applied sciences

MDPI

Article

Comparison of Image Normalization Methods for Multi-Site Deep Learning

Steffen Albert [1,*], Barbara D. Wichtmann [2], Wenzhao Zhao [3,4], Angelika Maurer [2], Jürgen Hesser [3,4,5,6], Ulrike I. Attenberger [2], Lothar R. Schad [1] and Frank G. Zöllner [1]

1 Computer Assisted Clinical Medicine, Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, 68167 Mannheim, Germany; lothar.schad@medma.uni-heidelberg.de (L.R.S.); frank.zoellner@medma.uni-heidelberg.de (F.G.Z.)
2 Department of Diagnostic and Interventional Radiology, University Hospital Bonn, 53127 Bonn, Germany; barbara.wichtmann@ukbonn.de (B.D.W.); angelika.maurer@ukbonn.de (A.M.); ulrike.attenberger@ukbonn.de (U.I.A.)
3 Data Analysis and Modeling in Medicine, Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, 68167 Mannheim, Germany; wenzhao.zhao@medma.uni-heidelberg.de (W.Z.); juergen.hesser@medma.uni-heidelberg.de (J.H.)
4 Central Institute for Scientific Computing (IWR), Heidelberg University, 69120 Heidelberg, Germany
5 CZS Heidelberg Center for Model-Based AI, Central Institute for Scientific Computing (IWR), Heidelberg University, 69120 Heidelberg, Germany
6 Central Institute for Computer Engineering (ZITI), Heidelberg University, 69120 Heidelberg, Germany
* Correspondence: steffen.albert@medma.uni-heidelberg.de

Abstract: In this study, we evaluate the influence of normalization on the performance of deep learning networks for tumor segmentation and the prediction of the pathological response of locally advanced rectal cancer to neoadjuvant chemoradiotherapy. The techniques were applied to a multicenter and multimodal magnet resonance imaging data set consisting of 201 patients recorded at six centers. We implemented and investigated six different normalization methods (setting the mean and standard deviation, histogram matching, percentiles, combining percentiles and histogram matching, fixed window and an auto-encoder with adversarial loss using the imaging parameters) and evaluated their impact on four deep learning tasks: tumor segmentation, prediction of treatment outcome, and prediction of sex and age. The latter two tasks were implemented as a reference test. We trained a modified U-Net with different normalization methods in multiple configurations: on all images, images from all centers except one, and images from a single center. Our results show that normalization only plays a minor role in segmentation, with a difference in Dice of less than 0.02 between the best and worst performing networks. For the prediction of sex and treatment outcomes, the percentile method combined with histogram matching works best for all scenarios. The biggest difference in performance, depending on the normalization method, occurs for classification. In conclusion, normalization is especially important for small data sets or for generalizing to different data distributions. The deep learning method was superior to the classical methods only in a minority of cases, probably due to the limited amount of training data.

Keywords: normalization; MRI; medical imaging

check for updates

Citation: Albert, S.; Wichtmann, B.D.; Zhao, W.; Maurer, A.; Hesser, J.; Attenberger, U.I.; Schad, L.R.; Zöllner, F.G. Comparison of Image Normalization Methods for Multi-Site Deep Learning. *Appl. Sci.* **2023**, *13*, 8923. https://doi.org/10.3390/app13158923

Academic Editor: Shengsheng Wang

Received: 26 June 2023
Revised: 21 July 2023
Accepted: 1 August 2023
Published: 3 August 2023

1. Introduction

Colorectal cancer is the third most lethal cancer in Europe, with a 5-year survival rate of 68% in Germany [1]. The recommended treatment for locally advanced rectal cancer is radiotherapy or chemoradiotherapy followed by total mesorectal excision [2]. Neoadjuvant therapy can result in pathological complete remission. Accurate prediction of pathological treatment responses is essential to decide which tumors should be surgically resected and which patients should qualify for a watch-and-wait strategy.

For staging, magnetic resonance imaging (MRI) plays a key role [3]; however, T2-weighted (T2w) morphological imaging has low sensitivity while diffusion-weighted imaging (DWI) shows controversial results. Therefore, emerging techniques from machine learning are currently being investigated to overcome these limitations in predicting the pathologic treatment response. Thus, tumor segmentation and classification approaches have been studied.

Tumor segmentation in rectal cancer is intrinsically challenging [4], even for state-of-the-art deep neural networks, due to hard-to-delineate tumors. Similar problems arise for the prediction of treatment outcomes [5]. One problem is that data sets are relatively small and often include only one or two centers [6]. To apply such models in a broader clinical setting, they must show good generalization, which means that the network should be trained on a set of images from different centers. For this, it is important to implement a standardized and harmonized protocol. Nevertheless, certain difference arising from, e.g., vendor-specific differences between scanners and sequence implementations, exist [7]. To mitigate such effects to a certain extent, image normalization methods could be employed as a preprocessing method prior to training a machine learning algorithm.

There are different normalization methods, which could help reduce such data inhomogeneity. Statistical methods [8,9] can also be challenging because the intensity distribution can vary due to variations in the anatomy of the patient, the chosen field of view, and the content of the bladder and bowels. In addition to classical approaches, deep learning methods for normalization have emerged [10,11].

In the literature, there are various methods for domain adaptation [12], which go beyond simply changing the mean and the scale. A basic technique is histogram matching [13]. Many newer approaches are based on deep learning. Cycle GANs can be used to translate images from one acquisition protocol and scanner to another [14]. This can be extended to include more than two sites by using a StarGAN, which uses multiple encoders and decoders [15]. Other methods, such as ComBat [16,17] try to estimate and remove batch-dependent effects. ComBat can be applied directly to the images or to the resulting features.

The problem with most approaches is that they require batches of data with similar characteristics (domains) to remove batch-dependent effects or to transfer the style of one batch to the rest. This can be, for example, an individual scanner with a consistent acquisition protocol. There are some approaches that do not require a domain. ImUnity uses a reference image to specify the desired contrast [10].

