A Weakly Supervised Gradient Attribution Constraint for Interpretable Classification and Anomaly Detection

Valentine Wargnier-Dauchelle, Thomas Grenier, Françoise Durand-Dubief, François Cotton, Michaël Sdika

A Weakly Supervised Gradient Attribution Constraint for Interpretable Classification and Anomaly Detection

Valentine Wargnier-Dauchelle, Thomas Grenier, Françoise Durand-Dubief, François Cotton and Michaël Sdika

***Abstract*— The lack of interpretability of deep learning reduces understanding of what happens when a network does not work as expected and hinders its use in critical fields like medicine, which require transparency of decisions. For example, a healthy vs pathological classification model should rely on radiological signs and not on some training dataset biases. Several post-hoc models have been proposed to explain the decision of a trained network. However, they are very seldom used to enforce interpretability during training and none in accordance with the classification. In this paper, we propose a new weakly supervised method for both interpretable healthy vs pathological classification and anomaly detection. A new loss function is added to a standard classification model to constrain each voxel of healthy images to drive the network decision towards the healthy class according to gradient-based attributions. This constraint reveals pathological structures for patient images, allowing their unsupervised segmentation. Moreover, we advocate both theoretically and experimentally, that constrained training with the simple Gradient attribution is similar to constraints with the heavier Expected Gradient, consequently reducing the computational cost. We also propose a combination of attributions during the constrained training making the model robust to the attribution choice at inference. Our proposition was evaluated on two brain pathologies: tumors and multiple sclerosis. This new constraint provides a more relevant classification, with a more pathology-driven decision. For anomaly detection, the proposed method outperforms state-of-the-art especially on difficult multiple sclerosis lesions segmentation task with a 15 points Dice improvement.**

First submission on 30th August 2022. Resubmitted on 24th February 2023 and on 12th April 2023. This work was supported by the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR). We acknowledge the "Observatoire Français de la Sclérose en plaques" (OFSEP) for providing the data collected with ANR-10-COHO-002. This work was performed using HPC resources from GENCI-IDRIS (AD011012544/AD011012589). Finally, this work was partly funded by SIMAF: "Projet Emergence", CNRS-INS2I.

The authors are with Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69100, Lyon, France (e-mails: valentine.wargnier@creatis.insa-lyon.fr; thomas.grenier@creatis.insa-lyon.fr; francoise.durand-dubief@chu-lyon.fr; francois.cotton@chu-lyon.fr; michael.sdika@creatis.insa-lyon.fr)

F. Durand-Dubief is with Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France

F. Cotton is with Service de Radiologie, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, Pierre-Bénite, France

***Index Terms*— Anomaly detection, Attribution maps, Classification, Constrained learning, Interpretability**

I. Introduction

Deep learning methods have proven their effectiveness in medical image analysis through segmentation, detection, classification or registration tasks [1]–[4]. These methods are applied to a large range of medical imaging techniques such as MRI (Magnetic Resonance Imaging), CT (Computed Tomography), radiography, ultrasound, fundus images, etc. For example, it can be used for brain tumor segmentation on MRI [5], for image registration on chest CT [6] or to class healthy vs COVID-19 lungs X-ray images [7].

More specifically, classifiers are essential building blocks for various deep learning frameworks. They can be used for classification or characterization of samples for computer-aided diagnosis tasks [8]. They can also be used to detect error and measure classification uncertainties [9] or to detect outliers in order to clean a dataset before expert analysis or before training [10]. In addition, they are also essential building blocks of generative adversarial networks (GAN) [11] in which they model a metric learned from the dataset that will drive a generator network during the training. In this case, it might be crucial that the discriminator bases its decision on meaningful and relevant features.

Despite the success of these methods in medical imaging, the "black box" nature of deep learning restrains their diffusion in clinics for diagnosis or characterization as practitioners need to have confidence in the proposed automatic decision. Indeed, neural networks decision is difficult to interpret because of their large number of parameters and non-linearity. Moreover, due to their high capacity to extract features, their decision is not always based on relevant radiological signs as human experts use. Besides, most medical imaging techniques are not quantitative. For example, MRI acquisitions have intensity variations due to field strength and inhomogeneity, acquisition protocol, scanner brand, artifacts, etc. Thus, the network decision can be based on the acquisition signatures of the different datasets and not only on the pathology. Several methods have been proposed in order to standardize MRI and try to remove this signature hoping the network decision to be based on radiological signs [12]–[15]. This normalization

is not always sufficient for the network to focus on relevant features and it may find shortcuts in making its decision. In this case, performance metrics such as accuracy can be high but if we scrutinize at what the decision is made on, for example with attribution methods, the explanation might not match domain expert knowledge [16]. Especially in a critical domain like medicine, we expect a deep model to be accurate, interpretable and decision to be consistent with high-level clinical knowledge.

Attribution maps computation is state-of-the-art technique to explain deep networks decision [17]–[19]. From a trained network, these methods compute a heat map that indicates the positive or negative contribution of each voxel of the input image in the network decision. These methods are mostly used at inference to verify the interpretability of a trained network and check that these maps match high-level knowledge as in [20], [21]. For example, in a medical context, they can be used to check that the decision of the network matches anatomical abnormalities present in the image. Very few use them during the training to improve the interpretability of a classification network: in [19], [22], the classification task is trained jointly with a regularization loss on these maps to make them cleaner. These kind of regularization has also been used for anomaly detection on architectures such as auto-encoder (AE) in [23]. Constraining GradCam [18] attributions with respect to the latent space, they surpass literature methods for anomaly detection [24]. In these cases, the assumption is that models trained to reconstruct healthy images will not be able to reconstruct anomalies and thus, the reconstruction difference can be used as an anomaly segmentation (AE [25], Variational AE (VAE) [26], Generative Adversarial Network (GAN) [27], etc).

In this work, we propose to constrain a binary classification network between healthy and pathological subject images such that its decision (healthy or pathological image) is based on relevant radiological structures in an unsupervised way. To do so, we use attribution maps as network decision indicator and constrain the network attributions for normal images to be entirely relevant for the healthy classification. During inference, the trained network can be used for classification based on radiological signs but also as weakly supervised pathology segmentation network.

The main contributions of this work are the following: 1/ We propose an unsupervised method to constrain the attributions of a deep classifier to be negative outside the pathological areas (and consequently positive inside) through a new loss. 2/ Only the image label is necessary to reach good performances in terms of classification and segmentation, outperforming anomaly detection literature methods. 3/ Attribution constraints are integrated in the training such that the resulting network is invariant to the choice of the (gradient based) attribution method used at inference. 4/ We show that the constraint on Gradient attributions is, in most cases, equivalent to Expected Gradient with a easier and faster training. 5/ Our method was evaluated with numerical experiments on two classification and segmentation tasks: binary classification between healthy subjects and either brain tumor patients or challenging multiple sclerosis patients.

II. Related work

A. Interpretable classification

Most attribution methods [17], [28] have been proposed as posthoc procedure to visualize which pixels contribute positively or negatively to the network decision [20], [21]. In few works, they have also been used to regularize the training of a classification network. For example, in [19], it is assumed that neighboring pixels of the input image should have a similar impact on the decision. This constraint is implemented by adding a total variation loss on the Expected Gradients attributions. This loss effectively regularizes the attribution maps that, as a result, become cleaner and smoother. However, this loss is not related to the healthy vs pathology classification task i.e. not related to the existence of visible and characteristic pathological structures. In [22], the idea is to make the gradients with respect to the input small in non-interesting areas. Assuming a mask of these area is available, the L_2 norm of the gradient is penalized in this mask. In the context of a healthy vs pathological classification, the constraint proposed by [22] could be adapted to force the neural network to be insensitive to "healthy region". Nonetheless, voxels in healthy regions should not be neutral, they should drive the decision towards a healthy classification.

B. Anomaly detection

Anomaly detection state-of-the-art methods are often based on a reconstruction task: the network is trained to reconstruct healthy images and anomalies are segmented at inference by thresholding the reconstruction error for pathological subjects. AE [25] and VAE [26] are typically used for this kind of approach. They rely on the fact that pathological areas are out of the training distribution and should not be correctly reconstructed. Nevertheless, for (V)AE, the reconstructed image is often blurry making the reconstruction error map large around image edges and the anomalies difficult to threshold. The distribution of healthy images can also be learned through GAN architectures. In [27], the encoder-decoder architecture is trained for both image-to-image and latent space-to-latent space reconstruction. An adversarial loss is added such that generated images could not be differentiated from real images by the discriminator. However, it is well known that GANs are difficult to train and prone to mode collapse. Finally, in [23], attribution regularization is added to a reconstruction-based method: the GradCam [18] map of the first convolution block with respect to the latent space of an AE is maximized over the whole image. The anomaly detection is done by thresholding this GradCam map. Nevertheless, as described in [18], GradCam lacks semantics on the first layers: [23] ends up being a simple thresholding after a well-chosen convolution on the input image.

