ImUnity: A generalizable VAE-GAN solution for multicenter MR image harmonization

Stenzel Cackowski, Emmanuel Barbier, Michel Dojat, Thomas Christen

ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization

Stenzel Cackowski[*]
stenzel.cackowski@univ-grenoble-alpes.fr

Emmanuel L. Barbier[*†]
emmanuel.barbier@univ-grenoble-alpes.fr

Michel Dojat[*]
michel.dojat@univ-grenoble-alpes.fr

Thomas Christen[*]
thomas.christen@univ-grenoble-alpes.fr

ABSTRACT

ImUnity is an original deep-learning model designed for efficient and flexible MR image harmonization. A VAE-GAN network, coupled with a confusion module and an optional biological preservation module, uses multiple 2D-slices taken from different anatomical locations in each subject of the training database, as well as image contrast transformations for its self-supervised training. It eventually generates 'corrected' MR images that can be used for various multi-center population studies. Using 3 open source databases (ABIDE, OASIS and SRPBS), which contain MR images from multiple acquisition scanner types or vendors and a large range of subjects ages, we show that ImUnity: (1) outperforms state-of-the-art methods in terms of quality of images generated using traveling subjects; (2) removes sites or scanner biases while improving patients classification; (3) harmonizes data coming from new sites or scanners without the need for an additional fine-tuning and (4) allows the selection of multiple MR reconstructed images according to the desired applications. Tested here on T1-weighted images, ImUnity could be used to harmonize other types of medical images.

Keywords Brain · Deep Learning · Adversarial Network · Machine Learning · Self-supervised learning · Radiomic features

1 Introduction

Magnetic Resonance (MR) data acquired from the same patient but at different acquisition sites often lead to different MR images. This is due to the qualitative nature of the acquisitions which produces weighted images (such as T1w or T2w) that are sensitive to technical choices (hardware, sequence parameters) as well as scanner artifacts. Consequently, pooling images from multi-center MR studies in order to approach a particular clinical or biological question does not guarantee an increase in statistical power because of a parallel increase in non-biological variance. These unwanted variations in image intensities also prevent large dissemination of machine learning tools that are trained on a specific site and may not generalize their model to other image providers [Liu et al., 2020].

Several solutions have been proposed in the last decade to harmonize data coming from multi-site or multi-scanner MR studies. Their goal is to remove confounding site, scanner or protocol effects, while preserving the biological information contained in the images. Classical post-processing steps such as standardization, global scaling [Fortin et al., 2017] or intensity histogram matching [Shinohara et al., 2014] have been shown to reduce the influence of site or scanner biases. However, they also tend to remove informative local intensity variations. Statistical techniques, where image intensity and datasets bias are modeled in every voxel, have been proved to be more efficient. Ravel [Fortin et al., 2016], ComBat [Fortin et al., 2017], refined ComBat versions (Pomponio et al. [2019], Beer et al. [2020]) or dictionary learning [St-Jean et al., 2020] methods have been successfully used to analyze harmonization impacts on diffusion

---

[*]Université Grenoble Alpes, Inserm U1216, Grenoble Institut for Neurosciences, 38700 La Tronche, France
[†]Corresponding author

	Travelling subjects	Fine-tuning for new clinical question	Fine-tuning for unseen sites	Max. number of target sites
Zhu et al. [2018] (Cy-cleGAN)	not required	not required	required	N = 2
Dewey et al. [2019] (Deep-Harmony)	required	not required	required	N = 2
Zuo et al. [2021] (Calamity)	not required	not required	required	N = number of training sites
Dinsdale et al. [2020], Guan et al. [2021]	not required	required	not required	N > number of training sites
This study (ImUnity)	not required	not required	not required	N > number of training sites

Table 1: Versatility of deep-learning harmonization models

MRIs or longitudinal structural sequences. However, it can be noted that these techniques need to be adjusted every time new sites or scanners provide images to the database. Moreover, the same clinical individual information (such as patient age, sex, etc.) needs to be available in every center of the database. More recently, deep-learning models such as CycleGAN [Zhu et al., 2018], Deep-Harmony [Dewey et al., 2019] or Calamity [Zuo et al., 2021] have shown encouraging results for structural MR image (T1w, T2w or FLAIR) harmonization but have also shown limitations. Briefly, CycleGAN, which consists in two Generative Adversarial Networks (GANs) working together, is restricted to the harmonization of 2 sites, and needs to be fine-tuned for every pair of sites. Deep-Harmony, a U-Net [Ronneberger et al., 2015] harmonization adaptation network, has the disadvantage of requiring traveling subjects (subjects who have been scanned successively at different sites or scanners) for all scanners/sites for its training, a condition barely met in practice even in prospective studies. Similarly to CycleGAN, it is limited to two sites and needs to be fine-tuned. Calamity, an unsupervised deep-learning method, needs two different MR sequences as inputs for every subject and needs fine-tuning when data from new sites are considered. Finally, Dinsdale et al. [2020] and Guan et al. [2021] have proposed to include unlearning modules or domain discriminators directly into their classification networks. As such, they learn how to remove datasets biases during their analyses without reconstructing harmonized MR images. They have shown improvements in brain tissues segmentation and brain disorder classification after harmonization. However, these techniques clearly require to be trained for every new clinical question while the former approaches harmonize data once for all.

We propose in this paper a new type of harmonization method, called ImUnity, based on deep-learning, which extends previous techniques to offer a fast and flexible harmonization solution. ImUnity generates 'corrected' MR images that can then be utilized for various population imaging studies. To avoid the need for traveling subjects or multiple MR sequences in the database, our self-supervised Variational AutoEncoder (VAE-GAN) architecture uses for its training multiple slices from the same individual and randomized image contrast transformations. It also unlearns center bias using a confusion module connected to its bottleneck while an optional biological module can ensure that clinical features are preserved in the latent space. Once trained, this architecture should allow data coming from new sites or scanners to be harmonized without the need for fine-tuning. The architecture also allows estimates towards multiple target sites and then, users can choose multiple MR image reconstructions according to the chosen target domain (site or scanner).

To evaluate the efficiency and flexibility of our harmonization tool, we tested the approach using 3 open source databases that contain images from multiple acquisition sites, scanner vendors or strength of magnetic fields, and a large range of patients ages. For most of the experiments, ImUnity was trained using data from only one of the databases and then applied to the other two to evaluate generalisation of the model. Quality of the reconstructed images, capacity of removing site or scanner bias and ability to classify patients were evaluated after data harmonization.