For most MRIs, the acquisition parameters are known and saved in the DICOM image metadata. To our knowledge, no work has attempted to directly use these known parameters for training the normalization method.

In this study, we propose a deep learning algorithm that takes advantage of this information toward homogenization of the image data. Our method is evaluated against four classical methods from the literature. The techniques were applied to a retrospectively collected multimodal MRI data set of patients with locally advanced rectal cancer (LARC) recorded at six different centers. Furthermore, we evaluate the influence of different normalization technologies on the performance of deep learning networks for tumor segmentation and prediction of pathological responses.

2. Methods

2.1. Image Data

In this retrospective study, we used image data from 144 patients enrolled in six different centers within the CAO/RAO/AIO-12 study [18]. In addition, we retrospectively selected 57 patients from our institute as internal data. Our center participated in the study, so we pooled the study data with the internal data.

According to the study protocol, for each patient, a transverse 2D T2w sequence and a DWI sequence were acquired before therapy and between neoadjuvant therapy and operation. Although a study protocol was given, the obtained image data showed varying

acquisition parameters for the individual sequences. Figure 1 depicts, for example, two of these parameters for the T2w sequence. Similar differences in the acquisition parameters were observed for the DWI sequence. Example images can be seen in Figure 2.

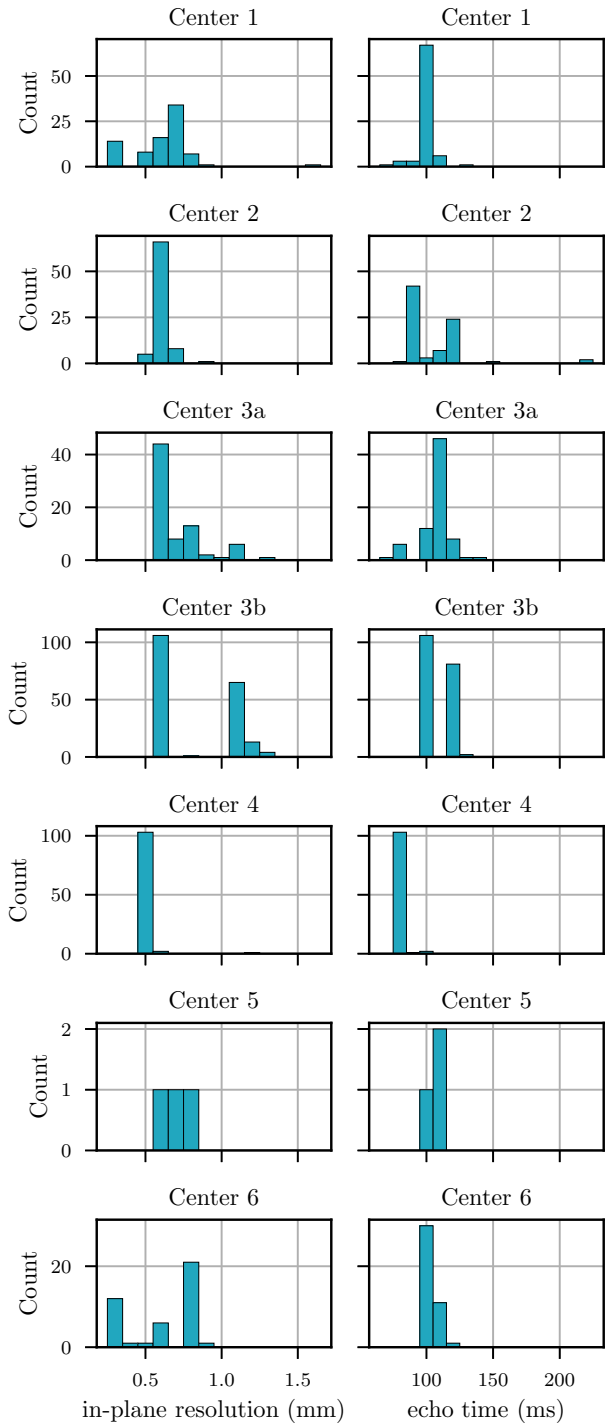

Figure 1. Distribution of acquisition parameters for the T2w axial images: The image acquisition parameters vary widely. The left column shows the in-plane resolution and the right the echo time. Each row shows the data from one center. The data from center 3 were split into the data from the study (3a) and from clinical routine (3b). The in-plane resolution is supposed to be 0.8 mm but varies between 0.26 mm and 1.64 mm. The echo time is supposed to be 110 ms but varies between 69 ms and 219 ms.

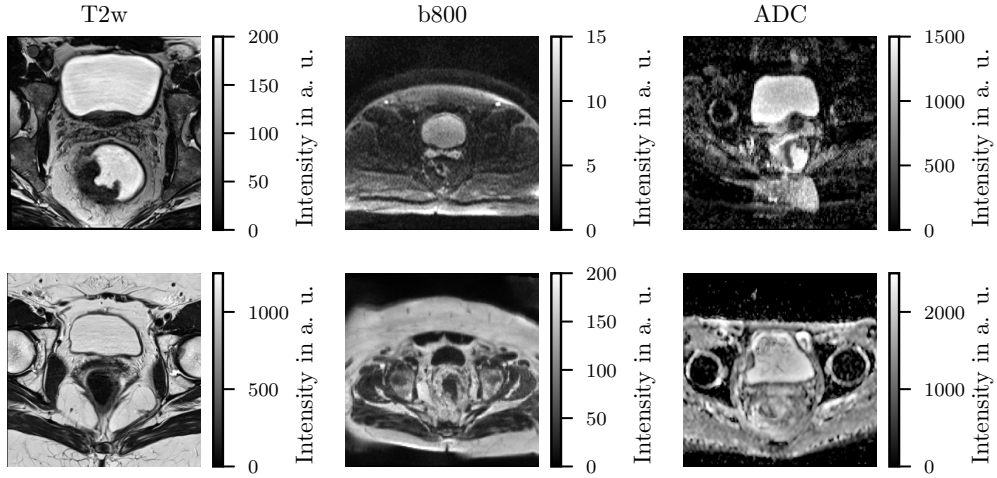

Figure 2. Examples of images from the study. Images with the lowest (top row) and highest mean intensities (bottom row) are shown for each modality. The range of intensity values varies largely, except for the ADC, which has a physical interpretation. Some of this variation is due to different acquisition parameters. The T2w image in the top left was acquired with an echo time of 219 ms instead of 110 ms. For the b800 and ADC images, both examples are from the same center, but different scanners.