Compared to anomaly detection methods, which are only trained on healthy images, interpretable classification methods described in Section II-A need pathological images and weakly supervision through their label (healthy/pathological). Nevertheless, pathological databases are often available and it allows to add pathological information in network features. More precisely, attributions of classification methods are based

on healthy vs pathological differentiation whereas in [23], attributions are based on a reconstruction task, with no information on the considered pathology. In addition, interpretable classification methods can be used for two tasks: segmentation and classification.

III. Methodology

A. Classification with attribution constrained training

In this work, we propose to train a deep network to classify healthy vs pathological subject images with the additional constraint that its decision satisfies high-level clinical properties. We assume that the decision of the network for a given input image is reflected by an attribution map with the same size as the input image (see details in Section III-C.1). During the training, the network learns to correctly classify healthy and pathological subjects but also to satisfy constraints on the attribution map produced by the network (see Figure 1). It is done by minimizing the following loss:

$$L = L_C + \alpha_A L_A \tag{1}$$

where L_C is the classification loss, L_A is the attribution loss aimed at penalizing unsatisfied constraints on attribution maps with a penalization coefficient α_A. L_C is a standard classification loss: in our implementation, it is the binary cross entropy. Details on L_A are given below.

The trained network can be used as an **interpretable and relevant classifier** whose decision is based on clinical features but also as a **pathology segmentation network** using inference attribution maps as pathology segmentation mask.

B. Unsupervised attribution constraint with binary cross entropy

Attribution maps aim at revealing regions in the input image that contribute positively or negatively to the network decision. For healthy subject images, no region should drive the network to pathological classification but each region should drive the decision towards healthy class. Thus, attribution maps of healthy subjects should be negative over the whole image. We propose to use the binary cross entropy as an effective way to impose this constraint. With this loss, attribution values are seen as logits of a virtual pixel-wise classifier. Voxels from healthy subject images should be classified as healthy, i.e. with negative logit. No additional input is necessary to use this constraint which is consequently unsupervised. The only annotation needed is the image-level label (healthy/pathological) already required for the classification. Consequently, our L_A loss can be used on all healthy subject images and can be written as:

$$L_A(x) = BCE\big(\sigma(A(x)), 0_{Size(A(x))}\big) \tag{2}$$

where σ is the sigmoid function, BCE is the binary cross entropy loss, x is the input image, $A(x)$ is the attribution map for the image x (same size as x) and $0_{Size(A(x))}$ is a all-zero image with same size as $A(x)$.

Although our constraint is designed to be used in weakly supervised learning, it can easily be extended to semi-supervised training. In this case, it is assumed that a pathology map $m(x)$, indicating where positive attributions should be, is available for some of the patient images x. This pathology map can just be a segmentation mask of the region of interest: brain tumors or multiple sclerosis lesions for example. In this case, the L_A loss becomes:

$$L_A(x) = BCE(\sigma(A(x)), m(x)). \tag{3}$$

Note that training a deep classifier with the additional L_A loss is strictly equivalent to the multi-task training of a very constrained encoder-decoder that would segment the pathology maps and whose latent space would be the classification output. The encoder branch of this encoder-decoder would be the classifier (forward pass, unbroken arrow in Figure 1). The decoder branch, parameterized with the encoder weights, would be the attribution computation (dotted arrow in Figure 1).

C. Training with gradient attribution constraints

1) Gradient based attributions: Intuitively, the gradient of the network output with respect to the input image indicates which pixels are the most critical in the network decision. If the derivative with respect to a pixel is the largest, changing its value will change the network output the most. Gradient-based attribution methods [29] are known to have interesting properties regarding interpretability but also to be integrated smoothly in the training procedure. Three gradient attribution methods are considered in this work. Note that we used local attributions [29], i.e. without multiplication of gradient by the input, but development with global attributions would be similar.

a) Gradient (G): Output gradient with respect to the input, or saliency, is the simplest way to evaluate voxels relevance for output [28]. It is defined as:

$$A_i^G(x) = \frac{\partial F(x)}{\partial x_i} \tag{4}$$

where x is the input, F is the network and i is the index (voxel) in the input image.

b) Integrated Gradients (IG): Integrated Gradients [17] sharp values are defined as:

$$A_i^{IG}(x, x') = \int_{\alpha=0}^{1} A_i^G\left(x' + \alpha(x - x')\right) d\alpha \tag{5}$$

where x is the input and x' is the baseline. The baseline x' corresponds to a null attribution input, often chosen as a null image (i.e. all-zero image) with the same size as image x. The integral parameter α allows to aggregate gradients for images along a path between the image and the baseline. It has been proved that IG have several guaranteed properties: sensitivity, implementation invariance, completeness, linearity and symmetry preservation (see [17]). In comparison, G does not respect the sensitivity axiom.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3282789

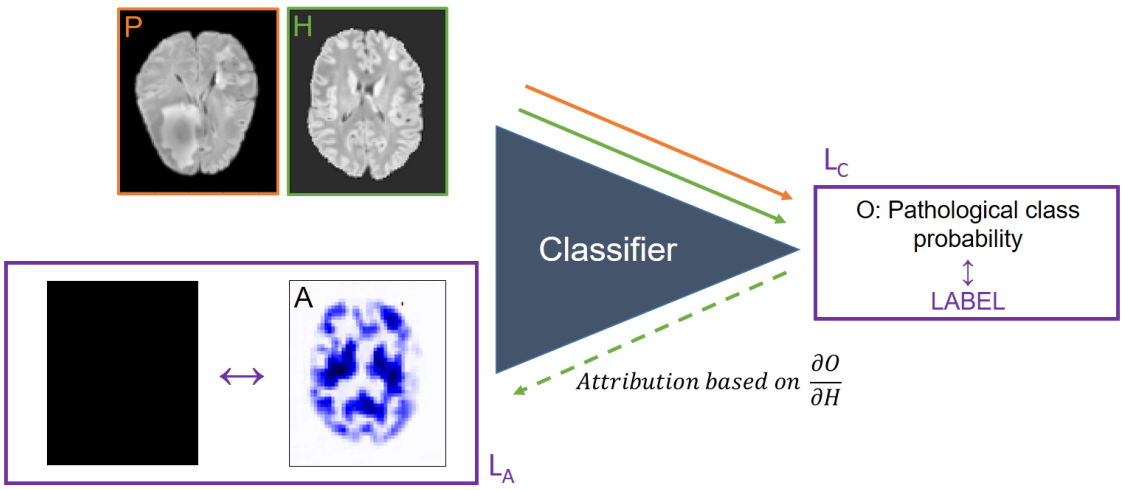


Fig. 1: Method overview for unsupervised training. The classification network is trained with classification loss L_C on both pathological (P, in orange) and healthy (H, in green) images. The loss L_A, which constrains gradient-based attributions A, is only applied on healthy images. During inference, the segmentation of pathological structures is obtained by thresholding these attributions.

c) Expected Gradients (EG): The output of Integrated Gradients is highly dependent on the baseline and the choice of a null image is debatable [30]. In Expected Gradients [19], this problem is solved by marginalizing Integrated Gradients over the baselines in training dataset distribution:

$$A_i^{EG}(x) = \int_{x'} A_i^{IG}(x, x')\ dx'. \tag{6}$$

It was also proposed in [19] to constrain EG attribution maps during the training. As this would be untractable, a stochastic version was proposed in which the double integration on α and x' in Equation 6 is replaced by the sampling of a pair (α, x') at each iteration. To further reduce the computation, the other images of the x mini-batch could be used as the baseline x'.

2) Easier training: are constraints on G sufficient for EG ?: The constrained training on EG maps proposed in [19] improves the interpretability of deep neural network at the expense of a substantial amount of computational burden. Indeed, training with EG requires at least a second *double backpropagation pass* on a different minibatch from the one used for the classification loss, built as $x' + \alpha(x - x')$ with x and x' samples from the training dataset and $\alpha \in [0, 1]$. In comparison, training with G only requires a double backpropagation *on the same minibatch* used for the classification loss, efficiently re-using the derivative computation for the gradient descent. Consequently, G needs less code, computation time and GPU memory than EG.

In this section, we defend the following conjecture:

Conjecture 3.1: Two models trained with constraints on either G or EG produce equivalent G or EG maps at inference.

This conjecture implies that constraining the training with EG instead of G is unnecessary.

Although we cannot provide a completely formal proof of this conjecture, we propose: a global sketch of this "proof", new theoretical elements to formally prove some points of this sketch and a discussion based on state-of-the-art to support the missing points. This claim is also backed by experiments (see section V-B.1). To explain our intuition, we need to define the following notions.

