Sacha Dedeken

Normalize MRI for robust radiomics in oncology using Deep Learning

supervised by John Klein, Jérémie Boulanger, David Pasquier, Alexandre Escande
Associate Professor, Associate Professor, Radiotherapist oncologist, Radiotherapist oncologist, CRIStAL, Centre Oscar Lambret

Abstract : Magnetic Resonance Imaging (MRI) is a major tool to allow diagnosis and prognosis of malignant tumors. Machine learning algorithms have successfully developed predictive models on MR images. However MR images of the same or different patients acquired using different devices or with different acquisition parameters can have large variations in intensity. This lack of standardization greatly hinders the reproducibility and generalization power of algorithms trained on a restricted dataset. This research project propose deep learning methods to reduce inter-machine variability of MRI intensities. The contributions are the cleaning of an new dataset adapted to the task, the development of a toolbox to different experiments and the design of an experimental protocol with a more robust evaluation criteria.

Keywords : *Radiomics, Generative Adversarial Networks, MRI, Normalization*

2021-2022 M2 Data Science

Contents

1 Context

1.1 Teams presentation

This research project is the result of the interaction between the CRIStAL laboratory (Centre de Recherche en Informatique, Signal et Automatique de Lille) and the Oscar Lambret Center. CRIStAL is a member of the interdisciplinary research institute IRCICA - USR CNRS 3380. CRIStAL appeared in January 2015 and is since 2018 under the supervision of the University of Lille, CNRS and Centrale Lille Institut, but also of Inria Lille Nord Europe and IMT Lille-Douai. Composed in 2022 of more than 400 members, the laboratory's research activities concern themes related to the major scientific and societal issues of the moment such as: BigData, software, image and its uses, human-machine interactions, robotics, control and supervision of large systems, intelligent embedded systems, bioinformatics... with applications in particular in the sectors of trade industry, health technologies, smart grids [1]. The Oscar Lambret Center is the Regional Cancer Center of Hauts-de-France, located in Lille. This private health establishment of collective interest groups together oncology services, numerous radiotherapy facilities, teaching and training units and also, since 2006, an integrated clinical research unit in oncology [2].

1.2 MRI principle

Magnetic resonance imaging (MRI) is a completely painless radiological examination that provides "sliced" images of the body. Unlike the scanner, which uses X-rays, the images obtained by magnetic resonance are the result of the interaction between the body's natural magnetism and that of the machine (in which there are large magnets). The magnetic field, by the energy it brings, orients all the hydrogen atoms that make up the tissues in the same direction. When the field is stopped, these atoms return to their initial state by giving back this energy. All the tissues of the body do not contain the same proportion of hydrogen. Therefore, the energy levels returned will be different from one organ to another. MRI provides images with high spatial resolution, from any angle and in 2 or 3 dimensions. It is applicable to the whole body and allows to visualize all tissues of the body. The succession of excitation pulses gradient fields constitutes what is called a pulse sequence. Depending on the tissue one wants to highlight, different pulse sequences can

be applied. In particular, two types of images can be distinguished: the T1-weighted images highlight fat tissue within the body, while T2-weighted images highlight both fat and water. MRI is a common examination, increasingly used in the diagnosis and follow-up of cancers. As with a CT scan, a contrast agent, gadolinium, may first be injected prior to an MRI. The patient is then asked to undress completely and lie down on the machine table. The microphone and camera system allows the medical team to hear and see the patient. Depending on the organs to be studied, an MRI scan lasts between 30 and 90 minutes [3].

Figure 1: MRI exam

After the examination, the machine uses a reconstruction algorithm to provide the radiotherapist with a 3-dimensional image, i.e. a succession of slices. The specialist, with the help of his experience, his watch in clinical research and assisted by an image exploration software, can then analyze the examination to make a diagnosis (determine the patient's disorder), a prognosis (predict the future state of a patient and his chances of recovery) and an adapted treatment. In the case of oncology, the primary challenge is to identify the existence of one or more tumors, and to assess its size, structure and various characteristics. Once the diagnosis is made, subsequent examinations can be used to monitor the evolution of the cancer (does the patient show an objective response to treatment, tumor progression or neither). Of course, all these decisions are always guided by information other than the examination alone, such as blood tests, biopsies or patient statements.

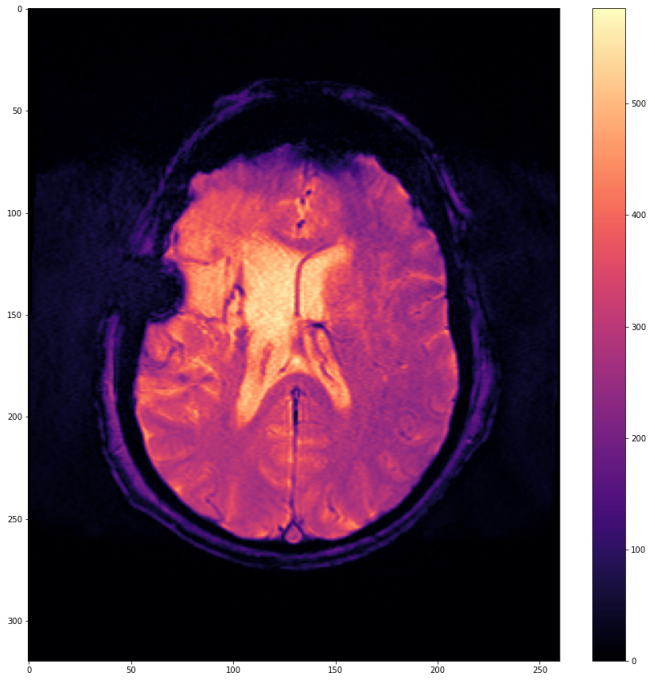

Figure 2: Example of MRI slice with a tumor on the left

1.3 Emergence of radiomics

Since the beginning of the 2010s, research in the area of medical imaging has seen a new field emerge, radiomics. The use of this neologism, composed of the prefix radio- (for radiology) and the suffix -omics (usually designating the field of study of the exploitation of genomic medical data), has continued to grow in scientific publications year after year. This new approach is based on two observations and a presupposition: The first observation is that the collection and analysis of MRI data is time-consuming and expensive, the second is that the radiotherapist although a specialist remains fallible, subject to fatigue and distractions. The presupposition is that the 3D image contains more useful information for analysis than the human eye is capable of extracting. The principle is to use mathematics and in particular machine learning on large MRI databases to automate diagnosis, treatment suggestion, patient follow-up and prognosis. This is part of the more global idea of precision medicine, going even further than science-based medicine and its consideration of clinical trials, by using Big Data to no longer treat each patient

as an average patient, but as an individual whose specific profile (e.g. sex, age, but also genome or imaging) will be able to guide decision-making. Radiomics can be defined as the quantitative mapping, that is, extraction, analysis and modelling of many medical image features in relation to prediction targets, such as genomic features and clinical end points [4]. The radiomics can be extracted from the histogram of the gray levels of an image, we speak then of radiomics "first order", or its matrix of co-occurrence of the gray levels, we speak then of "second order" radiomics.