2.2. Preprocessing

For preprocessing, we performed bias field correction using the N4 algorithm [19]. Then, we registered the diffusion-weighted images to the T2w image using ANTsPy [20]. For registration, we used the diffusion image with the lowest b-value (mostly b = 50, but b = 0 for some) and applied the same transformation to the other diffusion images and the apparent diffusion coefficient (ADC) image. Due to the high noise level in the diffusion-weighted images, we could not perform an elastic registration, so only a rigid registration was performed. Then, we normalized the images using the six methods described in the following.

2.3. Normalization Methods

2.3.1. Classical Methods

The percentile method (Perc) is very simple. We use 5th and 95th as the minimum and maximum values for the input of the network. We set any values outside this range to the corresponding minimum or maximum to eliminate outliers.

The second method is histogram matching (HM), originally developed for brain images [13]. The idea is to extract landmarks from each image and then average them over all images. As landmarks, we chose the 1st, 10th, 20th, ..., 90th and 99th percentiles of the voxel intensities. The average of the landmarks is used to define a standard histogram. Then, we interpolated the intensities to follow this standard histogram. We used the 1st percentile as the minimum value for the input of the neural network and the 99th percentile as the maximum input value.

The original paper suggests using Otsu thresholding to separate the brain from the background. Instead, we extracted the landmarks from the center volume (measuring $180 \times 180 \times 100\,\text{mm}^3$) because this region does not contain background voxels.

We also tested a combination of the percentile method and histogram matching (Perc-HM). We first normalized the images using the percentile method and then extracted the landmarks for histogram matching out of those images.

As the fourth normalization method, we set the mean and standard deviation to a fixed value (M-std). We set the mean to zero and the standard deviation to one.

The simplest method uses a fixed window (Win). We set the minimum to zero and the maximum to 3000 for the T2w and ADC images and 1000 for the b800 DWIs. We chose these values because nearly all images have a maximum intensity below them.

2.3.2. Deep Learning Method

As the last normalization method, we used an auto-encoder. We added multiple discriminators that we trained to predict the acquisition parameters of the DICOM headers.

The generator architecture is shown in Figure 3. It has a traditional CNN architecture, but we pass the edge information of the input image to the fully-upsampled output block to improve image quality. We apply a Gaussian filter before edge detection to propagate larger features but not noise.

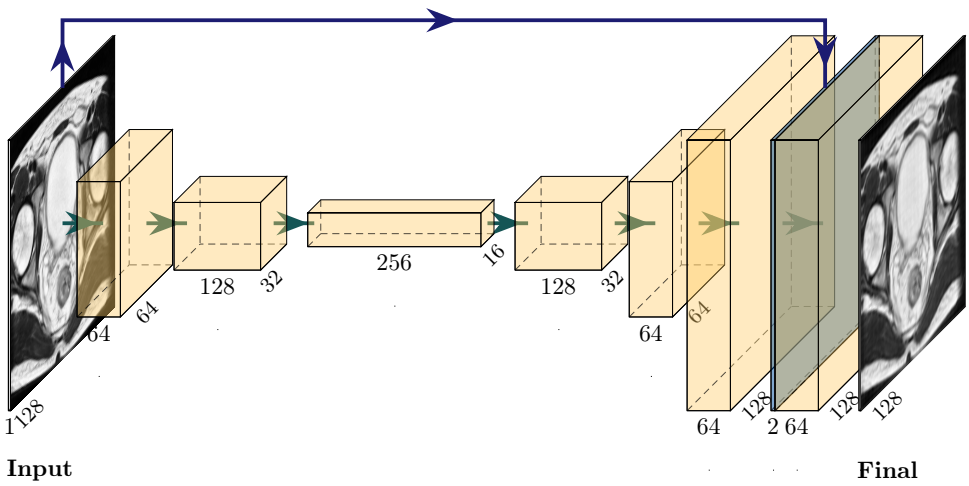

Figure 3. The architecture of the auto-encoder. Each yellow box represents a convolutional block consisting of batch normalization, a convolutional layer and activation. The numbers show the number of filters and the feature map size. The standard standard path of the features is shown by the green arrows and the path of the edge information is represented by the purple arrow. On the contracting path, we use a stride of two to decrease the feature map size. On the expanding path, we use transposed convolutions to increase the size again. Once we reach the full resolution again, we concatenate the edge information, and another convolutional block is applied.

Furthermore, we implemented three different discriminators: First, for acquisition parameters, the output of the discriminators applied to the generated image should match the value in the acquisition protocol. If no value was provided, we used the median value of all images. Second, for other variables, such as the scanner model or location, we attempted to remove the information using the discriminator as an adversarial loss on the latent features or the generated image. Therefore, the desired result is the same probability for each class in the classification tasks. Eventually, we added a discriminator that tries to detect which images are the original input images and which were generated by the generator. The generator tries to fool this discriminator.

In each training step, we first trained these discriminators. Then, we trained the image discriminator on the original input and the generator's output images. The latent space discriminator was trained on the original images. The generator already needs images in a certain range as input, so we used one of the classical normalization methods to normalize the images before training the auto-encoder. We individually trained the auto-encoder for each set of training images and each modality. For training, we used all the images available for that modality, not only the segmented images. We used different hyperparameters, as depicted in Table 1.