2 Materials and methods

2.1 Data

We used three open-source databases: (1) ABIDE, a multi-center project led by Di Martino et al. [2014], which focuses on Autism Spectrum Disorder (ASD). It gathers more than 1,000 autistic patients and controls. For this study, we used T1-weighted scans from 11 different sites and scanners from 3 different constructors (3T scanners at 10 different sites and one 1.5T scanner at one site). Sites presenting data from a large range of ages (from 6 to 47 years, mean age =

12 years) were selected. In total, 621 T1-weighted scans (309 patients and 312 controls) were collected. (2) OASIS (LaMontagne et al. [2019]) gathers T1-weighted scans from adult subjects who underwent several MR sessions on 4 different scanners from the same site. We used these traveling subjects (n = 769) to validate the ability of our model to perform multi-scanners harmonization.

(3) SRPBS (Tanaka et al. [2021]) is a multi-site database gathering multi-disorder subjects. We used 9 healthy adult traveling subjects to validate harmonization results between the different acquisition sites of the database (6 sites, 12 scanners from 3 different constructors). Note that OASIS and SRPBS images contain healthy adult brain scans while ABIDE mainly include healthy and pathological infant brain scans, leading to large anatomical differences between images in the databases.

For each subject in each database, the brain was extracted using Robex [Iglesias et al., 2011] and N4Bias [Tustison et al., 2010] was used to correct for intensity inhomogeneities. MR images were first co-registered to the publicly available and age specific 152-MNI templates [Sanchez et al., 2012]. Then, White-Stripe normalization [Shinohara et al., 2014] was run to align white matter (WM) peaks between all subjects (each WM peak was aligned to 0.7 after rescaling the whole image between [0:1]).

After visual inspection to detect images with ROBEX defects or other artifacts, we eventually included 545, 1072 and 81 T1-weighted scans from ABIDE, OASIS and SRPBS databases respectively.
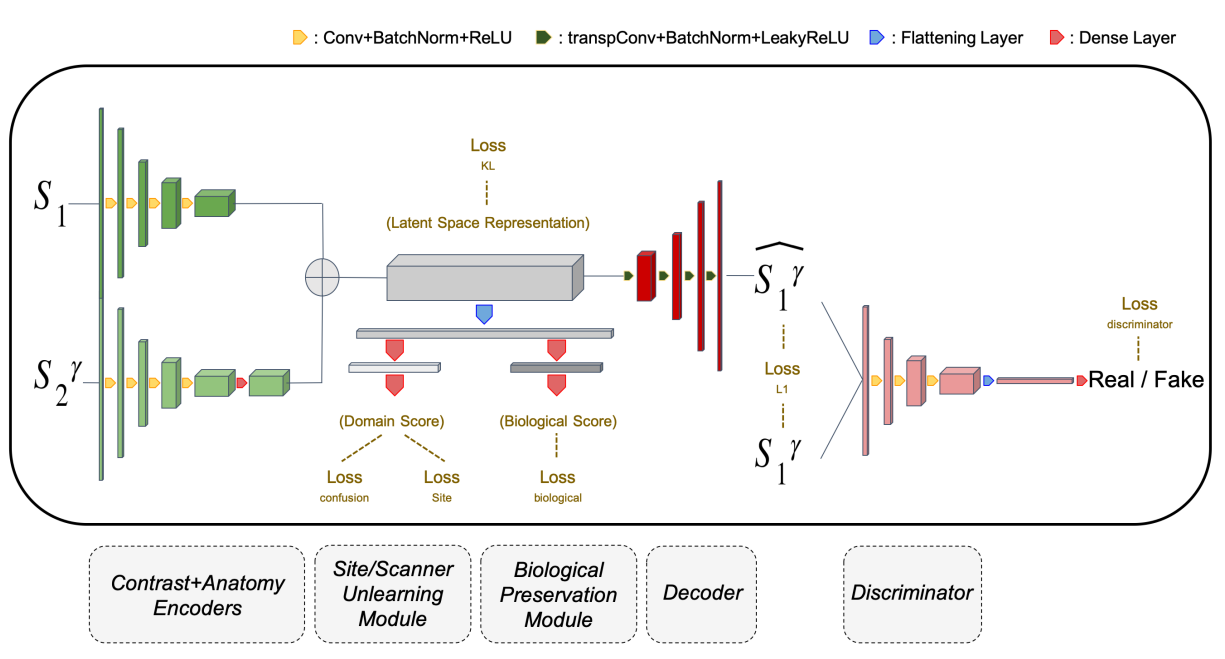

Figure 1: ImUnity's architecture. The model involves: a modified VAE generator (2.2.1), a CNN discriminator, an additional Site Unlearning module (2.2.2) and an optional Biological module (2.2.3).

2.2   ImUnity's model

The architecture of our model derives from convolutional VAE-GANs and is described in Figure 1. We adopted adversarial settings to ensure realistic outputs using a classical CNN as discriminator. The generator (here a VAE) learns how to represent input data into a lower dimension latent space (bottleneck). Information is then decoded to generate an output image. Inspired by Dinsdale et al. [2020], an unlearning center-bias module is connected to the bottleneck to limit the impact of site or scanner information. A biological preservation module can be inserted to maintain biological information in the latent space representation. Technical details are provided below.

2.2.1   Modified VAE generator

Inspired by Zuo et al. [2021], our generator takes two 2D-structural images as input, randomly taken at two different locations in the 3D-MR stack of images of each subject to consider. The first (S_1) image is used by the first CNN to encode the 'anatomical' information using only convolutional filters to ensure preservation of spatial information. The

second image (S_2), different from S_1 because randomly taken in another part of the brain, provides the initial 'contrast' information. Thus, S_1 and S_2 have different anatomy (different location in the brain) but similar contrast (same scan). S_2 contrast is modified using a gamma function with a random 'gamma' parameter sampled uniformly between 0.5 and 1.5 for each new input image. This modified S_2^γ slice is used as input to a second CNN to encode the 'contrast information' followed by a dense layer to reduce spatial information. An example of different gamma transformations applied to MR brain scans from the same subject is given in Figure 2. Once encoded, the two independent representations of S_1 and S_2^γ are concatenated to give a latent space representation which is decoded to create the output $\hat{S_1^\gamma}$ using transposed convolutional filters. Eventually, this output is compared to the reference gamma modified slice S_1^γ. Note that this generator is trained in a self-supervised fashion as it generates its own outputs. It does not require additional information such as scanner, center or biological information.