Important note: in this work we have concentrated on first-order radiomics, so it will be shown mainly histograms. These histograms calculate the proportion of gray level between 0 (for black) and 32 (for white) of a slice. They are therefore equivalent to a density and the sum of the bins is 1. Their generation and pre-processing is described in the section Approach.

Contrary to a biopsy which targets a part of a tumor at a single location, the radiomics approach allows to model and exploit the information of the whole tumor on all the concerned areas, as in the case of metastases. This approach is new and still at the research stage, but it is attracting great academic, medical and industrial interest because of its therapeutic claims and the results already available and very promising, as in the case of triple-negative breast cancer [5].

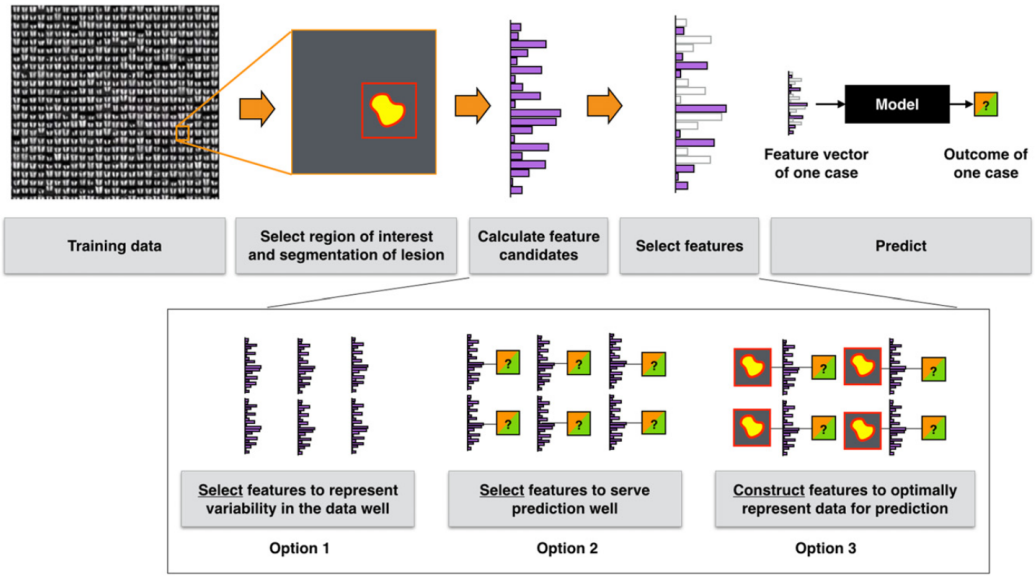

Figure 3: Radiomics workflow from Lambin et al [5]

1.4 Radiomics variability

But to be really efficient, radiomics research still has to overcome many problems. One of them is the robustness of the features created [6]. It can be defined as the variability of a score for a certain task (diagnosis, prognosis, ...) from one patient to another. The sources of this variability are very diverse: to begin with, the MRI examination being based on the electromagnetic field, it is not totally stable over time and is influenced by atmospheric conditions, which constitutes a base noise that is difficult to reduce. In addition to this, there is the inter-exam variability, from one exam to another the settings of the machine can vary and the same patient can be installed differently inside, giving a different image. There is also inter-patient variability, which is easily understood since patients differ from one another (morphology, physical condition, medical history). For radiomics from slices (therefore in two dimensions), there is also the problem of intra-patient variability, as the acquisition properties may vary from one slice to another. This variability can be represented by these graphs, which show for 6 examinations the superposition of the histograms of each slice:

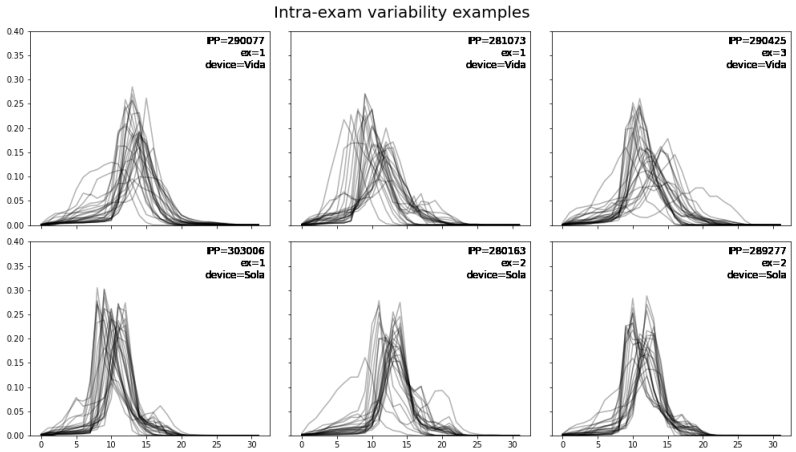

Figure 4: Intra-exam variability

Or by this one in the form of the distribution by component :

