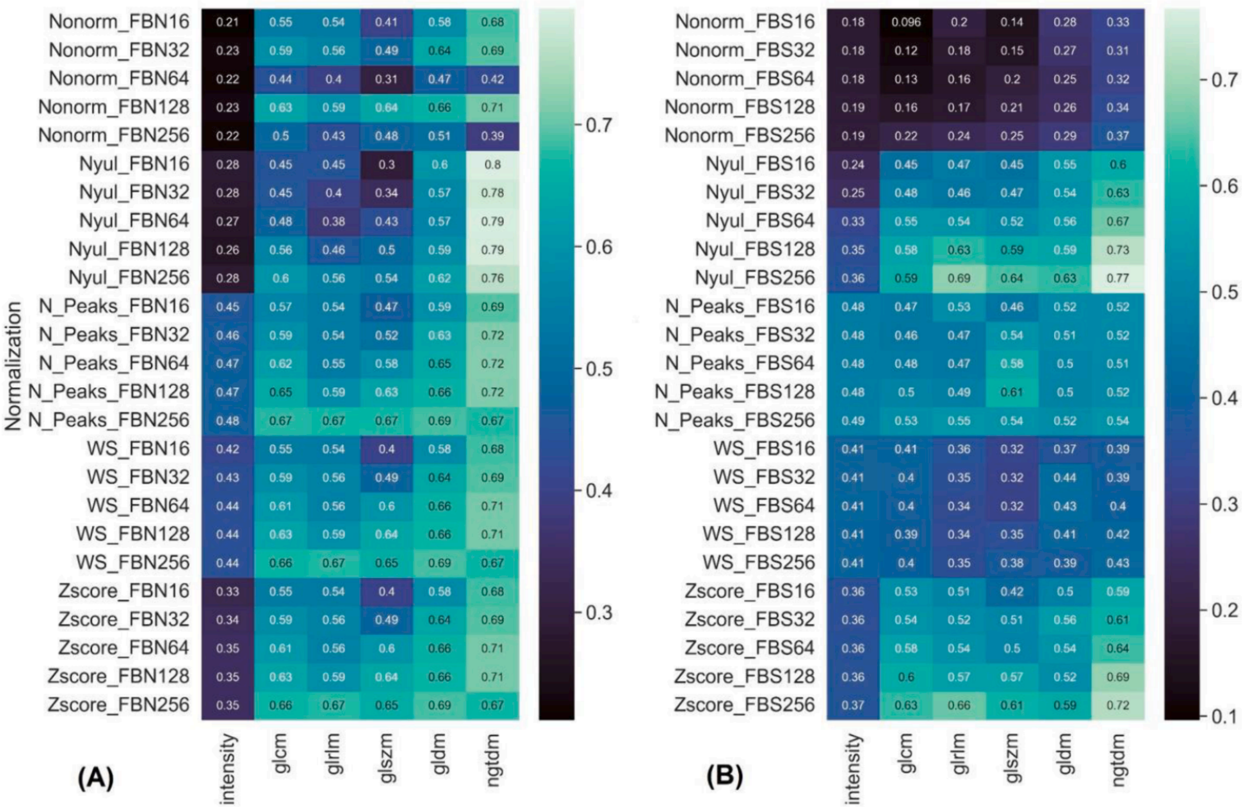

Fig. 2. Heatmap of average ICCs of original radiomics features for A) different intensity normalization methods and fixed bin number gray level discretization and B) fixed bin size gray level discretization method.

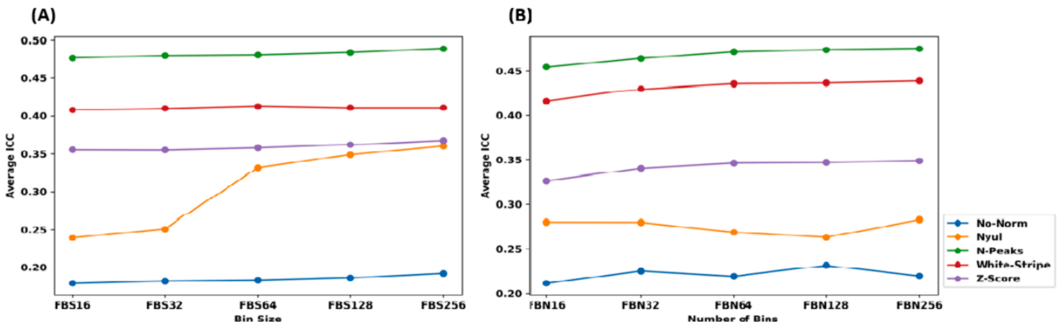

Fig. 3. Comparison the average ICC of intensity-based features between different normalization methods with A) FBS gray level discretization and B) FBN gray level discretization.