2.2.2 Site/Scanner-bias unlearning module

To ensure the task of "removing site or scanner bias", a module is directly connected to the encoders' outputs (latent space representation of inputs). The module can be seen as a domain (site or scanner) discriminator and is trained independently from the encoder to predict the scan's origin based on the latent space representation. On the other side, the encoder is trained in an adversarial fashion. A confusion loss is used to unlearn domain information. This principle has been introduced in the field of domain adaptation by Ganin et al. [2016] and has been adapted to medical imaging studies by Dinsdale et al. [2020]. Originally, the module was incorporated directly in the model to unlearn datasets bias and to improve predictions. Here, it is used in the bottleneck as a "datasets bias filter", forcing the encoder to learn a domain-invariant data representation. Thus, the generator learns a shared latent space that encodes all information needed to generate harmonized scans. Note that no skip connection is used in the generator architecture, preventing the presence of datasets bias in the final output. The loss function for the site/scanner unlearning module is:

$$l_{site}(P, Y) = -(1/N) \sum_{i=1}^{N} \sum_{s=1}^{S} \mathbb{1}(y_i = s) log(p_i^s) \tag{1}$$

While the confusion loss used in the encoders' training is :

$$l_{confusion}(P) = -(1/N) \sum_{i=1}^{N} \sum_{s=1}^{S} log(p_i^s)/S \tag{2}$$

Here $P = [p_1; ...; p_S]$ is the softmax output from the module, corresponding to the probability to belong to different sites (1, ..., S) , Y is the ground truth site affiliation vector, and N is the sample size.

2.2.3 Biological preservation module

An optional module ensures the "preservation" of biological information. It acts as a classifier of available biological information. For instance, features such as age or presence of diseases can be introduced. Contrary to the unlearning module, the encoder is trained to minimize its loss function. This module is not mandatory, and a fully self-supervised learning model can be adopted if it is turned off. The loss function of the biological preservation module for our particular application using the ABIDE database is:

$$l_{biological}(P, Y) = -(1/N) \sum_{i=1}^{N} \sum_{f \in features} \sum_{i=1}^{N} y_i^f log(p_i^f) + (1 - y_i^f) log(1 - p_i^f) \tag{3}$$

Here P represents module's predictions for biological features of interest, N is the sample size while Y is the ground truth vector. Note that in this study the binary cross entropy formulation was used for the loss function because only two features (age and patient status, i.e. ASD) were considered.

2.3 Training

Training the model involves several independent steps, due to the adversarial context and the use of the additional modules.

- Training the discriminator consists in minimizing binary cross-entropy $l_{discriminator}$ between its predictions and the labels corresponding to the nature of the inputs (real or fake). Adversarially, the generator learns how to maximize this loss function, forcing the generation of realistic outputs.

- Training the site/scanner unlearning module consists in minimizing the categorical cross-entropy (eq. 1) between its predictions and the site-affiliation labels. Adversarially, the generator is trained to minimize the confusion loss (eq. 2). It forces a site and scanner invariant representation of the dataset in the latent space, leading to uniform outputs of the unlearning module.

- In our ABIDE experiment, training the biological preservation module consisted in minimizing binary cross-entropy losses associated with each biological feature taken into account (here sex and patient status). Unlike the previous module, the loss $l_{biological}$ was directly integrated in the generator. This ensures the conservation of biological features in the latent space.

- Besides previous loss functions involved in training the generator, a l_1 loss function is used to ensure a good mapping between input $(S_1; S_2^\gamma)$ and the generated output $\hat{S_1^\gamma}$ ($l_1 = mean(|\hat{S_1^\gamma} - S_1^\gamma|)$). Moreover, the use of the Kullback-Leibler divergence l_{KL} between features distributions and a Gaussian distribution ensures a dense data representation in latent space. The global generator's loss function to minimize is therefore:

$$l_{generator} = -\lambda_1 l_{discriminator} + \lambda_2 l_{confusion} + \lambda_3 l_{biological} + \lambda_4 l_1 + \lambda_5 l_{KL} \qquad (4)$$

Where λ factors control the relative contribution of each loss. In our study we used : $\lambda_1 = 1$; $\lambda_2 = 1$; $\lambda_3 = 1$; $\lambda_4 = 100$; $\lambda_5 = 10^{-3}$ found empirically.

2.4 Experiments

The datasets extracted from the three databases were used to evaluate different aspects of our model. Impact on image quality in multi-site or multi-scanner harmonization was assessed using data from traveling subjects (ground truth) from the OASIS and SRPBS datasets. Ability to remove site information was evaluated using the ABIDE dataset. Finally, benefits of harmonization between data provider centers were assessed using the predictions of autism disorder in children from the ABIDE dataset. To demonstrate the flexibility of ImUnity, all experiments were performed with the same model trained on data coming from the ABIDE database, unless specified. OASIS and SRPBS were then used for the validation parts only. The model was trained on 2D axial slices with at least 1% of brain tissue voxels. Training was run on a Nvidia GeForce 2080 RTX for 2000 epochs using a learning rate of 10^{-4} and Adam optimizer.

2.4.1 Experiment 1 : Harmonization on traveling subjects (OASIS+SRPBS)

We first evaluated the ability of our model to transform images from one domain (site or scanner) to their equivalent in another domain. As SRPBS and OASIS databases contain traveling subjects, ground truth was available to assess ImUnity performances. In practice, one domain (acquisition site or scanner) was first selected as the reference for every subject. Individual scans were co-registered to their equivalent in the reference domain (to avoid variations between acquisitions due to movement). Then, all the images were transformed by the model into the reference domain. During this step, slices to be harmonized (anatomy) were fed to the model alongside the corresponding computed contrast slice from the reference domain. Finally, results obtained after transformation were compared to the ground truth, i.e. images acquired in the reference domain (traveling subject). Visual verification, histograms of image intensities, and the Structural Similarity Index Metric (SSIM, Wang et al. [2003]) were used to assess image likeness. The same model, trained on ABIDE data, was used for every site/scanner of the two other databases to evaluate the ability of ImUnity to generalize to sites never seen before. This experiment also evaluates ImUnity's versatility, either for the source domain or for the target domain (last two columns of Table 1).