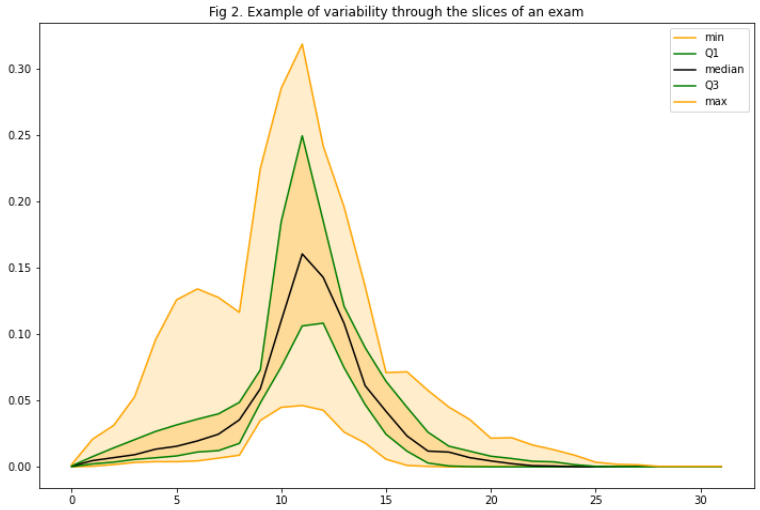

Figure 5: Intra-exam distribution

Finally, there is inter-machine variability, the examination of the same patient on two different devices will present significant variations, because the software for reconstruction of the 3D image from electromagnetic measurements depends on the manufacturer (for example Siemens or Philips). The generated image may vary in gray level intensity, brightness, contrast or even noise. This can be seen very well by viewing the average histogram of the examinations of each machine:

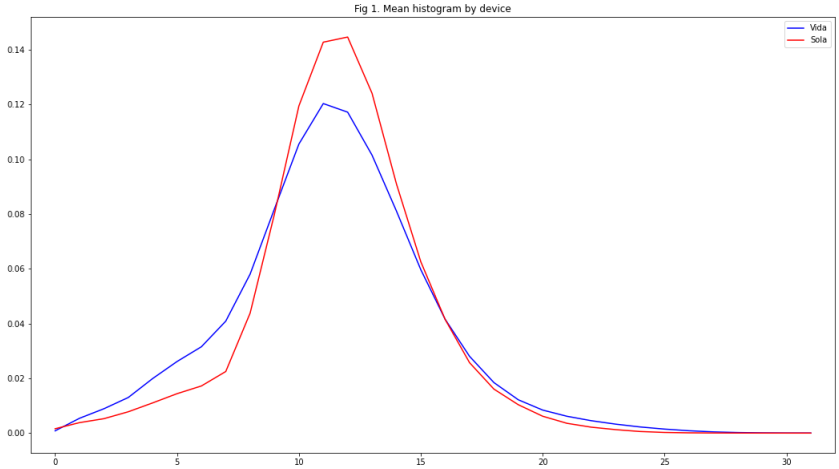

Figure 6: Mean histogram by class

This often implies that the training of a prediction algorithm is done on exams coming

from a single machine, with a low power of generalization to exams from other machines. These different sources of variability, and in particular the last one, impact the robustness of the measured radiomics and make the emergence of operational rapid learning healthcare (RLHC) networks difficult [5].

1.5 Objectives

The goal of this research project is to propose deep learning methods to reduce inter-scanner variability of MRI intensities. This research project is the following of my last internship, where for this purpose, I worked on a first dataset from Oscar Lambret center and trained GAN to learn histogram transport on exams from two different devices, evaluating it with a necessary but not sufficient criteria, and discussing some limitations and new ideas. This approach is quite recent in the literature on this problem, other more basic statistical methods being usually used [7]. All the code has been written in Python and will be freely available on Github [8].

In the continuity of the internship, the first months of the research project were used to go further into the literature review, especially on another very popular method called ComBat, and to elaborate a more proper experimental protocol which made it necessary to obtain another dataset with certain specificities that we looked for over several months online but also by hospital contact. The end of the research project has served to do a precise cleaning of this new dataset and to prepare a toolbox in Python to manipulate it, facilitating the realization of the experimental protocol for future work.

The following sections will present the first then the second dataset at our disposal, justify the chosen pre-processing, clarify the angle of attack used, then discuss the previous results and the new contributions.

2 Approach

2.1 Another classical method: ComBat

When they realized that reshuffling images to the same voxel size altered their quality, and that applying a z-score based on each machine or hospital's data was very limited, the researchers came up with the idea of re-using an algorithm developed in genomics by Johson et al. in 2007 called ComBat [9]. The method realigns all data in a single space in which the batch effect is discarded without altering the biological information.

In MR, the challenge is even more difficult as, unlike in PET and CT where images are expressed in kBq/mL and Hounsfield units, respectively, there is no standard MR intensity grayscale, implying the lack of a tissue-specific absolute intensity numeric meaning, even within the sameMR imaging protocol, for the same body region, for images obtained on the same scanner, and for the same patient. The standardization of image intensities among patients is therefore absolutely needed for comparing values of intensity-based features [10].

The ComBat method directly applies to the radiomic feature values and estimates the scanner effect by matching the statistical distributions of the feature values measured in VOI j for each scanner i:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i \epsilon_{ij}$$

where α is the average value for feature y_{ij}, X is the design matrix for the covariates of interest, β is the vector of regression coefficients corresponding to each covariate, γ_i is an additive scanner effect, and δ_i is a multiplicative protocol effect affected by an error term ϵ_{ij}. The model parameters are estimated using a maximum likelihood approach based on the set of available observations from the two machines. The corrected values are obtained using:

$$y_{ij}^{\text{ComBat}} = \frac{y_{ij} - \hat{\alpha} - X_{ij}\hat{\beta} - \hat{\gamma}_i}{\delta_i} + \hat{\alpha}$$

where $\hat{\alpha}, \hat{\beta}, \hat{\gamma_i}, \hat{\delta_i}$ are estimators of $\alpha, \beta, \gamma_i, \delta_i$, respectively. This approach has already been successfully validated for radiomic features measured from PET (positron emission tomography) and CT (computed tomography) images of patient or phantom data in studies supporting the relevance of harmonization [11].

2.2 Datasets

The first dataset is composed of 213 "Head and Neck" MRI examinations of 70 patients with or without tumors, stored in DICOM format. 180 exams were performed from the Vida machine (3T) and 33 from the Sola machine (1.5T). Sola produces fixed size images (496x512) while Vida generally produces images of size (260×320) but some exceptions make (270×320) or (256×256). An exam usually consists of 30-40 slices, and this variability is a barrier to comparing exams to each other at the same slice height. The minimum value of a voxel is always 0 but the maximum value does not seem to be fixed, it does not exceed 1000 on the Sola but can reach 2500 on the Vida. This dataset is also composed for each machine of 2 examinations of "phantoms", standardized objects allowing to check the good functioning of the machine, and also to measure the inter-machine variations, these objects being very stable in time and inanimate [12]. Phantoms were excluded from the dataset because they could bias the learning of the neural network based on real patient examinations.