Table 1. Hyperparameters for the different GANs used for normalization. The first is the default (GAN-Def). For GAN-Seg, we added segmentation as an additional task to preserve the important details. GAN-Img uses all discriminators on the images and not on the latent space. We trained GAN-Win and GAN-No-ed on images with window normalization with and without propagated edge information.

Network	Hyperparameters			
	Segmentation Loss	Train Only on Image	Initial Normalization	Skip Edges
GAN-Def	No	No	Perc	Yes
GAN-Seg	Yes	No	Perc	Yes
GAN-Img	No	Yes	Perc	Yes
GAN-Win	No	No	Win	Yes
GAN-No-ed	No	No	Win	No

2.4. Experiments

Ground truth tumor segmentations were obtained from a radiologist with five years of experience manually delineating the tumor on T2-weighted pre-therapy MRIs. For treatment response, we used the Dworak regression grade [21]. This system rates the tumor response on a scale of 0 (no regression) to 4 (no remaining tumor cells) and is used on the resected tumor. Sex and age are given in the patient data.

We trained one network for segmentation and another for classification and regression. We used a modified 2D U-Net [22] with batch normalization and residual connections for segmentation. Although published in 2015, U-Nets are still widely used for medical segmentation tasks [23]. As an architecture for classification and regression, we chose ResNet50 [24], which we used with random initialized weights. We only changed the last layer of ResNet to have the desired number of output neurons. We trained all networks for 100 epochs with 5-fold cross-validation. The three modalities (T2w, ADC and b800) were combined into a three-channel image, and we extracted 32 random patches per image. For segmentation, at least 40% of patches have their center inside the tumor volume. We augmented the patches by rotating them in-plane and uniformly scaling them.

We trained the network in three different configurations:

- All. In this configuration, we trained on all images from all centers and evaluated the network performance using cross-validation.
- Except-One. In the next experiment, we trained on all centers except one. We evaluated the performance of the training centers using cross-validation and evaluated the network of each fold on the remaining center.
- Single-Center. In the last configuration, we trained on one center only. Similarly to the Except-One experiment, we evaluated the performance on that center using cross-validation and applied all five networks to the other centers.

We normalized the three modalities (T2w, ADC, b800) individually. The Perc, M-std and Win methods do not need to be trained, so we only normalized the whole data set once. For HM, Perc-HM and the deep learning method, we trained the methods for each experiment on the patients included in the training and validation set.

As evaluation metrics, we used the Dice coefficient [25] of the tumor class for segmentation and the area under the receiver operator characteristic curve (AUC) for the prediction of sex and Dworak grade. A zero Dice indicates no overlap with the ground truth, and one indicates a perfect segmentation. For the AUC, a value of 0.5 is equal to pure chance and one means perfect prediction. In the case of several classes, we calculated the AUC by the average of the one-versus-others AUCs. We used the root mean square error (RMSE) for age prediction.

After training the networks, we evaluated them on previously unseen images from the same center and unseen data from all other centers, using the network from the epoch

with the best performance in the validation set. We used a Student's t test to determine the significance of the mean differences. We consider a p-value of less than 0.05 significant.

3. Results

Figure 4 shows examples of normalized vs. unprocessed slices using different normalization methods. The results obtained from the test data set are summarized in Figure 5.