2.4.2 Experiments 2 : Harmonization's effects on sites classification (ABIDE)

The second experiment evaluated the ability to detect the origin of data before and after harmonization. As no ground truth was available for this experiment, we considered harmonization impacts on classification algorithms. Standard Support Vector Machine (SVM) with a radial basis function kernel was used to classify ABIDE data. The classifier worked on all radiomic features (N=101) extracted using the pyradiomics python API [van Griethuysen et al., 2017]. These features aim to represent different aspects of MRI images such as shape, contrast or texture and are known to be sensitive to site effects [Orlhac et al., 2019]. The most 'correlated features' with sites affiliations before harmonization were selected for classification using Pearson tests (ran independently for each feature) using 10^{-3} as p-value threshold
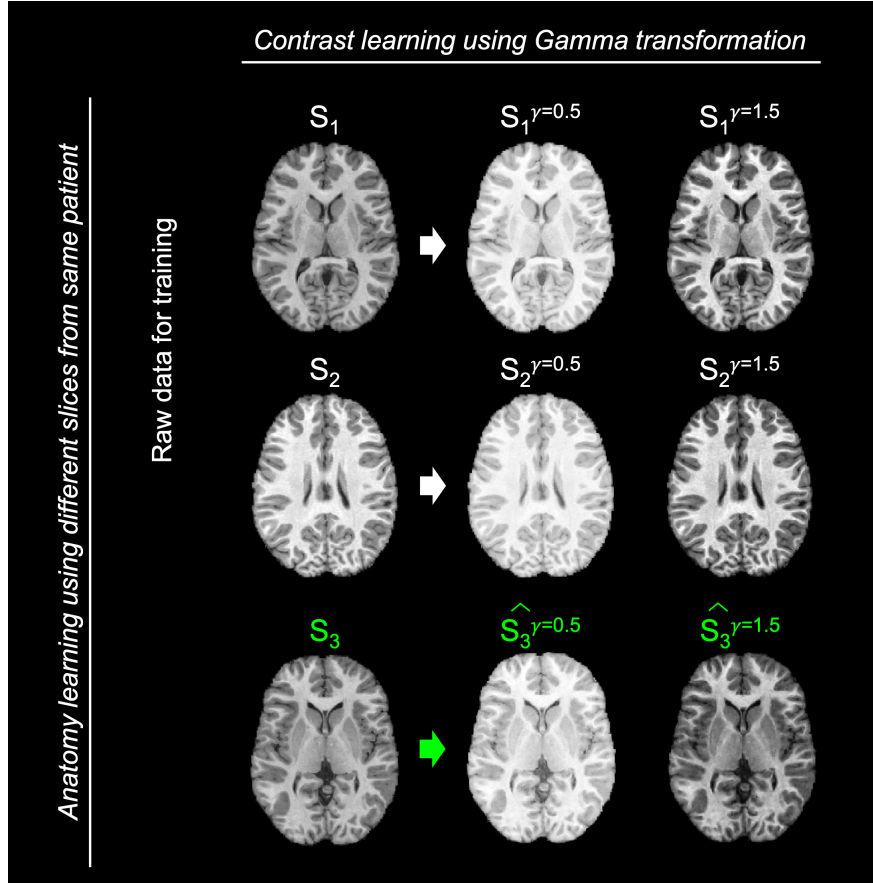
Figure 2: Inputs and outputs of the model: Different slices from the same patient are used to encode the anatomical information. Gamma transformations are used to encode the contrast information. Rows represent different anatomical slices (S_1, S_2 and S_3) taken from the same subject. The first two rows present Gamma transformations used to train the model (S_1^γ and S_2^γ). The last raw shows model outputs ($\hat{S}_3^\gamma$) for estimated Gamma transformations for the slice S_3 not present in the training set. From left to right columns: original slices, Gamma modified slices with parameter 0.5, Gamma modified slices with parameter 1.5.

(30 features in total). Accuracy and Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve were used to evaluate the specificity and sensitivity of the site classifier.

2.4.3 Experiments 3 : Harmonization's effects on autism syndrome disorder prediction (ABIDE)

Similarly to Experiment 2, Experiment 3 evaluated the ability of our classifier to detect patients with ASD from the ABIDE database, before and after harmonization. Here, results were obtained following a 10 fold-cross-validation procedure. The same trained model was used for different number of sites (and for different combinations of sites) included in the ABIDE database.

3 Results

Experiment 1: Figure 3-A shows the results obtained for one traveling subject from the SRPBS database (2.4.1). Images are shown for one acquisition at site (A) before harmonization (3-A, left), corrected by ImUnity to fit with acquisition at site B (3-A middle) and the corresponding ground truth acquired at site B (3-A right). One can notice the difference in image contrast between the 2 sites, highlighting the need for image harmonization, as well as the visual similarity between the harmonized image and the ground truth. It is interesting to observe that the anatomical structures of
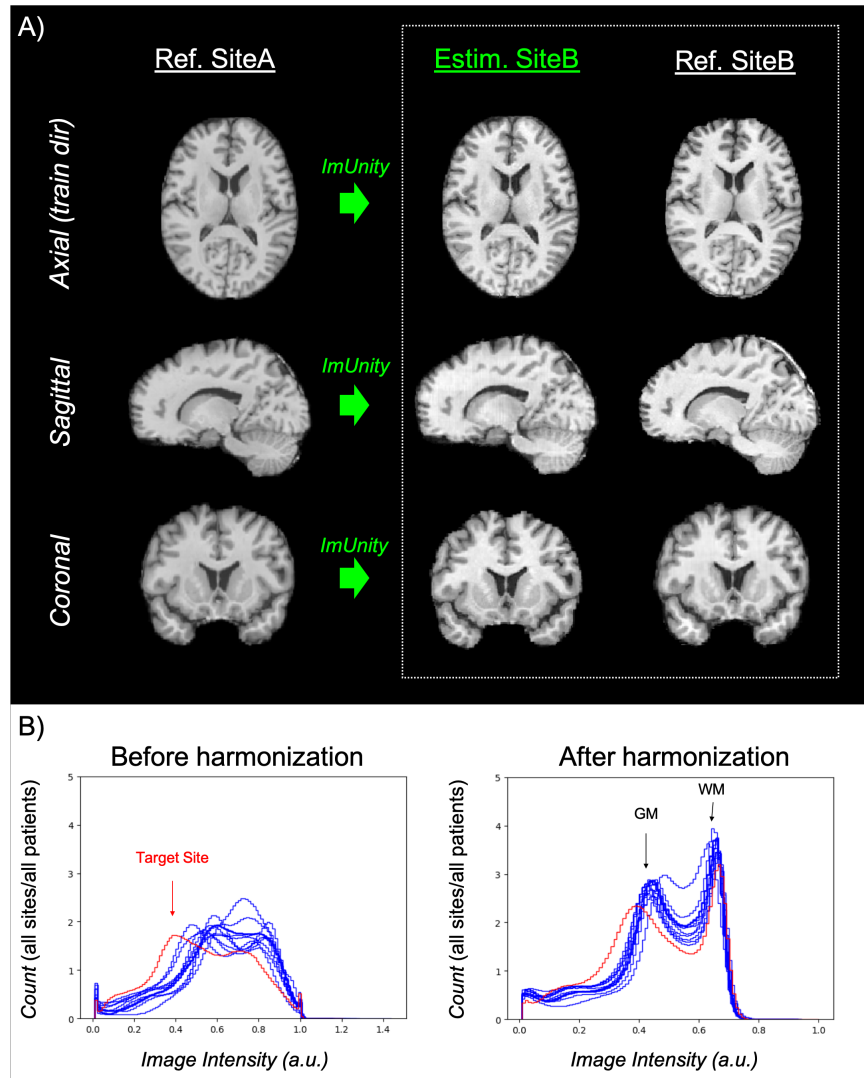
Figure 3: Harmonization result on travelling subjects from the SRPBS database. A) Left: 3D images from one patient (axial, sagittal, coronal views) acquired in site A before harmonization, Middle: ImUnity's harmonization to fit with acquisition at site B, Right: image acquired at site B (ground truth). Note that the ImUnity model was trained on axial slices only. B) Images intensity distributions (all patients) before (Left) and after ImUnity's harmonization (Right). The red histogram corresponds to the site taken as reference for the harmonization process (target site).
GM = gray matter; WM = white matter