Definition 3.2: If X is a subset of a vector space, we define the Line Segment set of X as:

$$LS(X) = \left\{\alpha x + (1-\alpha)x' \mid (x, x', \alpha) \in X^2 \times [0, 1]\right\}.$$

In words, $LS(X)$ is the set of points on line segments whose ends are in X, i.e. the images obtained by linear interpolation between two images of X. One can note that:

Remark 3.3: If $C(X)$ is the convex hull of X then $LS(X) \subset C(X)$.

Using this definition, the following proposition makes the link between constrained training with either G or EG.

Proposition 3.4: The stochastic constrained training with EG on X proposed in [19] is equivalent to a stochastic constrained training with G on $LS(X)$.

Proof: Formally, the genuine constraint on EG attribution maps is the addition of the following term to the classification loss:

$$\sum_{x \in X} L\left(\sum_{x' \in X, \alpha \in [0,1]} \nabla F\Big(\alpha x + (1-\alpha)x'\Big)\right) \tag{7}$$

where F is the network, x is the input image, x' is the baseline, X is the training dataset and L is a loss used to constrain the EG map of a single image. In the stochastic version proposed in [19], at each iteration, both sums are removed and the single term

$$L\left(\nabla F\Big(\alpha x + (1-\alpha)x'\Big)\right) \tag{8}$$

is added to the classification loss for a minibatch of x, x' and α. One can notice that this would be the same term that would be used if the overall constraint loss would be:

$$\sum_{(x, x', \alpha) \in X^2 \times [0,1]} L\Big(\nabla F\big(\alpha x + (1-\alpha)x'\big)\Big) \tag{9}$$

$$= \sum_{y \in LS(X)} L\Big(\nabla F(y)\Big) \tag{10}$$

As can be seen, training with the stochastic EG constraint on X is equivalent to the use of the same constraint using G attribution maps on the Line Segments set of X ($LS(X)$ introduced in Definition 3.2). ■

Using the Proposition 3.4, Conjecture 3.1 could be proven if we can show that constrained training on either X or $LS(X)$ have the same effect when the model is applied on test data. Although we do not provide formal proof of this part, it is reasonable to assume that this is the case. We base our intuition on the fact that in high dimension d, each point from a set of N points will be outside the convex hull of the other points with a probability close to 1 (unless N grow exponentially with d). This was proven for points on a hypersphere in [31]. It has also been heuristically validated for real data in [32]–[34]: for datasets such as MNIST, CIFAR10 or ImageNet, images from the test set are outside the convex hull of the training set. This led [32] to the conclusion that "the behavior of a model within a training set's convex hull barely impacts that model's generalization performance".

Regarding attribution constraints, our intuition is that constraining the gradient on the training set X only, and not like EG on $LS(X)$ (which is included in the convex hull $C(X)$), is sufficient to generate good EG maps at inference. This can also be understood as accounting that an image in the test set is very unlikely to be the result of a linear interpolation of two images of the training set. Using only G during the training results in a simpler implementation, a more direct path in the backward and a lower training time. As the constraint does not need to be evaluated at interpolation points, a second backward is not necessary as with EG.

3) Generic training: including all attribution maps in the training (IEG): In the previous section, we argued that the G constraint should be sufficient for good EG inference. Although IG is also based on gradient, training with G only might not be as sufficient. Indeed, the baseline is always the same and outside the convex hull of the training set. If the objective is not to be faster during training but to implement a constrained training generic for the attribution maps used during inference, we propose to use EG with a probability p to use a null baseline (i.e. to use IG) during training. Noticing the beginning of the integration path ($\alpha \approx 0$) for both IG and EG corresponds to G, training with constraints on IG and EG will also constrain G attributions. This setting should improve the invariance of the neural network to the choice of the attribution method used during inference.

IV. Experiments

Our method was evaluated on two brain pathologies: brain tumors and multiple sclerosis. In Sections V-A and V-B, the influence of different parameters of our method, and more specifically the attribution method used for the constraints, is evaluated. A comparison is made between • our unsupervised model (Unsup) based on Equation 2, • a model (Sup) where the constraint would be totally supervised ($m(x)$ in Equation 3 is the segmentation mask when available), • a model (UnsupTV) constrained with both our unsupervised constraint of Equation 2 and the total variation loss from [19], • [22] loss extrapolated in an unsupervised way on healthy images as in our proposition (still denoted as Ross) and finally, • [19] (denoted as Erion). For each method, G, EG, IG or IEG have been used for training and either G, EG or IG have been used at inference. In Section V-C, our method is compared to two state-of-the-art interpretable classification methods: Erion and Ross. The classification network trained without constraint and evaluated with GradCam [18] or Gradient attributions (respectively named NoConsGC and NoConsG) are also compared. For NoConsGC, unlike the original paper, no ReLU was applied on attributions to display both negative and positive relevance. In Section V-D, we compare our approach to state-of-the-art methods for anomaly detection: [25] (AE), [26] (VAE), [27] (f-AnoGAN) and [23] (Silva-Rodríguez).

A. Data

TABLE I: FLAIR MRI datasets. H refers to healthy dataset, T to tumors dataset, MS to multiple sclerosis dataset.

Dataset	N_{train}	N_{val}	N_{test}	H/T/MS	Annotated
MPI	64	15	15	H	No
kirby21	22	5	5	H	No
IBC	8	2	2	H	No
BraTS20	280	40	49	T	Yes
BraTS19 (2D)	2710	314	319	T	Yes
MSSEG	12	3	37	MS	Yes
OFSEP	401	50	50	MS	No

Three public FLAIR MRI datasets have been used for healthy images: MPI [35], kirby21 [36] and IBC [37]. MICCAI BraTS 2019 and 2020 [38]–[40] have been used for brain tumors images and MICCAI MSSEG 2016 [41] and the OFSEP/EDMUS dataset[1] from the "Observatoire français de la sclérose en plaques" [42], [43] for multiple sclerosis images. Images were acquired in different centers with multi-brand, 1.5T and 3T scanners. These datasets were split in training/validation/test as indicated in Table I.

MICCAI BraTS 2019 was used as in [23]. Other datasets were preprocessed using FSL FLIRT affine registration on MNI atlas MRI [44], [45], HD-BET brain extraction [46] and N4 bias field correction [47]. As bias field correction can be detrimental to tumors segmentation, the BraTS preprocessing pipeline[2] does not include this step, as opposed to MS lesions segmentation challenges [41]. In this work, bias field correction is always used for MS data and both versions of the preprocessing pipeline, with (simulating poor quality images with low contrast) and without bias field correction, are used on BraTS 2020 data.

Experiments were done on two voxel-size resolutions: 2mm and 1mm. The final image size is $91 \times 109 \times 91$ for a 2mm voxel size and $182 \times 218 \times 182$ for a 1mm voxel size.

[1] `http://www.ofsep.org`: The confidentiality and safety of OFSEP data are ensured by the recommendations of the French Commission Nationale de l'Informatique et des Libertés (CNIL). This study was covered by the Reference Methodology MR-004 of the CNIL.

[2] `https://cbica.github.io/CaPTk/preprocessing_brats.html`

TABLE II: G equivalence to EG. Pearson correlation coefficient between attribution maps A1 and A2 given in first two columns measured on 2mm brain tumors images. In the notation X_Y, X is the map constrained during training, Y is the inference map.

A1	A2	Sup	Unsup	UnsupTV	Ross	Erion	Average
EG_EG	NO_EG	0.09 ± 0.04	0.46 ± 0.07	0.45 ± 0.06	0.48 ± 0.08	0.30 ± 0.09	0.36 ± 0.07
	IG_EG	0.36 ± 0.09	0.59 ± 0.10	0.57 ± 0.07	0.83 ± 0.03	$\mathbf{0.54 \pm 0.07}$	0.58 ± 0.07
	G_EG	$\mathbf{0.82 \pm 0.06}$	$\mathbf{0.81 \pm 0.06}$	$\mathbf{0.65 \pm 0.09}$	$\mathbf{0.93 \pm 0.02}$	0.40 ± 0.06	$\mathbf{0.72 \pm 0.06}$
IG_IG	G_IG	0.64 ± 0.10	0.44 ± 0.14	0.64 ± 0.13	0.58 ± 0.17	0.50 ± 0.05	0.56 ± 0.12

TABLE III: IEG robustness. Pearson correlation coefficient between attribution maps A1 and A2 given in first two columns measured on 2mm brain tumors images. In the notation X_Y, X is the map constrained during training, Y is the inference map.