All the radiomics studied here are said to be "first-order", i.e. calculated from the histogram of the gray levels of the image. Each slice can be summarized by a histogram, without loss of information for the radiomics concerned. We will present here the different steps of pre-processing to generate the dataset, the algorithmic details being available on Github. The dataset can be summarized as a collection of histograms of variable size (between 1000 and 2500), each associated with a patient ID, an exam number, a slice number and a 1 if the exam was performed on Vida, 0 on Sola. To simplify things and make the histograms comparable, it is possible to apply a standardization (subtract the average histogram and divide by the standard deviation). Then a normalization on a given range (for example to bring all the values between 0 and 32). To keep the original meaning of a histogram it is of course necessary to finish with a discretization.

The standardization allows to remove the noise due to the fluctuations of scale between the examinations, the normalization on 32 values follows the recommendations of the literature, too few values making lose information and too many values adding noise [7]. Once these steps are done, we notice that the background of the image being represented by black and being dominant, all the histograms present a huge peak at its beginning, which makes all the histograms extremely similar and prevents any learning according to us. We therefore add a step at the very beginning of this pre-processing which is to convert these DICOM files into nifti files, to apply a brain extraction algorithm named HD-BET [13]. This one provides us with a mask that will allow us to isolate the voxels that are part of the brain from those symbolizing the cranial box or the background of the image. Thanks to this mask, the proportion of black voxels has been greatly reduced, allowing the shades to appear on the histograms from one slice to another. Nevertheless, this step revealed that many slices contained very little brain (in proportion of the voxels of the mask on the whole voxels of the slice). By representing the cumulative distribution function of the mask proportion in the image, we realized that requiring at least 1% of the slice to be brain voxels cut us off from nearly 20% of the dataset.

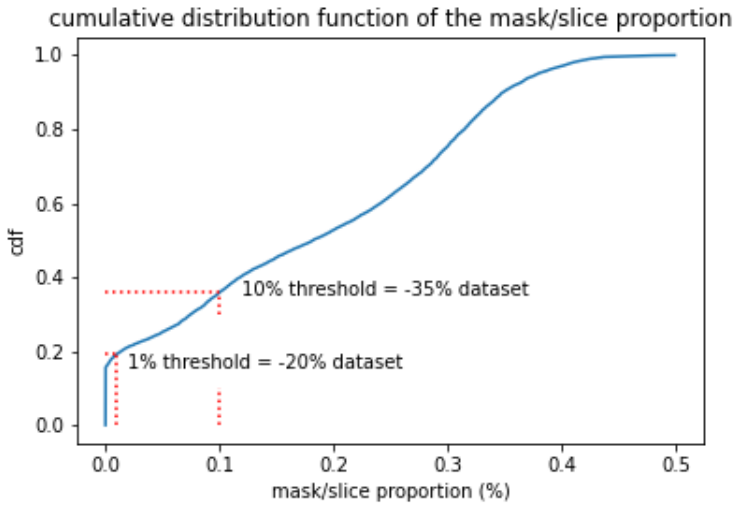

Figure 7: cumulative distribution function of the mask

However, we have applied this filter, to keep only slices of interest. After all these steps, the generated dataset is a table of 35 columns, with 1 row per histogram, of this form:

ID	exam	slice	device	h0	...	h31
291755	2	18	1	0.02	...	0.09

Table 1: Structure of the final dataset

As discussed in the part 3.2 "Evaluation criteria", we conclude at the end of the internship than the moments' convergence is a limited method to measure the performance of our neural network. We then developed an experimental protocol detailed in part 3.3, that need a dataset with different features: The first one is to contain several dozens of MRI exams on two different devices ; The second one is that all the exams have to be in the same modality (T1, T2, FLAIR) and of the same body region ; The third one is that the two groups of patients for each device had similar characteristics (not an unbalanced proportion of elderly or diseased patients) ; The fourth and last one but maybe the hardest to obtain is that these exams have to be associated with a verified medical data (cancer diagnostic, Gleason grades, etc) in order to create a radiomics-based predictor (in regression or classification settings).

This led to a lot of research within the Oscar Lambret center by the radiotherapists that helped me on that project and online on different medical open-source databases. After a long period of mail exchanges to verify and discuss each possibility, we find and agree to choose a database from the Cancer Imaging Archive called "Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy (Prostate-MRI-US-Biopsy)" at the end of January [14]. Because of the size of DICOM files the database weighs 80Go, so I got it by hand from Alexandre Escande that download it from the center and put it on a hard disk.

The database contains 937 MRI prostate exams of 776 patients, distributed on 8 different devices, including 641 exams on a Skyra machine, then 63 on a Vida machine, the rest of exams on the 6 others. All the exams are in a square format, the vast majority in (256×256) and some of them in (512×512). Each exam has between 60 and 80 slices, and has the advantage of being entirely in a tissues' region, avoiding the huge black peak

problem from the first dataset and so the difficult segmentation task to remove it. Each MRI exam is associated to a biopsy that concerns a dozen of 3D parallelepiped zones, defined by the MRI (x,y,z) coordinates of two opposite corners. And each 3D zone is associated to a primary and secondary Gleason grades.

The Gleason grading system is a prognosis tool specific to prostate biopsy, designed to evaluate the severity of a tumor. Depending on the observed pattern, the score goes between 1 for a well-differentiated carcinoma with small, well-formed, and closely packed glands, to 5 for anaplastic carcinoma with a tissue that does not have any or only a few recognizable glands. The Gleason score is decomposed into a primary grade which is the dominant pattern in the tumor, and a secondary grade which is the second most frequent pattern. The Gleason grading system is always expressed as a sum of the primary and secondary grades (e.g. 7 [3+4]). The majority of treatable cancer are of 5-7 Gleason scores, and cancer with 8-10 Gleason scores tend to be advanced neoplasms that are unlikely to be cured.