the input contrast reference are not propagated through the model, which explains small anatomical differences (e.g. superior sagittal sinus) between the model estimates and the ground truth. It is also worth noting that although the model was trained on 2D-axial slices, the 3D reconstructions of the estimates are of high quality in each orientation. Figure 3-B shows the effects of ImUnity's harmonization on image intensity distributions for all selected subjects from the SRPBS database. The model was used to harmonize every image to a target site (indicated in red). An alignment of histograms can be clearly observed after harmonization, with both gray and white matter peaks shifted. Changes in intensity distribution of the site of reference are due to pre-processing (see details in Supplementary Material (SM), Figure 6, top row). Images obtained after ImUnity's harmonization of the OASIS datasets are provided in SM, Figure 7.

Quantitative results obtained with the SSIM metric in all traveling subjects are summarized in Table 2. Both multi-site (SRPBS) and multi-scanner (OASIS) experiments are shown. For the latter, results from the literature are also given for reference. It can be seen that ImUnity increases the structural similarity in all cases and provides better performances compared to other deep learning approaches. Moreover, results from multi-scanner harmonization show that ImUnity

Task	Multi-scanner harmonization.		Multi-site harmonization
Dataset	OASIS scanner F $\rightarrow$ scanner E	OASIS all scanners $\rightarrow$ scanner E	SRPBS all sites $\rightarrow$ UTO site
Raw data	0.871 ± 0.045	0.845 ± 0.059	0.853 ± 0.021
Zhu et al. [2018] CycleGAN*	0.873 ± 0.046	-	-
Zuo et al. [2021] Calamity*	0.884 ± 0.046	-	-
ImUnity#	0.920 ± 0.024	0.919 ± 0.023	$\mathbf{0.907 \pm 0.024}$
ImUnity*	$\mathbf{0.943 \pm 0.003}$	$\mathbf{0.943 \pm 0.003}$	0.893 ± 0.025
ImUnity+	0.824 ± 0.064	0.827 ± 0.061	0.865 ± 0.019

Table 2: SSIM in travelling subjects for multi-scanner (OASIS database) and multi-site (SRPBS database) harmonization. Results are compared to the literature when available.
$*$: Model trained on OASIS database (n=1072); # : Model trained on ABIDE database (n=545); $+$: Model trained on SRPBS database (n=81)

performs well independently of the chosen reference domain. The last 2 lines present results obtained after training ImUnity on OASIS (to better match literature protocols) and SRPBS data. These models were used to harmonize OASIS as well as SRPBS data.

Experiment 2: Figure 4 shows ImUnity's harmonization effects on site classification on the ABIDE datasets (2.4.2) using tSNE [Maaten and Hinton, 2008], a dimension reduction algorithm, on radiomic features. Before harmonization, the presence of site clusters is clear. Once the data is harmonized using ImUnity, the points are shuffled and the accuracy of the SVM site prediction decreases from 0.70 to 0.38 (before and after harmonization respectively). This confirms the removal of site bias by ImUnity as the classifier is no longer able to correctly separate the sites. Additional results on the influence of the preprocessing step on sites classification are provided in SM Fig. 6.

Experiment 3: Figure 5 shows the capacity of our model to improve ASD prediction from the ABIDE datasets. Here, we used the same trained model to test the influence of different numbers of sites included in the database (from 2 to 11) as well as different combinations of those sites (for example 55 combinations of 2 sites taken among the 11 sites available). In every case, we observed a clear improvement of classification of autistic patients after harmonization as shown by increases in AUC provided by the SVM classifier. We show the results obtained with the best combination of sites as well as average and standard deviation of AUC with all combinations of sites. The preprocessing also has a positive impact on the prediction as shown in SM (Fig. 6, bottom row).

4   Discussion

We have presented ImUnity, an original harmonization tool for multi-center MRI databases. ImUnity shows high performances in term of quality of the generated harmonized images, as well as clear removal of the idiosyncratic bias attached to site-dependent image acquisition conditions. Moreover, the performed experiments clearly demonstrate ImUnity's versatility. By training ImUnity's model on datasets extracted from one database (here ABIDE) and looking at images harmonized from traveling subjects provided by two different databases (here OASIS and SRPBS) , we showed that ImUnity did not require new training phase to generalize to unseen sites or scanners (see Fig. 3). The performances were maintained independently of the site selected as reference (see Table 2). While the model was trained on ABIDE data only, it provided better results than the state-of-the-art methods in terms of image quality (+4%, see Table 2).

The last two rows of Table 2 still present SSIM metrics obtained when training ImUnity on the other 2 datasets (OASIS and SRPBS). As no biological features were available in these databases, the biological module was disabled and the model was trained in a self-supervised way. First, we noted additional improvements for scanner harmonization when the model was trained and applied on the same database (here OASIS, +2.5%). Second, the score obtained for multi-site harmonization (SRPBS) was the highest when trained on ABIDE (N = 545) data (with a slightly better score than with OASIS). It is interesting to observe the impact of the dataset size, the number of site/scanner involved in the training, the use of the biological module and the anatomical differences between datasets (ABIDE mainly contains children data while OASIS and SRPBS focus on adults) on these scores. ABIDE result suggests that anatomical differences could be compensated by a large training dataset presenting more site/scanner variability than OASIS (11 sites for ABIDE vs. 4 sites for OASIS). On the other hand, results from multi-scanners harmonization depict the difficulty of the model
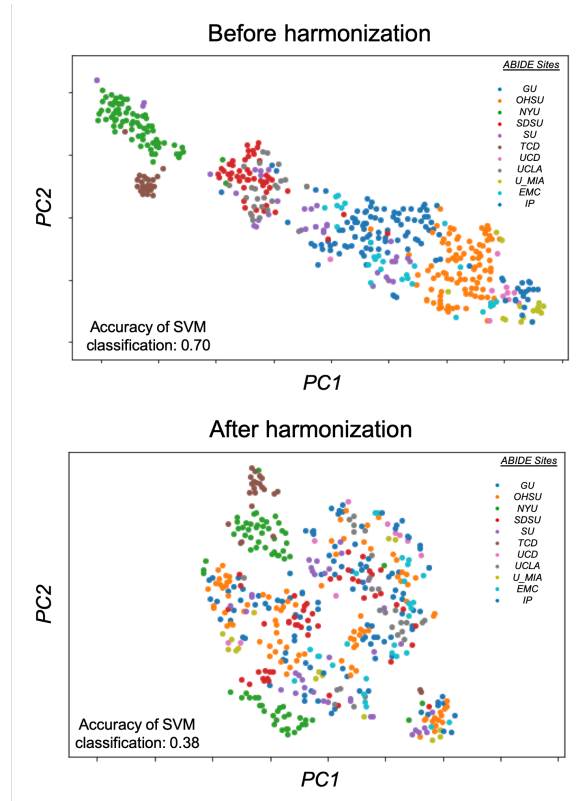
Figure 4: Harmonization effects on ABIDE sites classification. 2D scans representation of ABIDE database using tSNE reduction algorithm are presented before and after harmonization. Colors correspond to different sites. The separation of data provided by the different sites is clearly more difficult after harmonization.

trained on SRPBS data to generalize its training to the OASIS data. This indicates an over-fitting effect in this situation, as there was not enough training data (here N=81) and suggests that ImUnity may not be adapted to small sample size scenarios.