A1	A2	Sup	Unsup	UnsupTV	Ross	Erion	Average
EG_EG	IG_EG	0.36 ± 0.09	0.59 ± 0.19	0.57 ± 0.07	0.83 ± 0.03	$\mathbf{0.54 \pm 0.07}$	0.58 ± 0.09
	IEG_EG	$\mathbf{0.95 \pm 0.03}$	$\mathbf{0.86 \pm 0.07}$	$\mathbf{0.75 \pm 0.07}$	$\mathbf{0.87 \pm 0.03}$	0.27 ± 0.07	$\mathbf{0.74 \pm 0.05}$
IG_IG	EG_IG	0.36 ± 0.12	0.43 ± 0.09	0.52 ± 0.10	0.63 ± 0.23	$\mathbf{0.48 \pm 0.10}$	0.48 ± 0.13
	IEG_IG	$\mathbf{0.66 \pm 0.09}$	$\mathbf{0.67 \pm 0.05}$	$\mathbf{0.61 \pm 0.09}$	$\mathbf{0.79 \pm 0.09}$	0.25 ± 0.08	$\mathbf{0.59 \pm 0.08}$

Two different setups were used for anomaly detection: either using the three healthy datasets for reconstruction training and BraTS20 and MSSEG for anomaly detection at test or using the setup of [23] with BraTS19 middle slices without tumor as the healthy database and BraTS19 middle slices with more than 0.1% of tumor for the pathological database. In this last case only, our method was used in 2D.

Note that for supervised model on MS, both MSSEG and OFSEP datasets were used for L_C but only annotated MSSEG for L_A on pathological class.

B. Implementation details

Our network was implemented using Pytorch. The source code is available on GitHub[3]. We used a 3D 70x70 PatchGan [48] as classifier. This CNN is defined as $C64$-$C128$-$C256$-$C512$ where Ck denotes a Convolution-BatchNorm-LeakyReLU (slope 0.2) layer with k filters, except for the first layer on which no BatchNorm is applied. At the end, a convolution is applied to obtain a 1-dimensional output. Our model is trained with Yogi optimizer [49] with an initial learning rate of 1 and AMSGrad [50]. When no constraint is applied, the discriminator is trained with Adam optimizer with an initial learning rate of 10^{-3}. The optimizers were chosen for training stability in all experiments. State-of-the-art models were trained with their original optimizer. The batch size was set to 5 and data augmentation with random brightness variation, elastic deformation and mirroring along the mid-sagittal plane was used. IG and EG constraints were implemented with a single random α. A null baseline was used for IG. The null baseline probability of IEG was set to $p = 0.25$. In our experiments, the choice of this p parameter did not have much influence. IG was computed at inference using Captum[4]. EG was computed with the test database images as baseline at inference. The attribution coefficient α_A in Equation 1 was set to 10^8 (see V-A for experiments on this coefficient). For state-of-the-art methods, original coefficients were used. Early stopping was applied on classification accuracy metric on the validation set. For the comparison with anomaly detection methods, we used the code provided by [23]. Experiments were done using a Nvidia V100 GPU with 32 GB of memory.

[3]https://github.com/Valentine-Wargnier-Dauchelle/gradient-attributions-constraint

[4]https://captum.ai

C. Evaluation metrics

The image level classification performance was measured with the true negative rate (TNR, the proportion of well-classified healthy images) and true positive rate (TPR, the proportion of well-classified pathological images) of the healthy vs pathological classifier.

The quality of the attribution maps was measured using the Dice [51] between the thresholded attributions and ground truth pathology mask. We also used the area under the precision-recall curve (AUPRC) and the receptive-operative curve (AUROC, which plots false vs true positive rate) considering voxel level classification. As AUROC is not adapted for unbalanced classes, we also computed the area under the ROC up to a false positive rate of 10% (AUROC10) [52]. Indeed, increasing the false positive rate will rapidly degrade the segmentation quality. The resulting area is normalized according to the maximum attainable value. In Sections V-C and V-D, the thresholds were chosen for each method as the PRC operating point with the 2mm tumors validation dataset and then used for all test sets.

Correlation between attribution maps in Section V-B was measured with the Pearson coefficient defined as the covariance of the two maps divided by the product of their standard deviation.

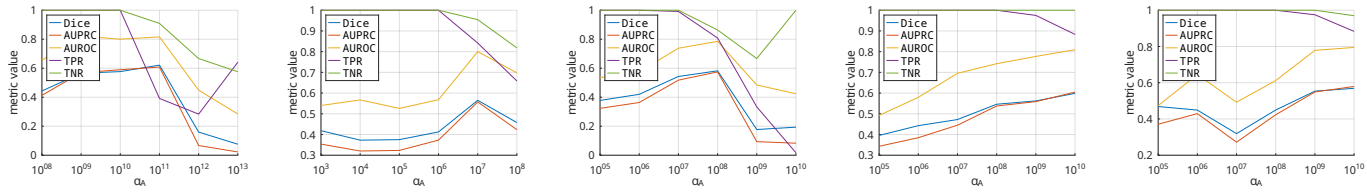
Statistical significance tests were performed comparing Unsup to each of the state-of-the-art methods. We used a permutation test with 10000 permutations, a 95% confidence and the Bonferroni correction.

V. Results

A. Loss weight

Optimal loss weight α_A introduced in Equation 1 was set on validation set for tumors dataset (without N4 bias field correction). We monitored all metrics by varying alpha for

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3282789



(a) EG (b) IG (c) IEG (d) G 2mm (e) G 1mm

Fig. 2: Loss weight α_A influence. Metrics (Dice, AUROC, AUPRC, TPR, TNR) versus α_A for different attributions methods. Experiments was performed on validation 2mm (+1mm for gradient) tumors set.

our unsupervised model. Results are presented in Figure 2. We choose the maximum coefficient keeping almost perfect classification (TPR and TNR close to 1) as segmentation metrics increase with α_A. With this choice, we aim to have a good classification and segmentation model. Nevertheless, segmentation performances can be improved by choosing a higher coefficient at the expense of classification performances. We note that pixel size (and so image size) does not impact optimal coefficient as TPR curves pick down for the same α_A for 2mm and 1mm voxel size images (Figure 2d and 2e). Thus, we set $\alpha_A = 10^{10}$ for EG, $\alpha_A = 10^6$ for IG, $\alpha_A = 10^7$ for IEG and $\alpha_A = 10^8$ for G. Note that, with EG, this coefficient was reduced to 10^9 for the supervised model (Sup) and when the TV loss is added (UnsupTV) as the model does not converge with the unsupervised model optimal coefficient. EG training seems less stable and more parameter dependent.

The optimal α_A coefficients, optimized on the validation set of the tumors dataset without N4 bias field correction, have been used in the remaining of the paper, for all the different experiments.

B. Influence of the attribution map method

1) Equivalence between Gradient and Expected Gradient constraint: In this section, we give some quantitative clues to answer the question raised in Section III-C.2: are constraints on G sufficient for EG ? In Table II, we report the Pearson correlation between inference attribution maps when different attribution maps are used in the constraint during training. One can first see that EG maps produced by EG constrained model (EG_EG) and unconstrained model (NO_EG) are not correlated. Adding any of the IG or G constraints drastically increases the correlation (between 12% and 73%). One can also see that the correlation is higher for G constrained maps than for IG ones in most cases and they are 14% more correlated in average.

G constrained training is also beneficial regarding its fit with pathology: if we consider the Dice between ground truth and positive attributions (presented in Figure 3) and the AUPRC (presented in Figure 5), one can see that results for both metrics are equivalent or improved when the constraint is applied on G instead of EG during the training.

As far as computation time is concerned, EG is about 50% slower than G for each iteration and convergence with G is generally easier to achieve.

Finally, Table II shows that G constrained maps are less correlated with IG than EG. For most tested cases, the correlation is lower between IG_IG and G_IG than EG_EG and G_EG. G constraint is not sufficient for robustness to IG at the inference but sufficient for good performances with EG inference, with easier training.