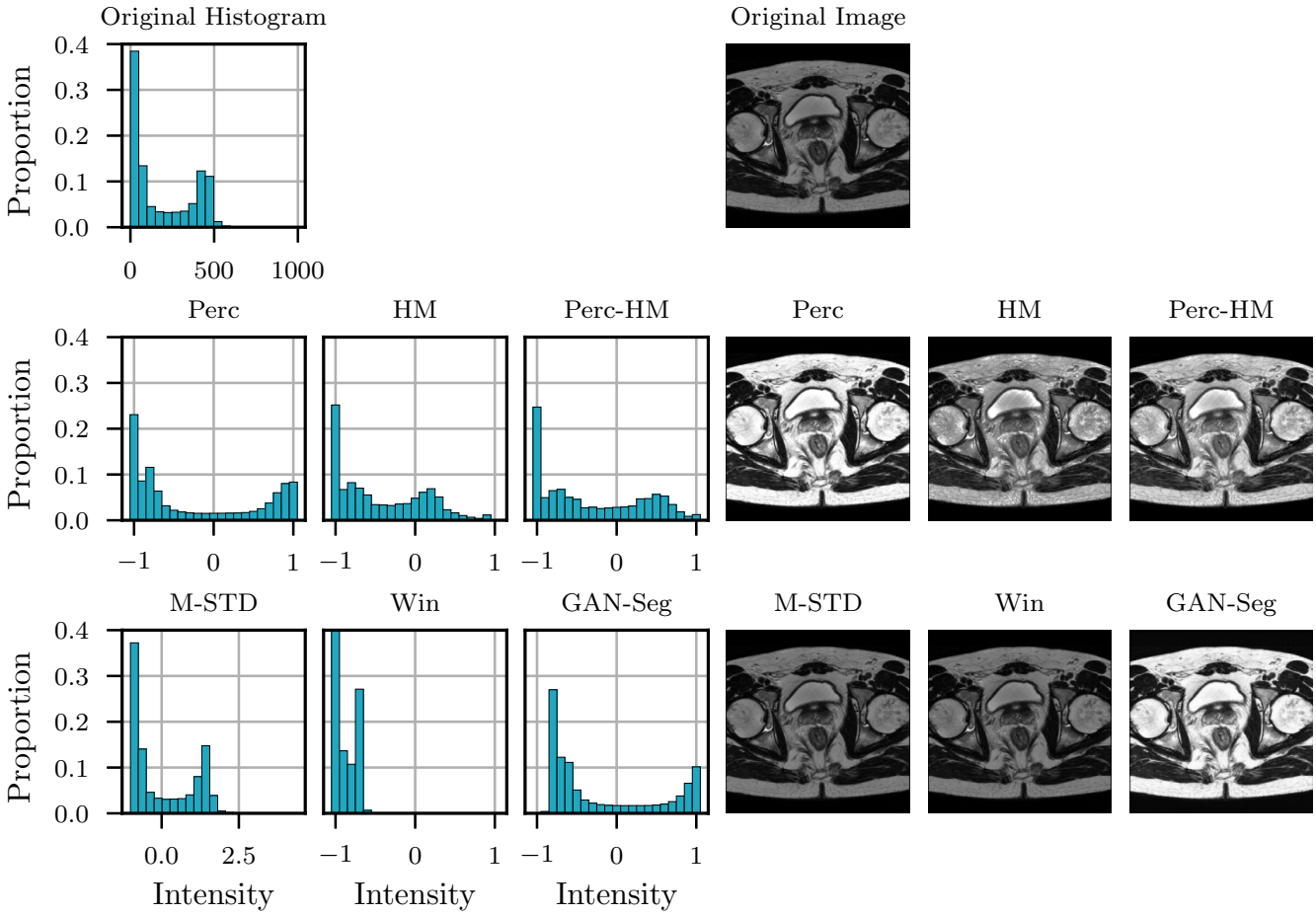

Figure 4. Comparison of the different normalization methods applied to one image. The upper row shows the original image and a histogram of the intensity. The following two rows show an image and a histogram for six exemplary selected normalization methods. The images are normalized to the minimum and maximum values of the resulting slice. Using a fixed window (Win) or subtracting the mean and dividing by the standard deviation (M-STD) only shifts and rescales the values; thus, the images look the same as the original image. For the fixed window, a maximum value must be selected that is higher than the intensity of most voxels in most of the images; therefore, many images only use a small part of the available range. The other methods result in intensities between −1 and 1 (other values can also be selected). They all remove outliers, especially at higher intensities, which increases the contrast in the visible images. Areas with high intensities are mostly fat, urine and bone marrow. For methods using histogram matching (HM and Perc-HM), all normalized images will have the same intensity distribution.

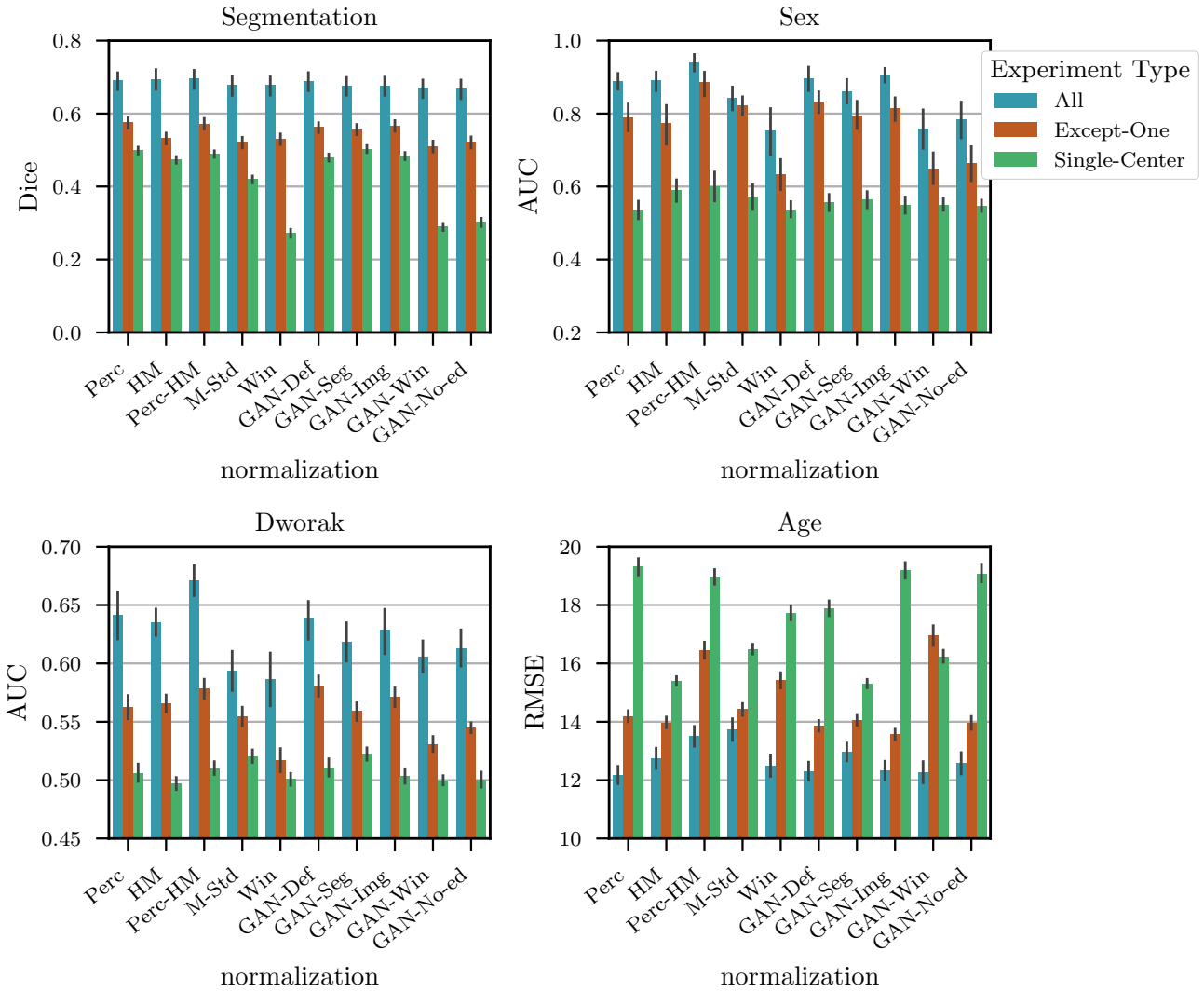

Figure 5. Performance of the different normalization methods for each task and training scenario (All, Except-One and Single-Center). For Except-One and Single-Center, only the results for images not from the training centers are shown.