In addition to the ability to remove sites or scanners biases, we evaluated the ability to improve subsequent clinical data analysis. The performances of ASD classification obtained on the ABIDE database were improved by ImUnity's harmonization of the datasets even when a simple classification procedure (classical SVM and radiomic features derived from structural MRI) was used. It is interesting to note that several studies have already tackled the issue of ASD patients classification using the ABIDE database [Gao et al., 2020, Sherkatghanad et al., 2019]. They used either more complex features (e.g. morphological brain networks) or more complex classifiers (Random Forest, Resnet, adapted deep-learning architecture, etc.) than in our study. However, the reported AUC metrics were close to ours (0.67 by Gao et al. [2020] and 0.75 by Sherkatghanad et al. [2019]). They also noted the difficulty of obtaining better results when more acquisition sites where included in the study. They did not however mention any harmonization procedure. Thus, we may expect an improvement of ImUnity's performances when introducing more informative radiomic features and choosing a more sophisticated classifier tool.

Because ImUnity is designed to reconstruct images and to create a new harmonized database, it does not need new training for new clinical or biological questions. Beyond classification, new clinical data investigation should be conducted with ABIDE (or other multi-center clinical databases) to have better understanding on the impact of our method on clinical research studies.

Like the majority of deep networks used for medical image analysis, the MR images used as inputs of our network were first pre-processed for intensity normalization, co-registration or brain extraction. Usually, the impact of these transformations is not examined in harmonization studies. In Fig. 6 (SM), we have highlighted the fact that these steps were already able to remove some of the sites and scanners biases with positive impacts on intensity distributions across sites or patients classification. Intrinsically, the use of White-Stripe normalization [Shinohara et al., 2014] forces the alignment of intensity distributions. Yet, a perfect alignment is not the ultimate goal of harmonization as we also
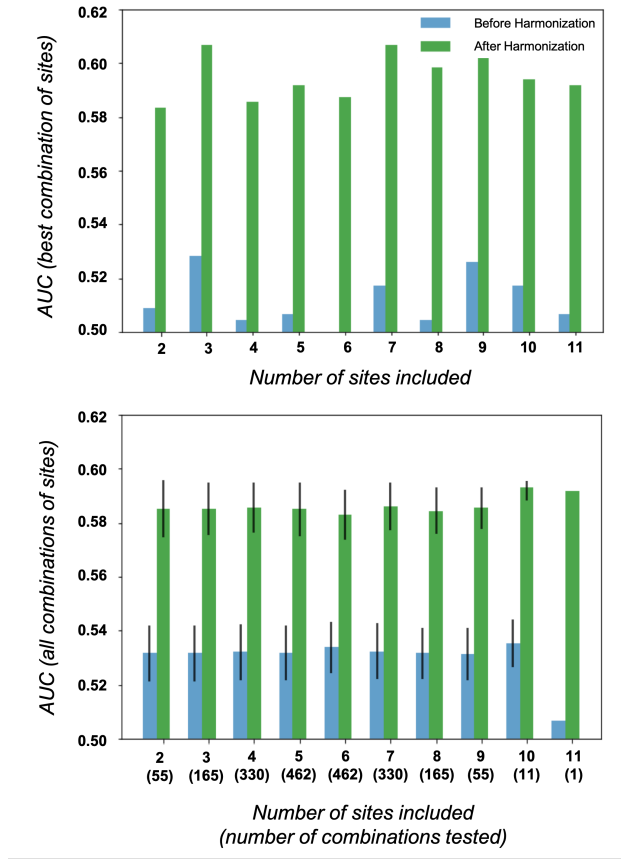
Figure 5: Harmonization effects on ABIDE patients classification. AUC metric for classification of patients with autism spectrum disorder (ASD) using SVM and extracted radiomic features (Experience 3, 2.4.3), are shown for different numbers of sites included in the database (from 2 to 11 sites).
Top row: Results obtained from the best (largest change in AUC before and after harmonization) combination of sites.
Bottom row : Average and standard deviation of AUC estimates for all combination of sites. Same trained model and harmonized data were used for different site combinations.

seek to preserve informative biological variations which vary independently across sites. Eventually, we observed that the best results were obtained for all experiments after the whole ImUnity process, with better intensity distributions alignments, removal of persistent datasets noises, and most importantly improvement in patients classification results. On the opposite, other experiments (not reported) also showed that the VAE-GAN network alone performed poorer when the pre-processing steps were omitted, suggesting that these steps are necessary to simplify the training process and improve generalization of the results.

In our study, we only worked with anatomical T1-weighted images. We showed that a single type of sequence, combined with computed image transformations with the Gamma function, are sufficient to learn contrast mapping. This greatly facilitates the use of our model because of few data requirements (the origin of each scan is the only pre-required information) and the possibility of self-supervised training. Yet, we believe that this approach is not only dedicated to T1 contrast harmonization and can easily be generalized to any MRI sequences. Presently, the model needs to be fine-tuned in order to harmonize a new medical imaging type. It could however be interesting to investigate its capacity to learn how to harmonize multiple sequences at once. This could be done by mixing sequence types in our training dataset and ensuring the conservation of this information by adding a new conservation module in the bottleneck. It could also be interesting to add other types of artificial contrast transformation for our training in order to account for other types of sites or sequences biases.

5 Conclusion

We presented ImUnity, an original and effective tool dedicated to MRI harmonization. Our proposed model derives from the VAE-GAN architecture. It ensures realistic outputs and allows removal of idiosyncratic datasets bias and the preservation of biological information. Our results show that the method reaches state-of-the-art results in term of image quality in traveling patients of the OASIS and SRPBS databases and improves autistic patients classification in the ABIDE database. The proposed model is versatile, requiring only one type of MR sequence without the need of matching subjects, can be generalized to sites unseen during the training phase and can be used to harmonize MR images to different reference domains without a new training phase.