To better understand this medical grading system, we can visualize micrograph of prostate cancer with different Gleason grades:

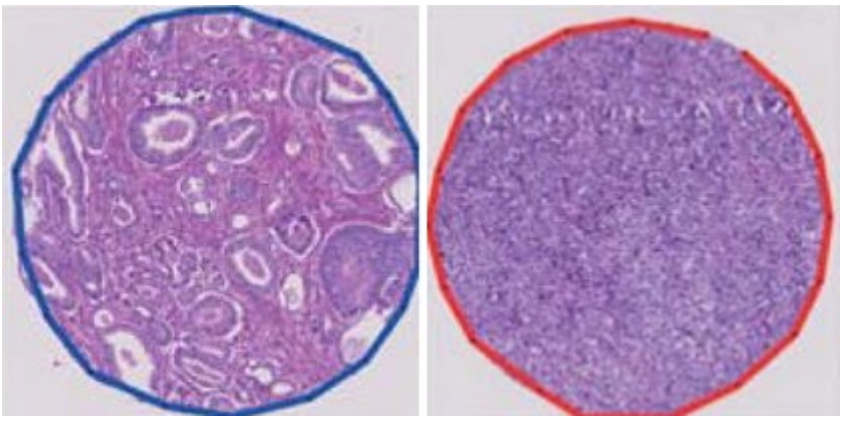

Figure 8: Micrographs of prostate cancer with Gleason 6 (left) and Gleason 10 (right)

2.3 Generative models applied to histogram transport

The objective is therefore to use this table to learn a histogram transport allowing us to simulate what an image generated by one machine (e.g. Vida) would have looked like if it had been produced on the other (e.g. Sola). From a theoretical point of view, we will consider that the histograms extracted from images acquired on the same machine come from the same distribution. Let $\mathcal{P}_0^{\mathcal{X}}$ and $\mathcal{P}_1^{\mathcal{X}}$ denote respectively these underlying distributions for the two classes 0 (Sola) and 1 (Vida) of histograms. In the following they will be respectively referred to as source and target distributions. Both distributions have a common support over the simplex $\mathcal{X}$. Indeed, we recall that the histograms satisfy the non-negativity and sum-to-one constraints:

$$\mathcal{X} = \{h \in \mathbb{R}^N | h[i] \geq 0 \forall i \in [0, N-1], \sum_{i=0}^{N-1} h[i] = 1\}$$

The goal is to learn a function $T : \mathcal{X} \longrightarrow \mathcal{X}$ such that for each histogram h coming from the source distribution $\mathcal{P}_0^{\mathcal{X}}$, $T(h)$ could have been sampled from the target distribution $\mathcal{P}_0^{\mathcal{X}}$, while still remaining similar to the original histogram h.

One way to achieve this goal is to train a generative neural network (GAN). Basically a GAN is a model that estimates the probability distribution of a training dataset and generates new data similar to this distribution. The architecture of a GAN consists of two neural networks: a generator G and a discriminator D trained in an adverse manner. In the classical setting, the generator learns to transform samples from a dense latent space into samples that can fool the discriminator. The latter is actually learning to determine whether a given sample comes from the real data distribution $p(z)$ or its has been created by the generator. The training procedure amounts to optimize the following loss, called adversarial loss:

$$\min_G \max_D \mathcal{L}_{GAN} = \mathbb{E}_{x \sim p(x)}[\ln D(x)] + \mathbb{E}_{z \sim p(z)}[\ln 1 - D(G(z))]$$

The goal of the discriminator is then to maximize this loss by successfully detecting the fake samples from the real ones, while the objective of the generator is the opposite: it tries to fool the discriminator [15]. In summary, the architecture of a GAN looks like this:

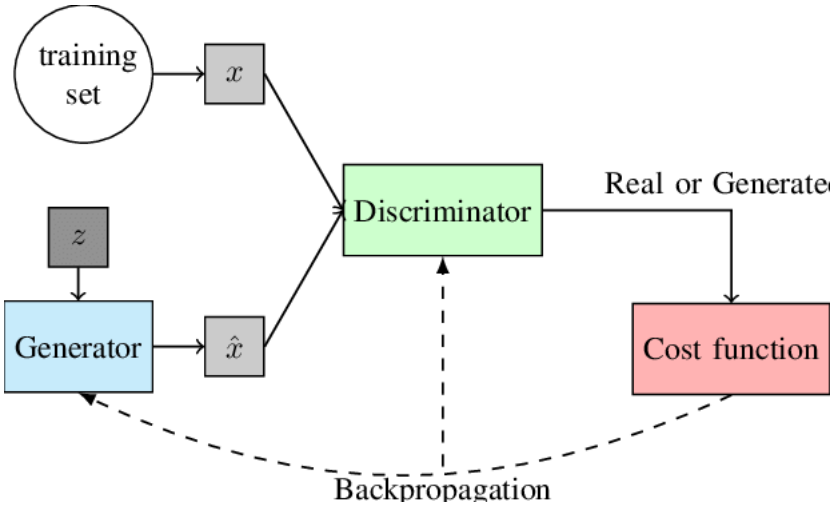

Figure 9: GAN structure [16]

It is thus possible to create a GAN whose generator takes as input histograms of the source distribution and tries to generate their transport in the target distribution, and a discriminator trying to distinguish the real target histograms from the target histograms created by the generator. In the following section, we will detail the creation of the GAN during the course, the criteria used to guide us and the final architecture chosen.

3 GAN architecture design

3.1 The variety of generative adversarial networks

The GANs form a family of architectures, and once this family has been chosen, we still have to determine what form to give to the network. There are many aspects to customize: First of all, we have to choose the number of layers of the generator and the discriminator as well as their types. A few dense layers of a few tens of neurons are capable of serving as both generator and discriminator. But convolution layers can also be used. These two main types can of course alternate, complement each other, each layer having a determinable number of neurons (and thus of connections/parameters).

Another architecture idea can be to condition the GAN (generator AND discriminator) to an additional information: the slice number. Indeed by separating the average histograms by slice (beginning, middle, or end of the examination) we notice that the slice number has a clear influence on the shape of the histogram.