3.1. All

When looking at all centers, there were no significant differences in segmentation. The best method is Perc-HM with a Dice of 0.69 ± 0.01. For the Sex and Dworak classification, Perc-HM is significantly better than all other methods with an AUC of 0.94 ± 0.02 and 0.67 ± 0.01. For age prediction, Perc, Win, GAN-Def, GAN-Img, GAN-Win and GAN-No-ed are the best methods without significant differences, with Perc being the best with an RMSE of 12.2 ± 0.2.

3.2. Except-One

When leaving out one center, all normalization methods achieve similar Dice scores between 0.66 and 0.69 on the training centers. For the test center, Perc, Perc-HM, GAN-Def, GAN-Seg and GAN-Img are the best methods with no significant differences. The best is Perc with a Dice of 0.58 ± 0.01.

When classifying the Dworak score for patients from the same centers used in training, there are no significant differences between the normalization methods. The mean AUC is 0.59. When evaluating the test center, Perc-HM, GAN-Def and GAN-Img are the best methods. GAN-Def has the highest Dice of 0.581 ± 0.004.

Sex classification works best if the images are normalized using Perc-HM for data from the training and test centers. For images from training centers, the AUC is 0.87 ± 0.04, but only HM, Win, GAN-Win and GAN-No-ed are significantly worse. For test centers, Perc-HM is significantly the best method, with an AUC of 0.88 ± 0.2.

For age prediction, GAN-No-ed performs best with an RMSE of 12.7 ± 0.2 years for patients from the same center, but it is not significantly better than GAN-Def, GAN-Img or M-Std. GAN-Img is the best method for data from the test center with an RMSE of 13.6 ± 1.

3.3. Single Center

When looking at the performance of the segmentation of images from the same center, Perc-HM achieves the highest mean Dice score of 0.66 ± 0.01. However, it is not significantly better than all other methods, except for M-Std, Win, GAN-Win and GAN-No-ed (see Figure 5). For all other centers, Perc and GAN-Seg are the best with a Dice score of 0.50 ± 0.01 (for both). However, Perc is not significantly better than Perc-HM (Dice of 0.49 ± 0.01) and GAN-Seg is just barely significantly better (with a p-value of 0.0496).

For the Dworak score classification, the best is GAN-Win with an AUC of 0.57 ± 0.01, but only Perc, M-Std and GAN-No-ed are significantly worse. For patients from the other centers, GAN-Seg and M-Std are the best, with an AUC of 0.522 ± 0.003 and 0.520 ± 0.003.

For sex classification, there were no significant differences for the training centers. For the test centers, Perc-HM was the best (AUC of 0.60 ± 0.02), but HM, M-Std, GAN-Def and GAN-Seg were not significantly worse. However, for age prediction, the best normalization method is HM with an RMSE of 13.6 years, but it is not significantly better than GAN-Win and GAN-Seg for training centers. HM and GAN-Seg were the best for test centers with RMSEs of 15.4 ± 0.08 and 15.3 ± 0.08.

4. Discussion

In this study, we proposed a deep-learning-based approach that incorporates image sequence parameters for image normalization. Furthermore, we investigated the influence of the implemented normalization strategies, including deep-learning-based approaches to rectal cancer segmentation, classification and regression from multimodal MRI acquired in a multicenter study.

For segmentation, the different normalization methods only lead to minor differences. For classification and regression, there are larger differences, and the best performing method is the percentile method combined with histogram matching.

The intensity of the MRI signal depends mainly on the tissue properties of the imaged voxel. The normalization methods use local information (for CNN-based methods) and/or global information (for most statistical methods) to standardize the images. This is not sufficient because the other voxels have different tissue properties. This limits how well the normalization models are able to correct anomalies. Thus, normalization probably mostly helps the neural network by adding prior information (for example, the mean intensity distribution when performing histogram matching), which is why it helps less for larger data sets.

In addition to different acquisition parameters, there are many other parameters that can hinder the generalizability of the trained network. There are differences in the patient population and treatment. For example, the time difference between the end of neoadjuvant treatment and the operation was (30 ± 6) d for Center 1 and (37 ± 6) d for Center 2. Differences like this cannot be corrected by normalization.

4.1. Classical Normalization Methods

The best-performing tumor segmentation network (Perc-HM) reaches a Dice of 0.69 ± 0.01 and is in a similar range (0.68–0.74) reported in the literature [4,26]. There are no significant differences in segmentation performance between normalization methods when training on all data. For the other scenarios, Perc, Perc-HM and GAN-Seg performed best.

The best-performing model for Dworak classification has a lower AUC of 0.67 ± 0.01 compared to [27] with an AUC of 0.82, but our data set is only a quarter of the size and is more heterogeneous. For the classification of sex and age, we could not find respective studies to compare. Classifying the sex of the patients results in a high AUC of 0.94 ± 0.02. Probably, since some sexual organs are visible in the images, this information might be picked up by the network.

Compared to the segmentation task, there are fewer examples, since the whole volume is classified, while for segmentation, each voxel is assigned a label and thus contributes to the overall performance, though certainly the voxels are not independent. This could explain why we see fewer differences between the normalization methods for segmentation compared to classification.

The performance decreases for all tasks as the size of the data set decreases, as expected. This is demonstrated by leaving out one center, and especially when training only on one center (see Figure 5). Here, the largest differences could be observed between the different normalization methods. For data from unseen centers, there is a large generalization error. One of the reasons for this error is that there are greater differences in the data acquisition parameters between centers than within one, as visible in Figure 1. Some normalization methods can better reduce these differences than others.

In summary, for tumor segmentation, Dworak and sex classification, Perc-HM is the best method in the All and Except-One scenarios. In the Single-Center scenario, Perc-HM also performs well for sex classification, but for segmentation and Dworak classification, GAN-Seg was significantly better. For age prediction, the results are inconclusive; no method is superior in all three scenarios.