6 Compliance with ethical standards

This research study was conducted retrospectively using human subject data made available by the following open sources: ABIDE, OASIS, SRPBS. Ethical approval was not required as confirmed by the license attached with the data.

7 Acknowledgments

References

Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. MS-Net: Multi-Site Network for Improving Prostate Segmentation with Heterogeneous MRI Data. *arXiv:2002.03366*, February 2020. URL http://arxiv.org/abs/2002.03366.

Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A. Elliott, Kosha Ruparel, David R. Roalf, Theodore D. Satterthwaite, Ruben C. Gur, Raquel E. Gur, Robert T. Schultz, Ragini Verma, and Russell T. Shinohara. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, November 2017. ISSN 1053-8119. doi:10.1016/j.neuroimage.2017.08.047. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5736019/.

Russell T. Shinohara, Elizabeth M. Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J. Mateen, Peter A. Calabresi, Samson Jarso, Dzung L. Pham, Daniel S. Reich, and Ciprian M. Crainiceanu. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin*, 6:9–19, August 2014. ISSN 2213-1582. doi:10.1016/j.nicl.2014.08.008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215426/.

Jean-Philippe Fortin, Elizabeth M. Sweeney, John Muschelli, Ciprian M. Crainiceanu, and Russell T. Shinohara. Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage*, 132:198–212, May 2016. ISSN 1053-8119. doi:10.1016/j.neuroimage.2016.02.036. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540379/.

Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M. Nasrallah, Theodore D. Satterthwaite, Yong Fan, Lenore J. Launer, Colin L. Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C. Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel H. Wolf, Raquel Gur, Ruben Gur, John Morris, Marilyn S. Albert, Hans J. Grabe, Susan M. Resnick, R. Nick Bryan, David A. Wolk, Russell T. Shinohara, Haochang Shou, and Christos Davatzikos. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage*, page 116450, December 2019. ISSN 1095-9572. doi:10.1016/j.neuroimage.2019.116450.

Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Christos Davatzikos, Yvette I. Sheline, Russell T. Shinohara, and Kristin A. Linn. Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, 220:117129, October 2020. ISSN 1053-8119. doi:10.1016/j.neuroimage.2020.117129. URL http://www.sciencedirect.com/science/article/pii/S1053811920306157.

Samuel St-Jean, Max A. Viergever, and Alexander Leemans. Harmonization of diffusion MRI data sets with adaptive dictionary learning. *Human Brain Mapping*, August 2020. ISSN 1097-0193. doi:10.1002/hbm.25117. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25117.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593*, November 2018. URL http://arxiv.org/abs/1703.10593.

Blake E. Dewey, Can Zhao, Jacob C. Reinhold, Aaron Carass, Kathryn C. Fitzgerald, Elias S. Sotirchos, Shiv Saidha, Jiwon Oh, Dzung L. Pham, Peter A. Calabresi, Peter C. M. van Zijl, and Jerry L. Prince. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magn Reson Imaging*, July 2019. ISSN 1873-5894. doi:10.1016/j.mri.2019.05.041.

Lianrui Zuo, Blake E. Dewey, Aaron Carass, Yihao Liu, Yufan He, Peter A. Calabresi, and Jerry L. Prince. Information-based Disentangled Representation Learning for Unsupervised MR Harmonization. *arXiv:2103.13283 [cs, eess]*, March 2021. URL `http://arxiv.org/abs/2103.13283`. arXiv: 2103.13283.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. URL `http://arxiv.org/abs/1505.04597`.

Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal. *bioRxiv*, page 2020.10.09.332973, October 2020. doi:10.1101/2020.10.09.332973. URL `https://www.biorxiv.org/content/10.1101/2020.10.09.332973v1`. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Hao Guan, Yunbi Liu, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis*, 71:102076, July 2021. ISSN 1361-8415. doi:10.1016/j.media.2021.102076. URL `https://www.sciencedirect.com/science/article/pii/S1361841521001225`.

A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry*, 19(6):659–667, June 2014. ISSN 1476-5578. doi:10.1038/mp.2013.78.

Pamela J. LaMontagne, Tammie LS Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. Technical report, December 2019. URL `https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1`. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press ISSN: 1901-4902 Label: Cold Spring Harbor Laboratory Press Type: article.

Saori C. Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, Ryuichiro Hashimoto, Go Okada, Yuki Sakai, Jun Morimoto, Jin Narumoto, Yasuhiro Shimada, Hiroaki Mano, Wako Yoshida, Ben Seymour, Takeshi Shimizu, Koichi Hosomi, Youichi Saitoh, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Yasumasa Okamoto, Okito Yamashita, Mitsuo Kawato, and Hiroshi Imamizu. A multi-site, multi-disorder resting-state magnetic resonance image database. *Sci Data*, 8(1):227, August 2021. ISSN 2052-4463. doi:10.1038/s41597-021-01004-8. URL `http://www.nature.com/articles/s41597-021-01004-8`. Bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;Neural circuits;Neurological disorders;Psychiatric disorders Subject_term_id: diagnostic-markers;neural-circuit;neurological-disorders;psychiatric-disorders.

Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M. Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging*, 30(9):1617–1634, September 2011. ISSN 1558-254X. doi:10.1109/TMI.2011.2138152.

Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, 29(6):1310–1320, June 2010. ISSN 1558-254X. doi:10.1109/TMI.2010.2046908.

Carmen E. Sanchez, John E. Richards, and C. Robert Almli. Age-specific MRI templates for pediatric neuroimaging. *Dev Neuropsychol*, 37(5):379–399, July 2012. ISSN 8756-5641. doi:10.1080/87565641.2012.688900. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3399736/`.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2016. URL `http://arxiv.org/abs/1505.07818`. arXiv: 1505.07818.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, November 2003. doi:10.1109/ACSSC.2003.1292216.

Joost J. M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G. H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J. W. L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.*, 77(21):e104–e107, 2017. ISSN 1538-7445. doi:10.1158/0008-5472.CAN-17-0339.

Fanny Orlhac, Frédérique Frouin, Christophe Nioche, Nicholas Ayache, and Irène Buvat. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology*, 291(1):53–59, 2019. ISSN 1527-1315. doi:10.1148/radiol.2019182023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9 (Nov):2579–2605, 2008. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Jingjing Gao, Mingren Chen, Yuanyuan Li, Yachun Gao, Yanling Li, Shimin Cai, and Jiaojian Wang. Multisite Autism Spectrum Disorder Classification Using Convolutional Neural Network Classifier and Individual Morphological Brain Networks. *Front Neurosci*, 14:629630, 2020. ISSN 1662-4548. doi:10.3389/fnins.2020.629630.

Zeinab Sherkatghanad, Mohammadsadegh Akhondzadeh, Soorena Salari, Mariam Zomorodi-Moghadam, Moloud Abdar, U. Rajendra Acharya, Reza Khosrowabadi, and Vahid Salari. Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network. *Front Neurosci*, 13:1325, 2019. ISSN 1662-4548. doi:10.3389/fnins.2019.01325.