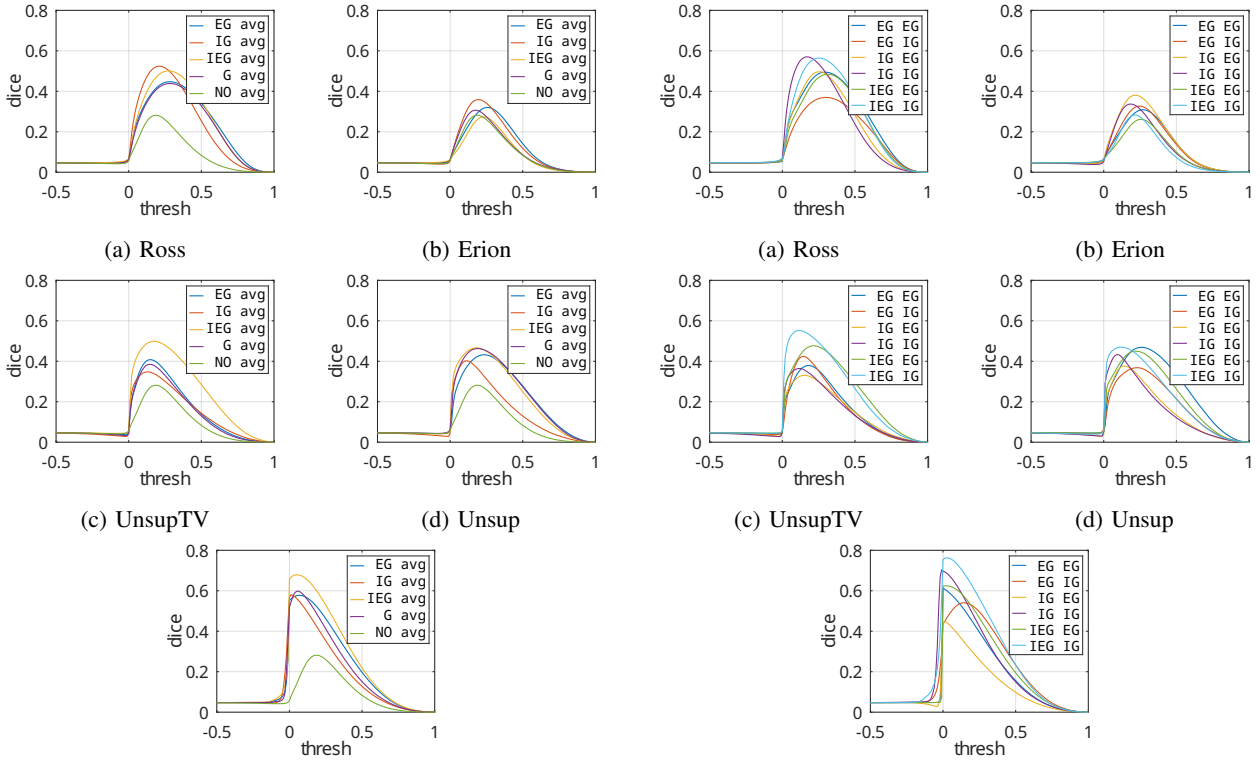
2) Robust training with integrated/expected gradient mix: Training with a mix between EG and IG (named IEG) is better than Gradient in terms of Dice (yellow curve in Figure 3) and AUPRC (green bar in Figure 5). This is the most efficient method in average: with this training, performances (Dice and AUPRC) remain constant according to the constraint choice (Sup, Unsup, UnspTV, Erion or Ross) whereas EG or IG reach very good performances for some and drop drastically for others. For example, training with IG is not efficient for UnsupTV and EG_IG map is the worst for Ross in terms of AUPRC.

This proposition provides a more robust model for the attribution inference method at the expense of a computational cost compared to G but no cost compared to EG or IG constraints. IEG is better or equivalent than EG constraint evaluated on IG and vice versa especially on Dice metric (Figure 4: light blue vs orange and green vs yellow curves) and AURPC (Figure 6: blue vs yellow and pigeon blue vs light green bars). In addition, with this training, attribution maps on IG are 11% more correlated to the reference IG_IG than when EG is used for the training and attribution maps on EG are 14% more correlated to reference than with IG only training (Table III).

C. Interpretable and relevant classification

A state-of-art comparison was made using the original attribution training method and G at the inference as it is included in every attribution method (EG beginning of the path). For our method, we use the G constraint for its speed.

As shown in Figure 7, our unsupervised constraint is clearly based on tumor area since attributions focus on it. Ross is visually equivalent to ours in the first example but less specific in the more difficult second example. Decisions with the other methods are less based on the region of interest. This is supported by the metrics presented in Table IV. Indeed, for BraTS 2mm, the Dice is much better: it is more than 20 points higher than the model with classification only training and Gradient as attributions evaluation (NoConsG) and around 40 points higher when GradCam (NoConsGC) is used. Our Unsup model outperforms state-of-the-art methods with a gain of at least 3 Dice points (statistically significant). Our model is more accurate with an AUPRC 6 points higher than the second-best



(a) Ross (b) Erion (c) UnsupTV (d) Unsup (e) Sup

Fig. 3: G equivalence to EG. Dice versus threshold value for different constraints and different attribution maps used during training for 2mm brain tumors data. Results are averaged on all attribution inference methods.

(a) Ross (b) Erion (c) UnsupTV (d) Unsup (e) Sup

Fig. 4: IEG robustness. Dice versus threshold value for different constraints and different attribution maps for 2mm brain tumors data. In the legend, left is the attribution maps used in the constraint during training, right is the attribution map used during inference.

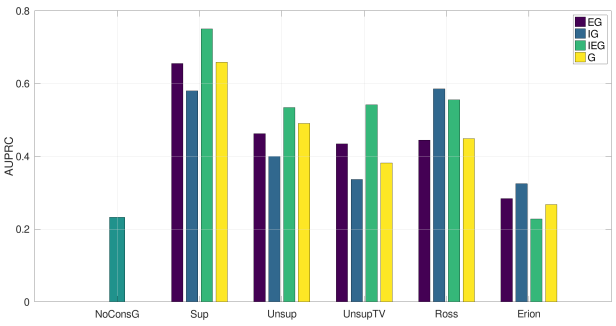


Fig. 5: G equivalence to EG. Influence of the attribution maps used in the constraint on the AUPRC for 2mm brain tumors data. Results are averaged on all attribution inference methods.

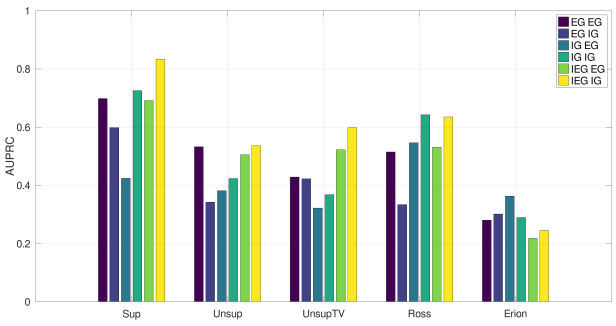


Fig. 6: IEG robustness. Influence of the attribution maps used in the constraint on the AUPRC for 2mm brain tumors data. In the legend, left is the attribution maps used in the constraint during training, right is the attribution map used at inference.

model (Ross). AUROC is a little bit lower but less adapted than AUPRC as pathological voxels are underrepresented compared to healthy ones. Looking at AUROC with a false positive rate of less than 10% (AUROC10), Ross and our unsupervised methods are equivalent. This difference between AUROC and AUROC10 shows that Ross could reach a higher sensitivity but with poor specificity. In words, more lesions could be detected with Ross but at the cost of an overwhelming number of false positives. When increasing the resolution, state-of-the-art method's performances increase as the segmentation task is easier. Especially, NoConsGC is the second-best method in terms of Dice whereas it is the worst for 2mm voxel size images. Nevertheless, our method is still better in terms of Dice and AUPRC. Using the BraTS dataset with N4 bias field correction, images with poorer contrast are considered and the segmentation task is more difficult. In this case, our