as a potential imaging biomarker. However, large number of radiomic features are sensitive to the acquisition, reconstruction parameters, and preprocessing methods [5,33–35]. The overall aim of our study was to investigate the impact of image preprocessing and both image- and feature-level harmonization on the stability of radiomics features. Furthermore, we assessed the performance of radiomic-based models for the classification of the primary site of cancer in a cohort of patients with BMs. We found that FBN results in higher number of stable features however using FBS with a tissue-based normalization leads to better results for classification.

Examination of GLD methods showed that increasing the number of bins and decreasing the bin sizes increased the percentage of stable features and, overall, using FBN resulted in more stable features in comparison to FBS. It should also be noted that the variance of the percentage of stable features between different normalization methods was higher for FBS discretization in comparison to FBN which means applying intensity normalization after FBN discretization does not have a strong benefit for radiomics features stability. Carré et al. [36] also investigated the effect of GLD and MR intensity normalization on the stability of radiomics features. They reported that using FBS discretization improved the number of robust features in all cases of applying MR intensity normalization compared to non-normalized which is consistent with the results of our study. Moreover, they reported that a higher number of bins was associated with a higher number of robust features. Based on their results, using FBN discretization, White-Stripe, and Z-score normalization resulted in a similar percentage of robust features as non-normalized images which is in line with our results. In their study, they stated that FBN discretization makes the use of MR intensity normalization unnecessary for second order radiomics features.

In another study, Li et al. [9] investigated the effect of MR intensity normalization, ComBat harmonization, and image resampling on radiomics features reproducibility. In that study, they used FBN = 32 for GLD. According to their results, the impact of intensity normalization on