Supplementary Material
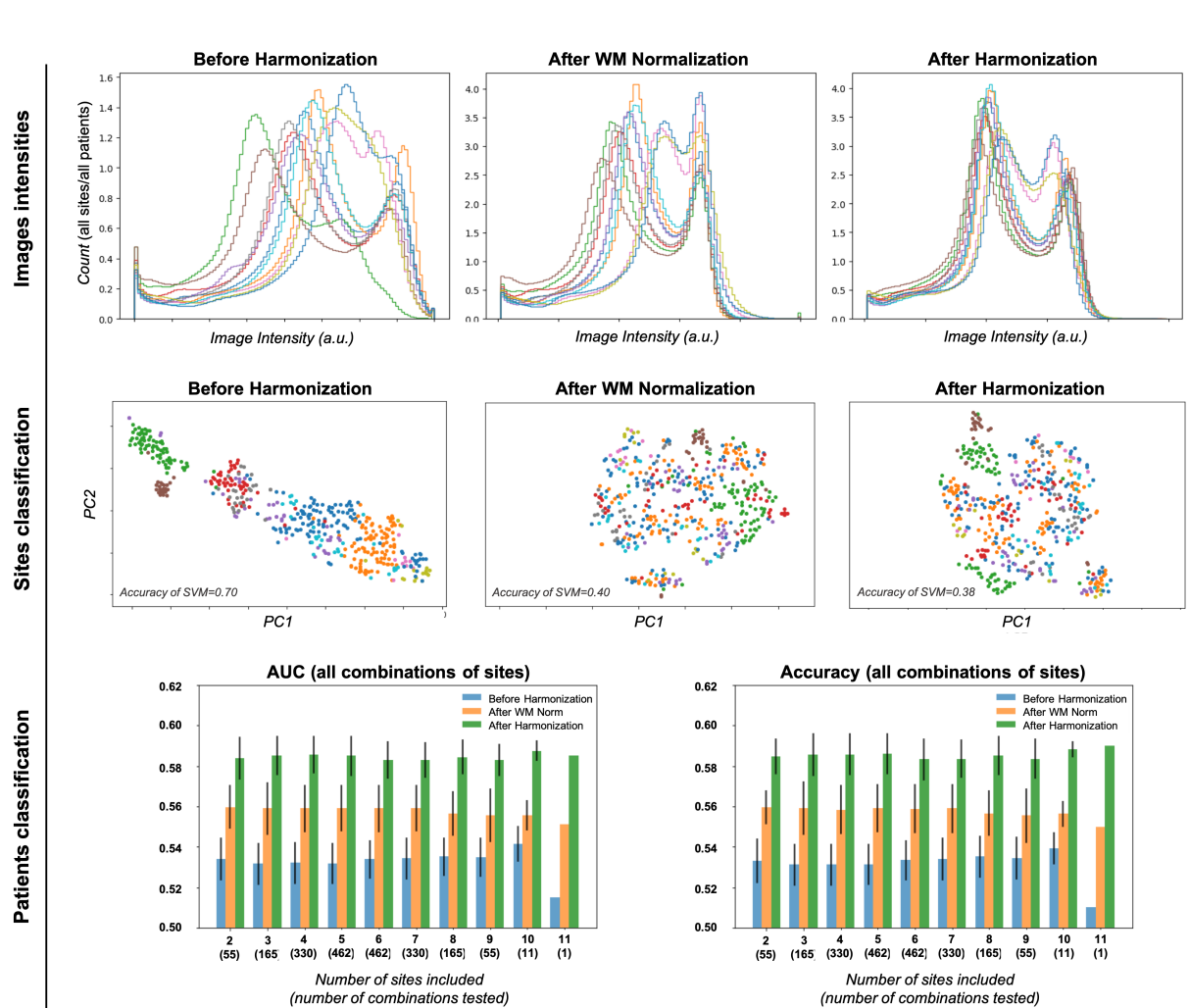
A   Impact of pre-processing steps

Figure 6: Impact of pre-processing steps (N4Biais, White-Stripe normalization) on our different experiments. From top to bottom row: impact on images intensity, impact on sites classification (2.4.2), impact on patients classification (2.4.3)
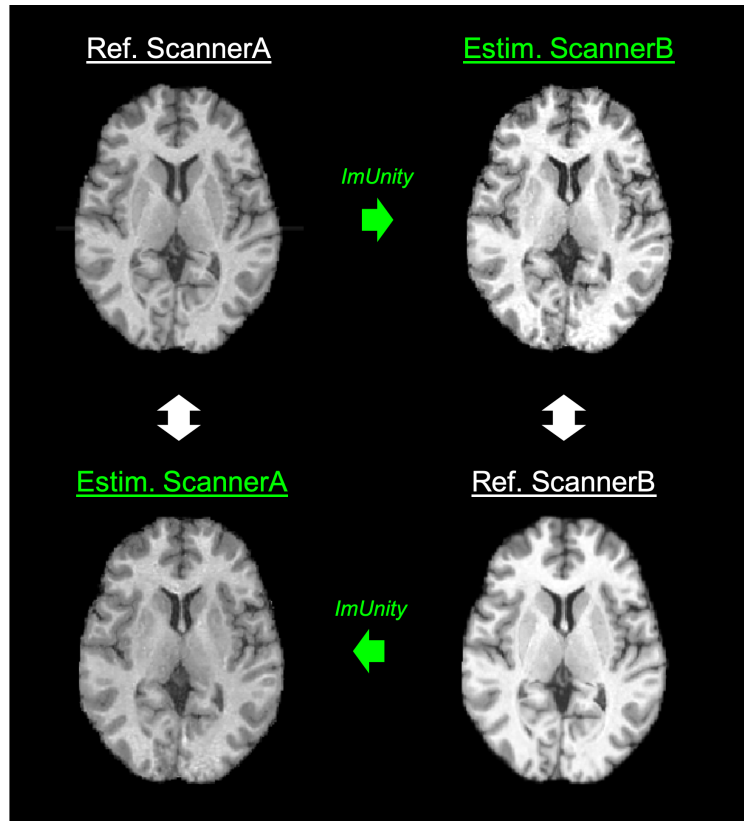
B Scanner harmonization

Figure 7: Multi-scanner harmonization (2.4.1) results between 2 scanners for the same subject extracted from the OASIS database. ImUnity's model was trained on datasets extracted from the ABIDE database. Top row: Slice acquired at site A (left) and corresponding harmonized image (right) matching acquisitions at site B. Bottom row: Slice acquired at site B (right) and corresponding harmonized image (left) matching acquisitions at site A. Left (resp. right) column allows to visually compare the ground truth and the estimated image for site A (resp. for site B).