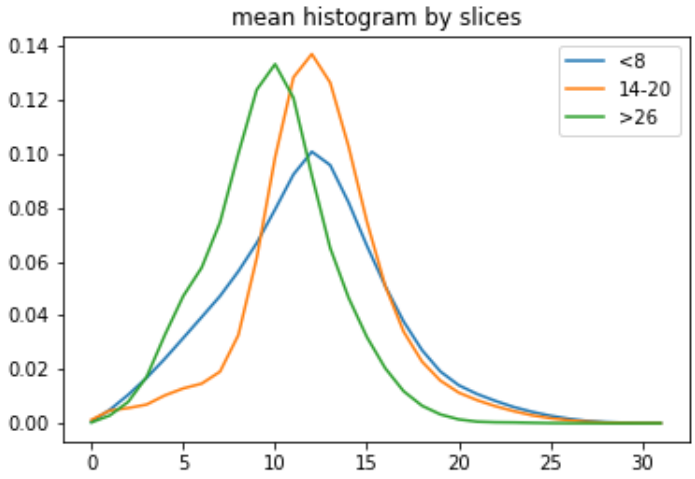

Figure 10: Slice position's influence

A Conditional GAN integrating the information of the slice may therefore be interesting [17]. Another possible architecture would be CycleGAN, which trains two GANs in parallel to learn a T transformation and its inverse T^{-1}. This type of generative model emerged in 2017 and has attracted a lot of interest via impressive graphical demonstrations:

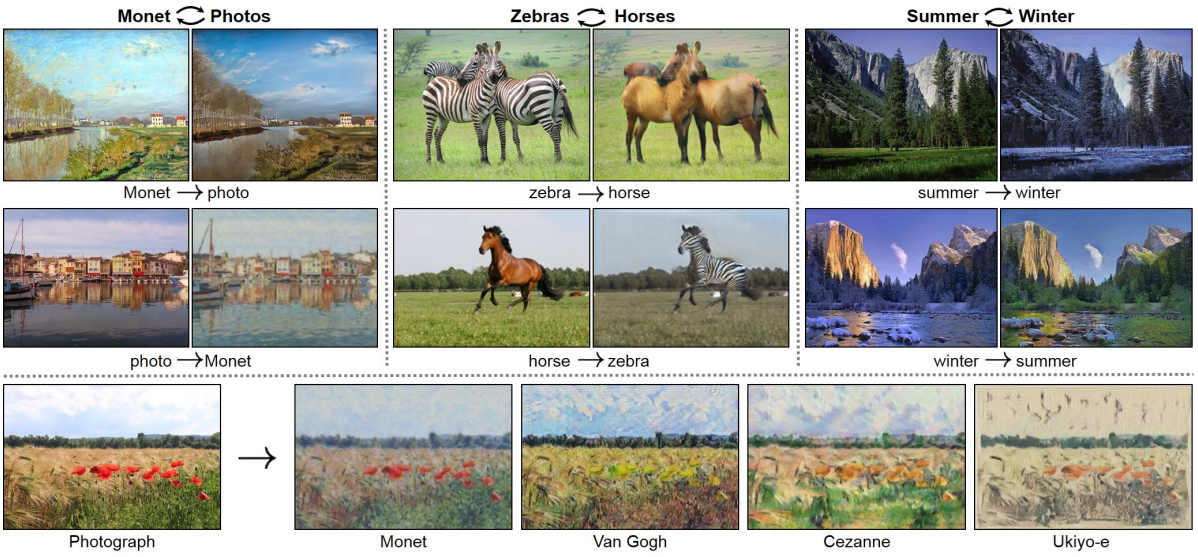

Figure 11: Visual demonstration of cycleGAN capacities [18]

A CycleGAN could in the problematic of this internship learn both the Vida to Sola transformation and at the same time the Sola to Vida transformation. Moreover, the cycleGAN methods are trained on two unpaired datasets and try to learn a transformation that preserves features of the original image, which is not guaranteed by a classical GAN [18].

But among this immense variety of possible architectures, each with innumerable parameters to be trained, how to evaluate their performance? This is a very complex question that came up several times during the previous internship and that leads to a deepening during the research project.

3.2 Evaluation criteria

It seems important to me to describe the different steps in our search for an evaluation criteria during the internship to better understand why the experimental protocol developed in the research project is important and non-trivial. First we tried to visually estimate the quality of the transformation by basing our expectations on the average histograms: if the transformation produced the same kind of shift as the transition from the average source histogram to the average target histogram, then the transformation was of good quality. In other words, we wanted each histogram to transform as an evolution from the Sola histogram to the Vida histogram on this figure:

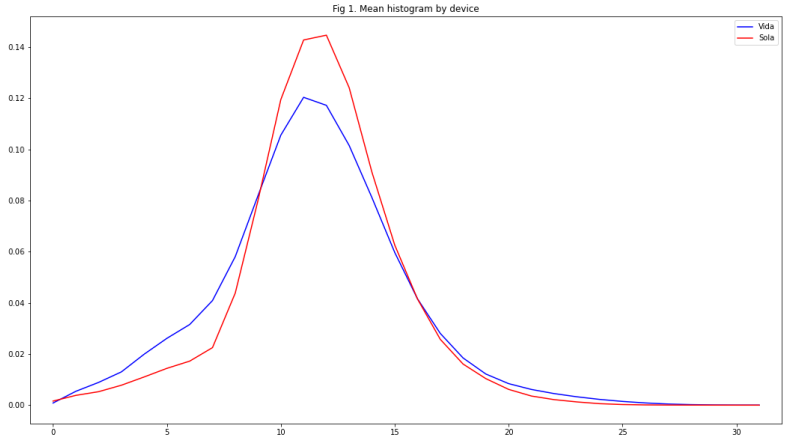

Figure 12: Mean histogram by device

This was a mistake, because in the case of histograms far from the mean histogram, it is necessary to take into account the total shape of the distribution to conclude:

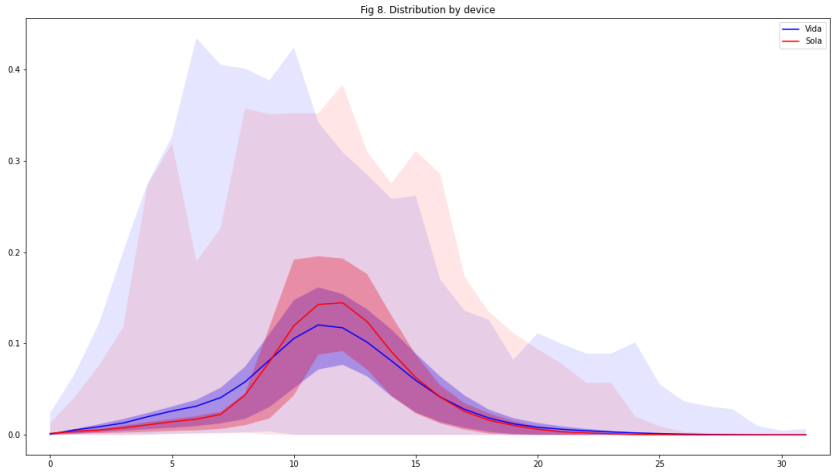

Figure 13: Histograms distributions by device

The main difficulty is that we try to learn a complex movement that cannot be visualized easily. When you try to train a GAN to generate images of puppies, you can check if the generated images seems relevant, realistic and diversified.

We finally decided that visual assessment was too difficult and subjective and that we should look for a quantitative measure. However, we were right on one point: a good transformation must bring the mean histogram of the source class closer to that of the target class. A first measure can therefore be the evolution of the distance (chi-square for example) before/after transformation, between the average histograms of each class. This is indeed a necessary but not sufficient condition, the mean could converge without preserving the variance. More precisely, it is true to say that the transformation we are looking for converges all the moments of the source distribution to the moments of the target distribution. The more moments we consider the more precise the control will be.

However, this has two limitations: this criterion is not sensitive to individual transformations. To better understand, consider each histogram of the dataset as a point in dimension 32. To each class is thus associated a point cloud, and we check that the

mapping from the source point cloud to the target point cloud area is satisfactory. The verification by moments ensures that the centers of the point clouds are close, that their dispersion is similar, etc. These are verifications on the point clouds. But if for a given mapping, we exchanged each pair of source point-transformation by another one at random, the convergence of the moments would be the same and yet the transformation would be totally changed. The second limitation being that very quickly, the verification of the convergence requires sample sizes that are far too large. Indeed, the 32x32 covariance matrix requires estimating more than 1000 parameters, and the skewness tensor more than 30 000. With a sample of a few hundred histograms it is not reasonable to try to estimate them.

We finally opted at the end of the internship for a simple and relatively robust compromise: to measure the relative evolution of the distances (chi-square) between the mean histograms and the variances per component (the diagonal of the covariance matrix). This allows us to obtain two scores, which are negative if the transformation makes the moments diverge globally, null if the transformation does not change the global distance and positive if the distance decreases, up to 100% if the distance is made null.

Once this criterion is defined, it is also necessary to separate the dataset according to a classical train-test division to avoid overfitting on the training data. We thus obtain 4 measures: $train_m$, $train_v$, $test_m$, $test_v$, all defined like this on the **s**ources, **t**argets et **g**enerated datasets:

$$score = \frac{\chi^2(h_s, h_t) - \chi^2(h_g, h_t)}{\chi^2(h_s, h_t)}$$

Using these criteria to compare the different architectures, we arrived at rather satisfactory results. The composition of the final architecture, available on the Github, is not decisive with regard to the duration of the previous internship and the performances obtained (see the following section), and even less important regarding the current research project. The important elements are mainly the proposal of a resolution approach via the GAN, and of a first evaluation method via the moment convergences. But this can be upgraded, and this is the road that the research project has begun to pave.

This is why at the beginning of the research project, I spent time trying to detail an experimental protocol that overcomes the limitations of the moments' convergence, that will be detailed in the following section.

3.3 Experimental protocol

The goal of this experimental protocol is to propose a clear evaluation method that can be easily adapt to any dataset that corresponds to our four conditions described in part 2.2 (like our second dataset), to test the performance of any MRI harmonization method (like ComBat, all the GAN family of model with its Conditional GAN and CycleGAN variants, or other future algorithms).

This protocol can be detailed in 8 principal steps:

1. Dataset analysis: As in any experiment, critical thinking about the input material is crucial and with this in mind, one should verify that our 4 conditions applies (sufficient size, same modality, same body region, similar medical features between groups, possibility to create a supervised predictor) without complication.

2. Pre-processing: As described in the first dataset manipulations, the dataset should be pre-process by a standardization, a normalization and a discretization to put each voxel between 0 and 32 as in literature recommandation, then extract the grey-level histogram of voxel for each 3D exam. If the image contain a lot of non-tissue region as air that creates a huge peak for black voxel, segmentation should be applied before the histogram extraction.

3. Subdatasets creation: To distinguish different results and measure the overfitting risk, one should split the dataset into 4 subgroups : $Skyra_{train}, Skyra_{test}, Vida_{train}, Vida_{test}$

4. Algorithm training: Train the GAN to learn the histogram transport from Vida to Skyra, for that give $Skyra(Skyra_{train} + Skyra_{test})$ to the generator and $Vida_{train}$ to the discriminator.

5. Histogram generation: Feed the generator of the trained GAN with $Vida_{train}, Vida_{test}$ to generate two new subgroups that we choose to call $Vida^G_{train}, Vida^G_{test}$.

6. Radiomics creation: Extract grey-level histograms from each biopsy zone to generate the list of 1D radiomics (mean, variance, etc).

7. Radiomics-based predictor: Train your supervised predictor, for example a Support Vector Classifier for binary classification of Gleason grades over 5, on radiomics associated with $Skyra_{train}$, then evaluate it on 6 conditions to answer the following questions:

    - $Skyra_{train}$: Does the classifier obtain a good score on its training data?
    - $Skyra_{test}$: Does the classifier obtain a good score on new data from the same device?
    - $Vida_{train} + Vida_{test}$: Does the classifier obtain a good score on new data from another device without harmonization?
    - $Vida^{G}_{train}$: Does the classifier obtain a good score on new but harmonized data, coming from the GAN training set?
    - $Vida^{G}_{test}$: Does the classifier obtain a good score on new but harmonized data, unseen by the GAN?
    - $Vida_{combat}$: Does the classifier obtain a good score on new but harmonized data via the ComBat method (or any other method used for comparison)?