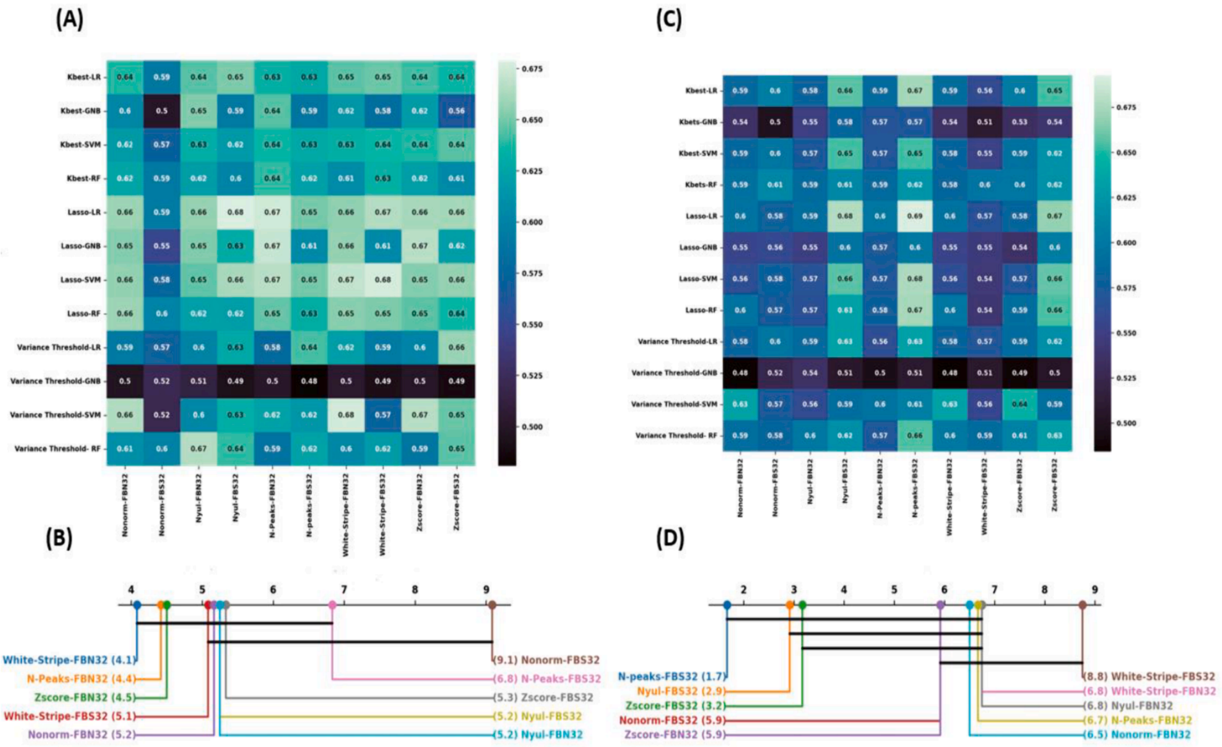

Fig. 4. Heatmap and critical difference plot of the performance of different classifiers based on extracted features from different intensity normalization methods. A) Heatmap of average test AUCs of different classifiers performance based on all radiomic features. B) Critical difference plot of average rank of different normalization methods based on section (A). C) Heatmap of average test AUCs of the performance of different classifiers based on preselection of stable features. D) Critical difference plot of average rank of different normalization methods based on the AUCs of section (C).

radiomics features reproducibility was not obvious compared to the case of non-normalized images. However, they showed that intensity normalization brings the images' intensities into the same scale and therefore makes the MR images more comparable.

In our study, N-Peaks intensity normalization could effectively improve the average ICC for first-order intensity features (from an average ICC of 0.21 to 0.48 for FBN and from 0.18 to 0.49 for FBS discretization (Fig. 3) however, none of the intensity-based features were stable and the average ICCs were lower in comparison to other feature categories. The results of Carré et al. [36] regarding the stability of intensity-based features were not consistent with our results. According to their results, applying intensity normalization increased the number of stable features (with ICC > 0.8) and the Nyul method resulted in the highest number of stable intensity-based features (16 out of 18 features were stable). One possible explanation is that the scanner variation in our study had a severe effect on image intensity ranges so that intensity normalization could significantly improve the average ICC of intensity-based features but did not make the feature stable.

Evaluation of the effect of ComBat harmonization on radiomics features stability showed that it could improve features stability for non-normalized, normalized images with WS, and Z-score normalization by a maximum of 3.4 %, and in the case of Nyul and N-Peaks normalization, it could not improve features stability in most of the cases. A study by Orlhac et al. [16] showed that applying ComBat harmonization in combination with WS normalization could significantly improve the similarity of distribution of radiomics features acquired from two different MRI scanners. In another study by Li et al. [9], the results show that ComBat harmonization can effectively improve the reproducibility of features extracted from T1-weighted MR images and remove the scanner effect. Based on their results, MR intensity normalization could make the images comparable, however, at the radiomics feature level it could not improve features reproducibility. More studies should be conducted regarding the joint effect of Intensity normalization and ComBat harmonization.

One possible approach for building robust and generalizable radiomic models is the preselection of stable and reliable features, which have the robustness to variabilities in image acquisition and reconstruction [5,37]. A significant reduction of radiomic features to be analyzed is one of the main advantages of this approach. However, there is the potential for information loss and the stable features might not necessarily contain clinically relevant information. In our study the pre-selection of stable radiomic features decreased the performance of the models however, using only stable features based on N-Peaks normalization did not decrease the performance of models which means our developed method is successful in keeping the clinical information. Compared to the other normalization techniques, this N-Peak method attempts to connect the intensity units to the tissue types more strongly. Using two normal tissue landmarks for the intensity normalization, the relationship of the intensity values with normal tissues should be stronger than when using only one tissue, as is the case for White Stripe. Furthermore, the fact that no width (or standard deviation) of an intensity distribution is used means that the technique is less subject to heterogeneity or noise in the region of interest, which would affect the standard deviations of intensity used for the Z-Score and White Stripe normalization techniques.

While IBSI recommends FBN discretization as preprocessing for MRI images, it should be considered that this method detached the relation between image intensities and physiological information while FBS discretization preserves it [30]. Therefore, using MRI intensity normalization with FBS gray level discretization seems to be a more reasonable pre-processing method.