4.2. Deep Learning Normalization Methods

For segmentation and Dworak classification, GAN-Def outperformed all other DL methods when training on all centers. The segmentation performance (0.69 ± 0.01) was comparable to that of classical methods and in the literature [4,26], but the Dworak classification (0.64 ± 0.01) is worse than that of classical methods and in the literature [27]. We observed that the DL methods are superior to the classical ones in only a few cases when used in the other two scenarios. For example, GAN-Seg, which uses segmentation as an additional task, was among the best methods for the Single-Center scenario and for segmentation in all scenarios. It achieves an AUC of 0.522 ± 0.003 for the Dworak classification in the Single-Center scenario, and scores lower than the respective best classical model and the literature reference [27]. One of the issues is that the data set used to train the GAN normalization is only slightly larger than the data set used to train the segmentation and classification. Thus, the U-Net and ResNet probably learn an encoding similar to that of the auto-encoder. The advantage of deep-learning-based methods is that they can be trained on a larger data set without the need for manual annotations.

4.3. Limitations

In this study, only two networks were tested for their performance using different normalization methods; while U-Net and ResNet are widely used, there are many other architectures. For segmentation, most follow the encoder–decoder structure and should behave similarly.

When predicting the Dworak score, there are not sufficient data to achieve a performance which would be sufficient to be used for clinical decision making. The best method achieved an AUC of 0.69 ± 0.01. This is not sufficient to inform treatment decisions. Even in studies with a larger training cohort of 592 patients, only an AUC of 0.82 was achieved [27].

One of the issues is that only about 20% of patients achieve a pathological complete response [2]. In the data set, there were only 37 patients with a pathological complete response. Therefore, to obtain clinically usable networks, the training data set would probably have to be larger by orders of magnitude. Collecting that much data faces many regulatory and technical challenges.

For segmentation, this is less of a problem because there are a lot more data points. There are approximately ten million voxels in the tumor class. Although these are not independent samples, there are still much more data available than for the classification. The maximum Dice of 0.69 ± 0.01 is in the range of the interobserver Dice of 0.71 ± 0.13 observed in [26] and 0.83 ± 0.13 in [4]. The performance could probably be further increased with more parameter tuning.

The difference in available data for the different tasks is probably also the reason why there are much larger differences in performance using different normalization methods for classification and regression than for segmentation.

5. Conclusions

The performance of six different normalization methods was evaluated for different deep learning tasks for LARC patients in a multicenter setting with data from six different centers. Different scenarios were tested with training on data from all centers, on data from all centers except one, and from data from a single center. In this way, the influence of the normalization method on generalizability can be assessed.

Normalization is important if the data are inhomogeneous, especially if the data set is small and the network is applied to data from a site not included in the training set. It plays a larger role in classification and regression than in segmentation.

Our results show that percentile normalization followed by histogram matching performed the best for tumor segmentation and prediction of treatment outcomes in locally advanced rectal cancer. Setting the mean and standard deviation to a fixed value, which is often done for images, performed significantly worse than most other methods.

Using deep learning approaches did not lead to any improvements over classical methods in most cases, and was only slightly better when training on data from a single center.

Normalization improved the generalization of the network for different tasks, but there are limits to what can be corrected with normalization. It is essential to standardize data acquisition for routine clinical imaging for the widespread application of deep learning in clinics. With the current data available, it is not possible to predict clinically relevant outcomes of neoadjuvant treatment for patients with LARC.

Author Contributions: Conceptualization, S.A. and F.G.Z.; Methodology, S.A.; Software, S.A.; Validation, S.A.; Formal analysis, S.A.; Investigation, S.A.; Data curation, S.A., B.D.W. and A.M.; Writing—original draft, S.A.; Writing—review & editing, S.A., B.D.W., W.Z., J.H. and F.G.Z.; Visualization, S.A.; Supervision, J.H., U.I.A., L.R.S. and F.G.Z.; Funding acquisition, J.H., U.I.A. and F.G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: We gratefully acknowledge the support by German Research Foundation (DFG) through the grant 428149221 as part of the Radiomics priority program (SPP 2177) and data storage SDS@hd supported by the Ministry of Science, Research, and the Arts Baden-Württemberg (MWK) and the DFG through the grants INST 35/1314-1 FUGG and INST 35/1503-1 FUGG.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee II of Heidelberg University (2017-571N-MA).

Informed Consent Statement: Informed consent was waived due to the retrospective nature of this work.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to patient privacy.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MRI	magnetic resonance imaging
T2w	T2-weighted
DWI	diffusion weighted imaging
ADC	apparent diffusion coefficient
LARC	locally advanced rectal cancer