8. Conclusion: We will deduce that the GAN improves the situation if the classifier obtains a better score on $Vida^{G}_{test}$ that on $Vida_{train} + Vida_{test}$, and we will have the possibility to compare it with $Vida_{combat}$. It is easy to also check the moment convergence of radiomics as a secondary criterion.

Mainly due to lack of time and difficulties concerning the biopsies 3D zones, the full experimental protocol couldn't be done on our recent dataset, but the description of a method and the creation of associated toolbox in Python will certainly facilitate future work in the continuity of this research project.

In the following part we will present the obtained results and will take the necessary critical distance to discuss them.

4 Results and discussion

The evaluation of a neural network is a stochastic process on at least two levels: First, the initialization of the network weights is a random process that can be determined by a first random seed. Second, the separation of the dataset into training and test data is also random and determined by a second random seed (which can of course be the same).

Ideally, we are looking for an architecture that performs better than the others on all the seeds in one case as in the other. For practical reasons, only about ten seeds have been evaluated. The GAN architecture trained during the previous internship often succeeds in reaching scores of 60% improvement for the mean and 20% improvement for the variance, both in train and in test. According to these criteria, we have found a GAN to be an interesting candidate for our problem of standardizing exams from various machines.

However, let us take a moment to present the limits of these results and thus distinguish the aspects that could be improved in the pursuit of this problem. A first remark is to say that the number of architectures to be explored being almost infinite, the best solution found will strongly depend on the time spent exploring, on the methods used and on the available machine capacity. By searching for a few weeks, by trial and error on my personal computer, it is clear that the architecture search can be improved.

A second point to be explored is that of GAN architecture sub-genres: as mentioned above, creating a conditional GAN on the slice number can surely improve the relevance of the algorithm given the importance of this data. It was discussed the interest of conditioning the GAN on other information at our disposal such as patient characteristics (age, sex), this could also be pushed by additional studies. The cycleGAN, also presented above, could help the stability of the individual transformations, bringing more architectural guarantees that a single transformation keeps its properties (because necessarily reversible), and also applicable to more than two different machines.

Regarding the experimental protocol, it is in my opinion a well better way to evaluate performance, but it is not perfect. It is dependent of the type of task that we try to learn on radiomics but also of the type of model we choose to do it. Ideally one should choose

a task that can be learned from radiomics by a sophisticated model (so with a good score on new data from the same device). And it is of course necessary to replicate this work, in particular to test the same method on different MRI datasets, for different medical tasks based on different radiomics.

To conclude on a more personal level, I am very happy that I was able to pursue my internship as a research project. Firstly because the supervisors are a very important aspect of a project and I would like to thank John Klein, Jérémie Boulanger, David Pasquier and Alexandre Escande for their full support. After this research project I am fully convinced that I want to work in data science and machine learning applied to medical research, because of the multidisciplinary issues with technical and human aspects, and also because of the alignment of my ethical principles with the emergence of clinical decision support systems. This research project contains its part of frustration because of the time needed to find a correct dataset (and I would like to thank again Alexandre for his great help), and because I would have really liked to conduct all the experiments to present the results myself. I keep that as a lesson about the difficulty to obtain precise medical data and to clean them, and I hope that this literature review, Python toolbox and cleaned dataset will help for future work, by me in a thesis or by another student research project.

In direct link with this project, I found a six months internship in SOPHiA Genetics Radiomics research team near Bordeaux, where I'll continue to do research in machine learning for radiomics. After that I would really like to do a thesis in the field, maybe in multicenters harmonization techniques using deep learning.

References

[1] Cristal website. https://www.cristal.univ-lille.fr/. Accessed: 2022-03-06.

[2] Oscar lambret website. https://www.centreoscarlambret.fr/. Accessed: 2022-03-06.

[3] Mri information. https://www.nhs.uk/conditions/mri-scan/. Accessed: 2022-03-06.

[4] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to Radiomics. *Journal of Nuclear Medicine*, 61(4):488, April 2020.

[5] Philippe Lambin, Ralph T.H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E.C. de Jong, Janita van Timmeren, Sebastian Sanduleanu, Ruben T.H.M. Larue, Aniek J.G. Even, Arthur Jochems, Yvonka van Wijk, Henry Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12):749–762, December 2017.

[6] Stefania Rizzo, Francesca Botta, Sara Raimondi, Daniela Origgi, Cristiana Fanciullo, Alessio Giuseppe Morganti, and Massimo Bellomi. Radiomics: The facts and the challenges of image analysis. *European Radiology Experimental*, 2(1):36, December 2018.

[7] Alexandre Carré, Guillaume Klausner, Myriam Edjlali, Marvin Lerousseau, Jade Briend-Diop, Roger Sun, Samy Ammari, Sylvain Reuzé, Emilie Alvarez Andres, Théo Estienne, Stéphane Niyoteka, Enzo Battistella, Maria Vakalopoulou, Frédéric Dhermain, Nikos Paragios, Eric Deutsch, Catherine Oppenheim, Johan Pallud, and Charlotte Robert. Standardization of brain MR images across machines and protocols: Bridging the gap for MRI-based radiomics. *Scientific Reports*, 10(1):12340, July 2020.

[8] Github project with all the python code. https://github.com/Sacha-Dedeken/M2_RP. Accessed: 2022-03-06.

[9] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 04 2006.

[10] Fanny Orlhac, Augustin Lecler, Julien Savatovski, Jessica Goya-Outi, Christophe Nioche, Frédérique Charbonneau, Nicholas Ayache, Frédérique Frouin, Loïc Duron, Irène Buvat, and et al. How can we combat multicenter variability in mr radiomics? validation of a correction procedure. *European Radiology*, 31(4):2272–2280, 2020.

[11] 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, and et al. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170, 2017.

[12] Phantoms for mri. https://www.nist.gov/topics/physics/what-are-imaging-phantoms. Accessed: 2022-03-06.

[13] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H. Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019.

[14] Link to the second dataset of 80go with complete scans and biopsy. https://wiki.cancerimagingarchive.net/plugins/servlet/mobilecontent/view/1966258. Accessed: 2022-03-06.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[16] Everything you wanted to know about deep learning for computer vision but were afraid to ask - figure 15. https://www.researchgate.net/figure/Generative-Adversarial-Networks-Framework$_fig3_322413149$. *Accessed* : $2022 - 03 - 06$.

[17] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]*, November 2014.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.