The findings of our study show the effectiveness of harmonization methods in increasing the stability of contrast-enhanced T1-weighted radiomic features in patients with brain metastases. In the case of using

FBS discretization, applying MR intensity normalization is mandatory. In order to develop a generalizable radiomic model we recommend using FBS discretization with a tissue-based intensity normalization method which can preserve the biological information.

Our study also bears some inherent limitations. First, only contrast-enhanced T1-weighted scans have been considered. However, the joint effect of image preprocessing and normalization on other sequences such as T2-weighted and T2 FLAIR may be different. Secondly, we had a relatively small dataset for the classification performance evaluation. However, we used nested cross-validation, which yields an unbiased estimate of a model's generalization performance [38].

Acknowledgment

This study was supported by the Clinical Research Priority Program (CRPP) Grant on "Artificial Intelligence in Oncological Imaging", University of Zurich.

CRediT authorship contribution statement

Zahra Khodabakhshi: Investigation, Data curation, Methodology, Formal analysis, Writing – original draft. **Hubert Gabrys:** Supervision, Methodology, Writing – review & editing. **Philipp Wallimann:** Data curation, Methodology, Writing – review & editing. **Matthias Guckenberger:** Resources, Writing – review & editing. **Nicolaus Andratschke:** Resources, Conceptualization, Supervision, Writing – review & editing. **Stephanie Tanadini-Lang:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2024.100585.

References

[1] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563–77. https://doi.org/10.1148/radiol.2015151169.

[2] Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O'Connor JPB. Radiomics in oncology: a practical guide. Radiographics 2021;41(6):1717–32. https://doi.org/10.1148/rg.2021210037.

[3] Khodabakhshi Z, Amini M, Mostafaei S, Haddadi Avval A, Nazari M, Oveisi M, et al. Overall survival prediction in renal cell carcinoma patients using computed tomography radiomic and clinical information. J Digit Imaging 2021;34(5):1086–98. https://doi.org/10.1007/s10278-021-00500-y.

[4] Zhu X, Dong D, Chen Z, Fang M, Zhang L, Song J, et al. Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. Eur Radiol 2018;28(7):2772–8. https://doi.org/10.1007/s00330-017-5221-1.

[5] Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 2020;65(24):24TR02. https://doi.org/10.1088/1361-6560/aba798.

[6] Shiri I., Hajianfar G., Sohrabi A.Abdollahi H, P Shayesteh S, Geramifar P, et al. Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: test-retest and image registration analyses. Med Phys 2020;47(9):4265–4280. doi:10.1002/mp.14368.

[7] Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol 2017;56(11):1544–53. https://doi.org/10.1080/0284186X.2017.1351624.

[8] Saltybaeva N., Tanadini-Lang S., Vuong D. Burgermeister S, Mayinger M, Bink A, et al. Robustness of radiomic features in magnetic resonance imaging for patients with glioblastoma: multi-center study. Phys Imaging Radiat Oncol 2022;22:131–136. doi:10.1016/j.phro.2022.05.006.

[9] Li Y, Ammari S, Balleyguier C, Lassau N, Chouzenoux E. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features. Cancers (Basel) 2021;13(12):3000. https://doi.org/10.3390/cancers13123000.

[10] Crombé A, Kind M, Fadli D, Le Loarer F, Italiano A, Buy X, et al. Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients. Sci Rep 2020;10(1):15496. https://doi.org/10.1038/s41598-020-72535-0.

[11] Moradmand H, Aghamiri SMR, Ghaderi R. 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):179–90. https://doi.org/10.1002/acm2.12795.

[12] Zakaria R, Das K, Bhojak M, Radon M, Walker C, Jenkinson MD. The role of magnetic resonance imaging in the management of brain metastases: diagnosis to prognosis. Cancer Imaging 2014;14(1):8. https://doi.org/10.1186/1470-7330-14-8.

[13] Mayerhoefer ME, Szomolanyi P, Jirak D, Materka A, Trattnig S. 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):1236–43. https://doi.org/10.1118/1.3081408.