References

1. Fitzmaurice, C.; Dicker, D.; Pain, A.; Hamavid, H.; Moradi-Lakeh, M.; MacIntyre, M.F.; Allen, C.; Hansen, G.; Woodbrook, R.; Wolfe, C.; et al. The Global Burden of Cancer 2013. *JAMA Oncol.* **2015**, *1*, 505. [CrossRef]
2. Benson, A.B.; Venook, A.P.; Bekaii-Saab, T.; Chan, E.; Chen, Y.J.; Cooper, H.S.; Engstrom, P.F.; Enzinger, P.C.; Fenton, M.J.; Fuchs, C.S.; et al. Rectal Cancer, Version 2.2015. *J. Natl. Compr. Cancer Netw.* **2015**, *13*, 719–728. [CrossRef]
3. Horvat, N.; Carlos Tavares Rocha, C.; Clemente Oliveira, B.; Petkovska, I.; Gollub, M.J. MRI of Rectal Cancer: Tumor Staging, Imaging Techniques, and Management. *RadioGraphics* **2019**, *39*, 367–387. [CrossRef] [PubMed]
4. Trebeschi, S.; van Griethuysen, J.J.M.; Lambregts, D.M.J.; Lahaye, M.J.; Parmar, C.; Bakers, F.C.H.; Peters, N.H.G.M.; Beets-Tan, R.G.H.; Aerts, H.J.W.L. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. *Sci. Rep.* **2017**, *7*, 5301. [CrossRef] [PubMed]
5. Wichtmann, B.D.; Albert, S.; Zhao, W.; Maurer, A.; Rödel, C.; Hofheinz, R.D.; Hesser, J.; Zöllner, F.G.; Attenberger, U.I. Are We There Yet? The Value of Deep Learning in a Multicenter Setting for Response Prediction of Locally Advanced Rectal Cancer to Neoadjuvant Chemoradiotherapy. *Diagnostics* **2022**, *12*, 1601. [CrossRef]
6. Wong, C.; Fu, Y.; Li, M.; Mu, S.; Chu, X.; Fu, J.; Lin, C.; Zhang, H. MRI-Based Artificial Intelligence in Rectal Cancer. *J. Magn. Reson. Imaging* **2023**, *57*, 45–56. [CrossRef]
7. Mayerhoefer, M.E.; 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: Effects of MRI Acquisition Parameters on Texture Analysis. *Med. Phys.* **2009**, *36*, 1236–1243. [CrossRef]
8. Reinhold, J.C.; Dewey, B.E.; Carass, A.; Prince, J.L. Evaluating the Impact of Intensity Normalization on MR Image Synthesis. In *Medical Imaging 2019: Image Processing*; Angelini, E.D., Landman, B.A., Eds.; SPIE: San Diego, CA, USA, 2019; p. 126. [CrossRef]
9. Shah, M.; Xiao, Y.; Subbanna, N.; Francis, S.; Arnold, D.L.; Collins, D.L.; Arbel, T. Evaluating Intensity Normalization on MRIs of Human Brain with Multiple Sclerosis. *Med. Image Anal.* **2011**, *15*, 267–282. [CrossRef] [PubMed]
10. Cackowski, S.; Barbier, E.L.; Dojat, M.; Christen, T. ImUnity: A Generalizable VAE-GAN Solution for Multicenter MR Image Harmonization. *Med. Image Anal.* **2023**, *88*, 102799. [CrossRef]
11. Tax, C.M.; Grussu, F.; Kaden, E.; Ning, L.; Rudrapatna, U.; John Evans, C.; St-Jean, S.; Leemans, A.; Koppers, S.; Merhof, D.; et al. Cross-Scanner and Cross-Protocol Diffusion MRI Data Harmonisation: A Benchmark Database and Evaluation of Algorithms. *NeuroImage* **2019**, *195*, 285–299. [CrossRef]
12. Guan, H.; Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1173–1185. [CrossRef]
13. Nyul, L.G.; Udupa, J.K.; Zhang, X. New Variants of a Method of MRI Scale Standardization. *IEEE Trans. Med. Imaging* **2000**, *19*, 143–150. [CrossRef] [PubMed]
14. Modanwal, G.; Vellal, A.; Mazurowski, M.A. Normalization of Breast MRIs Using Cycle-Consistent Generative Adversarial Networks. *Comput. Methods Programs Biomed.* **2021**, *208*, 106225 .
15. Bashyam, V.M.; Doshi, J.; Erus, G.; Srinivasan, D.; Abdulkadir, A.; Singh, A.; Habes, M.; Fan, Y.; Masters, C.L.; Maruff, P.; et al. Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. *J. Magn. Reson. Imaging* **2022**, *55*, 908–916. [CrossRef] [PubMed]
16. Fortin, J.P.; Parker, D.; Tunç, B.; Watanabe, T.; Elliott, M.A.; Ruparel, K.; Roalf, D.R.; Satterthwaite, T.D.; Gur, R.C.; Gur, R.E.; et al. Harmonization of Multi-Site Diffusion Tensor Imaging Data. *NeuroImage* **2017**, *161*, 149–170. [CrossRef]
17. Mali, S.A.; Ibrahim, A.; Woodruff, H.C.; Andrearczyk, V.; Müller, H.; Primakov, S.; Salahuddin, Z.; Chatterjee, A.; Lambin, P. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J. Pers. Med.* **2021**, *11*, 842. [CrossRef]
18. Rödel, C.; Liersch, T.; Becker, H.; Fietkau, R.; Hohenberger, W.; Hothorn, T.; Graeven, U.; Arnold, D.; Lang-Welzenbach, M.; Raab, H.R.; et al. Preoperative Chemoradiotherapy and Postoperative Chemotherapy with Fluorouracil and Oxaliplatin versus Fluorouracil Alone in Locally Advanced Rectal Cancer: Initial Results of the German CAO/ARO/AIO-04 Randomised Phase 3 Trial. *Lancet Oncol.* **2012**, *13*, 679–687. [CrossRef]
19. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **2010**, *29*, 1310–1320. [CrossRef]
20. Avants, B.B.; Tustison, N.J.; Song, G.; Cook, P.A.; Klein, A.; Gee, J.C. A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration. *NeuroImage* **2011**, *54*, 2033–2044. [CrossRef]

21. Dworak, O.; Keilholz, L.; Hoffmann, A. Pathological Features of Rectal Cancer after Preoperative Radiochemotherapy. *Int. J. Color. Dis.* **1997**, *12*, 19–23. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
23. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
25. Dice, L.R. Measures of the Amount of Ecologic Association between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
26. Wang, J.; Lu, J.; Qin, G.; Shen, L.; Sun, Y.; Ying, H.; Zhang, Z.; Hu, W. Technical Note: A Deep Learning-Based Autosegmentation of Rectal Tumors in MR Images. *Med. Phys.* **2018**, *45*, 2560–2564. [CrossRef]
27. Shin, J.; Seo, N.; Baek, S.E.; Son, N.H.; Lim, J.S.; Kim, N.K.; Koom, W.S.; Kim, S. MRI Radiomics Model Predicts Pathologic Complete Response of Rectal Cancer Following Chemoradiotherapy. *Radiology* **2022**, *303*, 211986. [CrossRef]