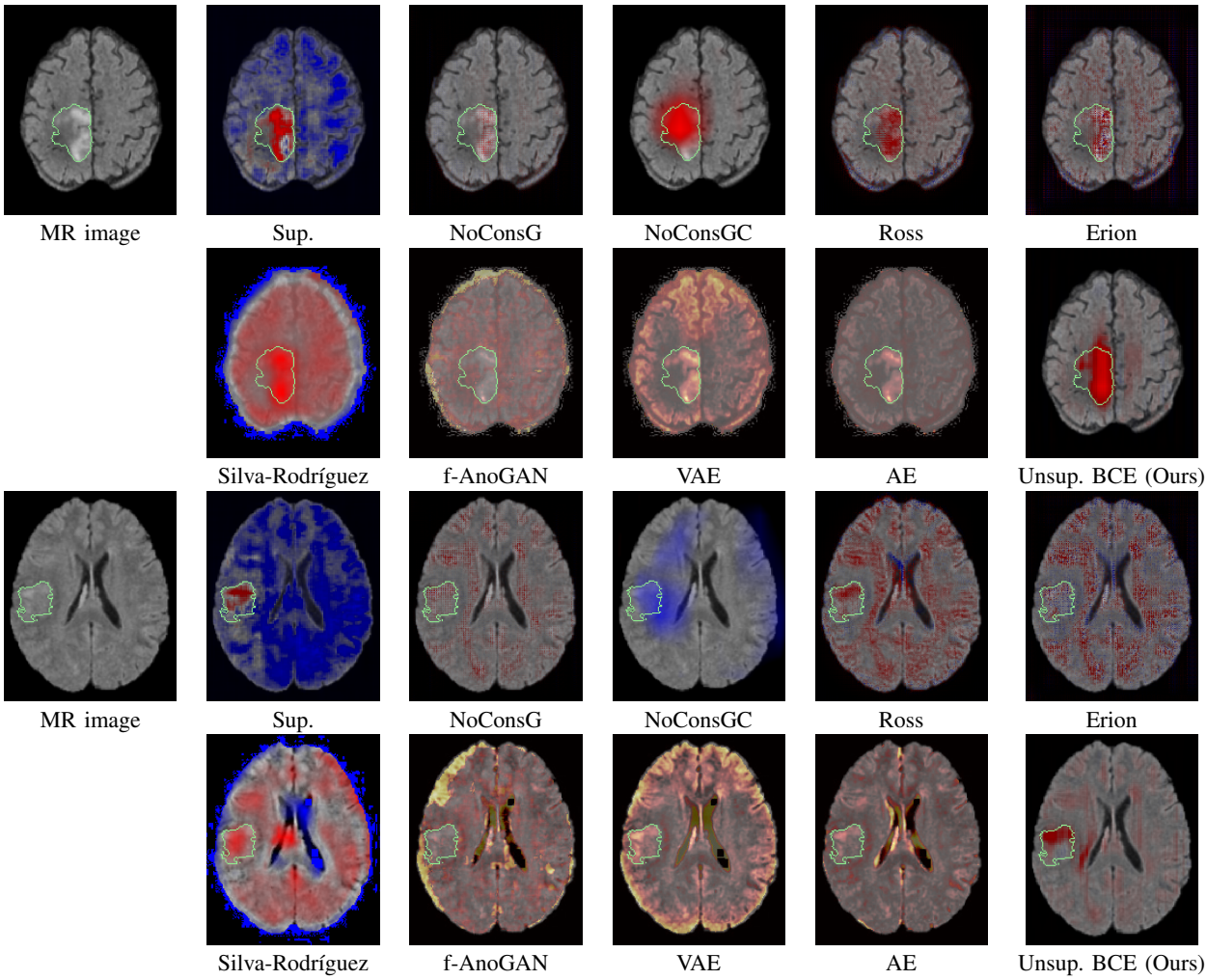


Fig. 7: Segmentation maps (attributions or reconstruction error) for different methods on brain tumors images (1mm with N4 correction). Manual annotation is drawn in green. Blue represents healthy relevance and red pathological relevance for classification attribution methods. High attributions are in red for Silva-Rodríguez. For reconstruction methods, reconstruction error scale is from black to yellow. From left to right, top to bottom: MR image, supervised, NoConsG, NoConsGC, Ross, Erion, Silva-Rodríguez, f-AnoGAN, VAE, AE, our proposed unsupervised methods.

method is much better with a 20 points gap for Dice and an AUPRC twice better than the second-best method. Under these conditions, all metrics are statistically significant. Our constraint allows for a more relevant classifier in the meaning that network decision is more based on clinically relevant structures, brain tumors, without degrading the classification performances with a TPR and TNR higher than 95%.

In multiple sclerosis, lesions are smaller than brain tumors and even in this case, our unsupervised proposed method stands out through its performance compared to literature methods. Visual examples in Figure 8 shows that our proposition detects more lesions (second example) and it is more specific: in the first example, with our constraint, the areas of high attributions are focused on MS lesions whereas with Ross, high attributions are spread out around ventricles. Thus, Dice with our method is 25 points higher than NoConsG which discriminates healthy and MS images regardless of lesions. In comparison with the second-best method, our proposition achieves a three times better Dice and an AUPRC five times higher. Classification performances are not too degraded with an accuracy of about 90%. We notice that the constraints (supervised and unsupervised) make the classification harder as accuracy is lower and training harder. It is likely that without the constraint, the network uses some shortcut based on the datasets signature instead of clinical features. Note that the split in train/validation/test traditionally used to detect overfitting will not detect this kind of problem: if the global dataset has a bias of some sort, the bias will be present in each of the subsets and high test accuracy might be based on this bias.

Therefore, using unsupervised BCE loss for constrained training allows an interpretable classification with decision more based on the pathology area, outperforming classification without constraint and state-of-the-art constraints. Our method is particularly efficient for difficult segmentation tasks (MS or low contrast tumor images) for which the difference in performance with state-of-the-art is large and significant.

Moreover, with our method, the voxels of healthy images

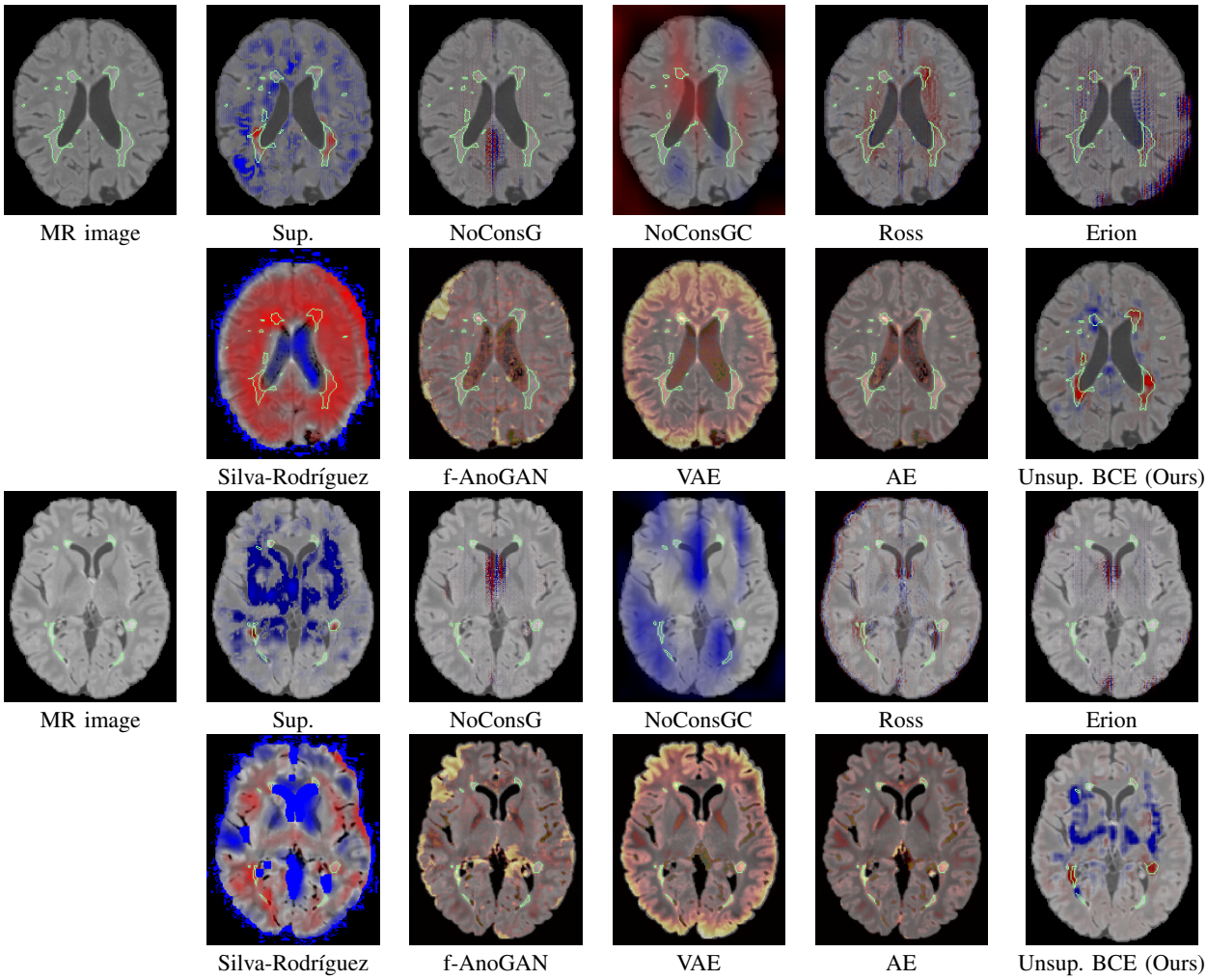


Fig. 8: Segmentation map (attributions or reconstruction error) for different methods on MS images (1mm). Manual annotation is drawn in green. Blue represents healthy relevance and red pathological relevance for classification attribution methods. High attributions are in red for Silva-Rodríguez. For reconstruction methods, reconstruction error scale is from black to yellow. From left to right, top to bottom: MR image, supervised, NoConsG, NoConsGC, Ross, Erion, Silva-Rodríguez, f-AnoGAN, VAE, AE, our proposed unsupervised methods.

contribute to the healthy classification decision as the attributions of healthy images are negative, that is to say, relevant for healthy class. Indeed, for our method, the histogram of healthy attributions (Figure 9 in blue) vanishes in the positive area and the pathological curve (in orange) is both negative, for healthy regions, and positive, for tumors areas as shown with segmentation evaluation. In comparison, healthy attributions of other methods are partially positive and both classes histogram curves are mingled.