[14] Buch K, Kuno H, Qureshi MM, Li B, Sakai O. Quantitative variations in texture analysis features dependent on MRI scanning parameters: a phantom model. J Appl Clin Med Phys 2018;19(6):253–64. https://doi.org/10.1002/acm2.12482.

[15] Shinohara RT, Sweeney EM, Goldsmith N, Shiee N, Mateen FJ, Calabresi PA, et al. Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin 2014;6:9–19. https://doi.org/10.1016/j.nicl.2014.08.008.

[16] Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol 2021;31(4):2272–80. https://doi.org/10.1007/s00330-020-07284-9.

[17] Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 2000 Feb;19(2):143–50. https://doi.org/10.1109/42.836373.

[18] Reinhold JC, Dewey BE, Carass A, Prince JL. Evaluating the impact of intensity normalization on MR image synthesis. Proc SPIE Int Soc Opt Eng 2019;10949:109493H. https://doi.org/10.1117/12.2513089.

[19] Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics 2012;13(3):539–52. https://doi.org/10.1093/biostatistics/kxr034.

[20] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med 2018;59(8):1321–8. https://doi.org/10.2967/jnumed.117.199935.

[21] Shiri I, Amini M, Nazari M, Hajianfar G. Haddadi Avval A, Abdollahi H, et al. Impact of feature harmonization on radiogenomics analysis: prediction of EGFR and KRAS mutations from non-small cell lung cancer PET/CT images. Comput Biol Med 2022;142:105230. https://doi.org/10.1016/j.compbiomed.2022.105230.

[22] Mahon RN, Ghita M, Hugo GD, Weiss E. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. Phys Med Biol 2020;65(1):015010. https://doi.org/10.1088/1361-6560/ab6177.

[23] Juntu J, Sijbers J, Van Dyck D, Gielen J. Bias field correction for MRI images. In Computer Recognition Systems. Proceedings of the 4th International Conference on Computer Recognition Systems. 2005; 3: 543–551. https://doi.org/10.1007/3-540-32390-2_64.

[24] Duron L., Balvay D., Vande Perre S. Bouchouicha A, Savatovsky J, Sadik JC, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One 2019;14(3):e0213459.

[25] Buchner JA, Peeken JC, Etzel L, Ezhov I, Mayinger M, Christ S, et al. Identifying core MRI sequences for reliable automatic brain metastasis segmentation. medRxiv 2023. https://doi.org/10.1101/2023.05.02.23289342.

[26] Isensee F., Schell M., Pflueger I. Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp 2019;40(17):4952–4964. doi:10.1002/hbm.24750.

[27] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–20. https://doi.org/10.1109/TMI.2010.2046908.

[28] Shah M, Xiao Y, Subbanna N Francis S, Arnold DL, Collins DL, et al. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. Med Image Anal 2011;15(2):267–82. https://doi.org/10.1016/j.media.2010.12.003.

[29] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[30] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295(2):328–38. https://doi.org/10.1148/radiol.2020191145.

[31] Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005;19(1):231–40. https://doi.org/10.1519/15184.1.

[32] Lecler A., Duron L., Balvay D., J. Savatovsky, O. Bergès, M. Zmuda,et al. Combining multiple magnetic resonance imaging sequences provides independent reproducible radiomics features. Sci Rep 2019;9(1):2068. doi:10.1038/s41598-018-37984-8.

[33] Ibrahim A, Refaee T, Primakov S, Barufaldi B, Acciavatti RJ, Granzier RWY, et al. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. Cancers 2021;13(8):1848. https://doi.org/10.3390/cancers13081848.

[34] Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 2018;288(2): 407–15. https://doi.org/10.1148/radiol.2018172361.

[35] Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 2017;44(3):1050–62. https://doi.org/10.1002/mp.12123.

[36] Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Sci Rep 2020;10(1):12340. https://doi.org/10.1038/s41598-020-69298-z.

[37] Ronrick D, Francois L, Ingrid M, Ronan A, Joanne A, Caroline R, et al. Pre-selecting radiomic features based on their robustness to changes in imaging properties of multicentre data: impact on predictive modeling performance compared to ComBat harmonization of all available features. J Nucl Med. 2021: Supplemental 40–40.

[38] Stone, M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Series B 1974; 36(2):111–147. http://www.jstor.org/stable/2984809.