D. Anomaly detection

The proposed constraint model can be used for weakly supervised anomaly detection with only image-level label. In Table V, metrics for the comparison of our method to state-of-the-art on different datasets are reported. Our method outperforms other methods for Dice, AUPRC and AUROC10 (always statistically significant for the first two). For low resolution images, the proposed method surpasses state-of-the-art by almost 15 points of Dice and AUPRC. By increasing resolution to 1mm voxel-size, Dice gap between our method and the best state-of-the-art method (Silva-Rodríguez) increases by 5 points. For more difficult tasks (BraTS with N4 correction and MS), the best literature method is AE but our method is still more efficient with 15 points higher Dice and AUPRC.

Visually, in Figure 7 and 8, our unsupervised method seems more specific than others especially VAE and Silva-Rodríguez which detect anomalies in healthy tissue. Silva-Rodríguez seems competitive only with the second setup (middle slices extracted from the same Brats2019 dataset for both healthy and pathological). Nevertheless, it reaches 8 points lower Dice and still less precise with an AUPRC 20 points lower than our constrained training. Note that in most real applications, the whole image is given and the segmentation algorithm must be able to handle slices from the whole brain and especially outputs zero mask segmentation for healthy slices.

TABLE IV: Comparison to state-of-the-art for interpretable classification on brain tumors and MS. Statistical difference with Unsup is indicated with †.

		Attributions segmentation				Images class.	
Dataset	Method	Dice	AUROC	AUROC10	AUPRC	TPR	TNR
BraTS 2020 2mm	Supervised	$0.71 \pm 0.17^{\dagger}$	$0.85^{\dagger}$	$0.76^{\dagger}$	$0.73^{\dagger}$	1.00	0.95
	NoConsG	$0.29 \pm 0.16^{\dagger}$	$0.61^{\dagger}$	$0.34^{\dagger}$	$0.16^{\dagger}$	**1.00**	**1.00**
	NoConsGC	$0.12 \pm 0.16^{\dagger}$	$0.62^{\dagger}$	$0.15^{\dagger}$	$0.04^{\dagger}$	**1.00**	**1.00**
	Ross	$0.48 \pm 0.20^{\dagger}$	**0.80**	**0.63**	0.39	**1.00**	**1.00**
	Erion	$0.29 \pm 0.15^{\dagger}$	$0.70^{\dagger}$	$0.40^{\dagger}$	$0.19^{\dagger}$	**1.00**	**1.00**
	Proposed (Unsup)	$\mathbf{0.51 \pm 0.16}$	0.73	0.62	**0.45**	**1.00**	0.95
BraTS 2020 1mm	Supervised	$0.70 \pm 0.15^{\dagger}$	$0.78^{\dagger}$	$0.68^{\dagger}$	$0.66^{\dagger}$	0.93	0.95
	NoConsG	$0.27 \pm 0.13^{\dagger}$	$0.70^{\dagger}$	$0.38^{\dagger}$	$0.18^{\dagger}$	0.86	**1.00**
	NoConsGC	$0.48 \pm 0.20^{\dagger}$	**0.90**†	0.65	0.40	0.86	**1.00**
	Ross	$0.40 \pm 0.19^{\dagger}$	$0.89^{\dagger}$	**0.66**†	0.44	**1.00**	**1.00**
	Erion	$0.29 \pm 0.18^{\dagger}$	$0.78^{\dagger}$	$0.36^{\dagger}$	$0.19^{\dagger}$	**1.00**	**1.00**
	Proposed (Unsup)	$\mathbf{0.52 \pm 0.17}$	0.69	0.56	**0.45**	**1.00**	**1.00**
BraTS 2020 1mm with N4 correction	Supervised	$0.55 \pm 0.18^{\dagger}$	$0.70^{\dagger}$	0.55	$0.49^{\dagger}$	0.98	1.00
	NoConsG	$0.18 \pm 0.07^{\dagger}$	$0.71^{\dagger}$	$0.33^{\dagger}$	$0.14^{\dagger}$	**1.00**	**1.00**
	NoConsGC	$0.20 \pm 0.17^{\dagger}$	$0.79^{\dagger}$	$0.19^{\dagger}$	$0.08^{\dagger}$	**1.00**	**1.00**
	Ross	$0.19 \pm 0.11^{\dagger}$	**0.77**†	$0.38^{\dagger}$	$0.17^{\dagger}$	**1.00**	**1.00**
	Erion	$0.16 \pm 0.07^{\dagger}$	$0.66^{\dagger}$	$0.23^{\dagger}$	$0.08^{\dagger}$	**1.00**	**1.00**
	Proposed (Unsup)	$\mathbf{0.38 \pm 0.15}$	0.73	**0.50**	**0.30**	0.94	**1.00**
MS 1mm	Supervised	0.24 ± 0.18	$0.53^{\dagger}$	0.46	$0.24^{\dagger}$	0.77	0.77
	NoConsG	$0.001 \pm 0.002^{\dagger}$	$0.63^{\dagger}$	0.41	$0.02^{\dagger}$	**1.00**	**1.00**
	NoConsGC	$0.0003 \pm 0.0007^{\dagger}$	$0.34^{\dagger}$	$0.01^{\dagger}$	$0.002^{\dagger}$	**1.00**	**1.00**
	Ross	$0.09 \pm 0.09^{\dagger}$	**0.70** †	0.45	$0.04^{\dagger}$	**1.00**	**1.00**
	Erion	$0.01 \pm 0.01^{\dagger}$	0.60	$0.35^{\dagger}$	$0.01^{\dagger}$	**1.00**	**1.00**
	Proposed (Unsup)	$\mathbf{0.25 \pm 0.16}$	0.60	**0.51**	**0.20**	0.89	0.91

TABLE V: Comparison to state-of-the-art for anomaly detection on brain tumors and MS. Statistical difference with Unsup is indicated with †.

Dataset	Method	Dice	AUROC	AUROC10	AUPRC
BraTS 2020 2mm	Silva-Rodríguez	$0.37 \pm 0.17^{\dagger}$	**0.92**†	0.56	$0.32^{\dagger}$
	AE	$0.26 \pm 0.11^{\dagger}$	$0.90^{\dagger}$	$0.36^{\dagger}$	$0.16^{\dagger}$
	VAE	$0.25 \pm 0.14^{\dagger}$	$0.91^{\dagger}$	$0.30^{\dagger}$	$0.15^{\dagger}$
	f-AnoGAN	$0.17 \pm 0.10^{\dagger}$	0.79	$0.06^{\dagger}$	$0.06^{\dagger}$
	Proposed (Unsup)	$\mathbf{0.51 \pm 0.16}$	0.73	**0.62**	**0.45**
BraTS 2020 1mm	Silva-Rodríguez	$0.33 \pm 0.20^{\dagger}$	**0.91**†	0.52	$0.32^{\dagger}$
	AE	$0.28 \pm 0.11^{\dagger}$	**0.91**†	0.43	$0.17^{\dagger}$
	VAE	$0.24 \pm 0.13^{\dagger}$	**0.91**†	$0.34^{\dagger}$	$0.15^{\dagger}$
	f-AnoGAN	$0.16 \pm 0.10^{\dagger}$	$0.84^{\dagger}$	$0.16^{\dagger}$	$0.09^{\dagger}$
	Proposed (Unsup)	$\mathbf{0.52 \pm 0.17}$	0.69	**0.56**	**0.45**
BraTS 2020 1mm with N4 correction	Silva-Rodríguez	$0.18 \pm 0.09^{\dagger}$	**0.88**†	$0.19^{\dagger}$	$0.14^{\dagger}$
	AE	$0.23 \pm 0.10^{\dagger}$	**0.88**†	0.35	$0.15^{\dagger}$
	VAE	$0.18 \pm 0.10^{\dagger}$	$0.87^{\dagger}$	$0.26^{\dagger}$	$0.11^{\dagger}$
	f-AnoGAN	$0.16 \pm 0.09^{\dagger}$	$0.84^{\dagger}$	$0.18^{\dagger}$	$0.09^{\dagger}$
	Proposed (Unsup)	$\mathbf{0.38 \pm 0.15}$	0.73	**0.50**	**0.30**
BraTS 2019 1mm middle slices	Silva-Rodríguez	0.54 ± 0.22	**0.97**	**0.78**	$0.52^{\dagger}$
	Proposed (2D Unsup)	$\mathbf{0.62 \pm 0.14}$	0.96	**0.78**	**0.72**
MS 1mm	Silva-Rodríguez	$0.10 \pm 0.08^{\dagger}$	$0.93^{\dagger}$	0.50	$0.05^{\dagger}$
	AE	$0.10 \pm 0.09^{\dagger}$	**0.95**†	0.67	$0.05^{\dagger}$
	VAE	$0.03 \pm 0.03^{\dagger}$	$0.92^{\dagger}$	0.36	$0.01^{\dagger}$
	f-AnoGAN	$0.02 \pm 0.02^{\dagger}$	$0.81^{\dagger}$	$0.10^{\dagger}$	$0.01^{\dagger}$
	Proposed (Unsup)	$\mathbf{0.25 \pm 0.16}^{\dagger}$	0.60	**0.51**	**0.18**

VI. Conclusion

In this paper, we proposed an unsupervised method to constrain the decision of a classification network to be based on the pathology using attribution maps as a reflection of the decision. We show on two brain pathologies that attributions, a marker of the network decision, are more focused on the pathological areas and are consequently, more in accordance with high-level medical knowledge when the new constraint is used during the training. Moreover, these attribution maps could be used for weakly supervised anomaly detection. Thus, the proposed method outperforms both state-of-the-art interpretable classification and anomaly detection methods.

A deep analysis of the attribution choice for constrained training is also made. We proved that by using the Gradient attributions, which are more easily integrated and faster to compute, the constrained training reaches similar performances to

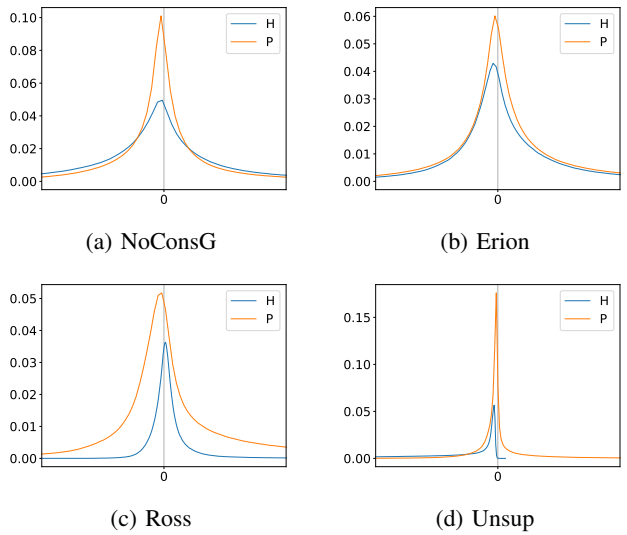


Fig. 9: Attributions histogram for different methods. Experiments were done on 2mm healthy (H) and tumors (P) sets without N4 correction.

that obtained by using Expected Gradient, a complex high-level attribution. We also propose a new attribution integration during training in order to be invariant to gradient-based attribution method used for inference. This constraint formulation integrates all existing gradient methods in one constraint.

As our proposition is a simple loss to be applied during training, it can be easily integrated into all deep models without changing their architecture. Thereby, this work could be used in several domains. As discriminators are a building block of adversarial networks, we can use attribution constraints to increase the performances and relevance of GAN methods for anomaly detection for instance. This work could also be extended to other kinds of architectures and not only to classifiers like regression and prediction networks: disease grade estimation which focuses on pathology, body registration transformation estimation with anatomical prerequisites, etc.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017.

[2] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annual review of biomedical engineering, vol. 19, pp. 221–248, 2017.

[3] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," Annals of translational medicine, vol. 8, no. 11, 2020.

[4] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," Physics in Medicine & Biology, vol. 65, no. 20, p. 20TR01, 2020.

[5] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnu-net for brain tumor segmentation," in International MICCAI Brainlesion Workshop. Springer, 2020, pp. 118–132.

[6] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3d convolutional neural networks," in International conference on medical image computing and computer-assisted intervention. Springer, 2017, pp. 232–239.

[7] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network," Applied Intelligence, vol. 51, no. 2, pp. 854–864, 2021.

[8] J. Gao, Q. Jiang, B. Zhou, and D. Chen, "Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview," Mathematical Biosciences and Engineering, vol. 16, no. 6, pp. 6536–6561, 2019.

[9] Y. Bahat, M. Irani, and G. Shakhnarovich, "Natural and adversarial error detection using invariance to image transformations," arXiv preprint arXiv:1902.00236, 2019.

[10] F. Alharbi, K. El Hindi, S. Al Ahmadi, and H. Alsalamn, "Convolutional neural network-based discriminator for outlier detection," Computational Intelligence and Neuroscience, vol. 2021, 2021.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), 2014, pp. 2672–2680.

[12] S. Cackowski, E. L. Barbier, M. Dojat, and T. Christen, "Imunity: a generalizable vae-gan solution for multicenter mr image harmonization," arXiv preprint arXiv:2109.06756, 2021.

[13] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur et al., "Harmonization of multi-site diffusion tensor imaging data," Neuroimage, vol. 161, pp. 149–170, 2017.

[14] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu et al., "Statistical normalization techniques for magnetic resonance imaging," NeuroImage: Clinical, vol. 6, pp. 9–19, 2014.

[15] V. Wargnier-Dauchelle, T. Grenier, F. Durand-Dubief, F. Cotton, and M. Sdika, "A more interpretable classifier for multiple sclerosis," in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021, pp. 1062–1066.

[16] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," Nature Machine Intelligence, vol. 3, no. 7, pp. 610–619, 2021.

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International Conference on Machine Learning. PMLR, 2017, pp. 3319–3328.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[19] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," Nature Machine Intelligence, pp. 1–12, 2021.

[20] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," arXiv preprint arXiv:2111.02398, 2021.

[21] L. Dumortier, F. Guépin, M.-L. Delignette-Muller, C. Boulocher, and T. Grenier, "Deep learning in veterinary medicine, an approach based on cnn to detect pulmonary abnormalities from lateral thoracic radiographs in cats," Scientific reports, vol. 12, no. 1, pp. 1–12, 2022.

[22] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: training differentiable models by constraining their explanations," in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 2662–2670.

[23] J. Silva-Rodríguez, V. Naranjo, and J. Dolz, "Looking at the whole picture: constrained unsupervised anomaly segmentation," in BMVC, 2021.

[24] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study," Medical Image Analysis, vol. 69, p. 101952, 2021.

[25] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in International MICCAI brainlesion workshop. Springer, 2018, pp. 161–169.

[26] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 289–297.

[27] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," Medical image analysis, vol. 54, pp. 30–44, 2019.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," arXiv preprint arXiv:1312.6034, 2013.

[29] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," arXiv preprint arXiv:1711.06104, 2017.
[30] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," Distill, vol. 5, no. 1, p. e22, 2020.
[31] I. Bárány and Z. Füredi, "On the shape of the convex hull of random points," Probability theory and related fields, vol. 77, no. 2, pp. 231–240, 1988.
[32] R. Balestriero, J. Pesenti, and Y. LeCun, "Learning in high dimension always amounts to extrapolation," 2021. [Online]. Available: https://arxiv.org/abs/2110.09485
[33] R. Yousefzadeh, "Deep learning generalization and the convex hull of training sets," CoRR, vol. abs/2101.09849, 2021. [Online]. Available: https://arxiv.org/abs/2101.09849
[34] ——, "Decision boundaries and convex hulls in the feature space that deep learning functions learn from images," ArXiv, vol. abs/2202.04052, 2022.
[35] A. Babayan, M. Erbey, D. Kumral, J. D. Reinelt, A. M. Reiter, J. Röbbig, H. L. Schaare, M. Uhlig, A. Anwander, P.-L. Bazin et al., "A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults," Scientific data, vol. 6, no. 1, pp. 1–21, 2019.
[36] B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso et al., "Multi-parametric neuroimaging reproducibility: a 3-t resource study," Neuroimage, vol. 54, no. 4, pp. 2854–2866, 2011.
[37] A. L. Pinho, A. Amadon, T. Ruest, M. Fabre, E. Dohmatob, I. Denghien, C. Ginisty, S. Becuwe-Desmidt, S. Roger, L. Laurier et al., "Individual brain charting, a high-resolution fmri dataset for cognitive mapping," Scientific data, vol. 5, no. 1, pp. 1–15, 2018.
[38] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," Scientific data, vol. 4, no. 1, pp. 1–13, 2017.
[39] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," arXiv preprint arXiv:1811.02629, 2018.
[40] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest et al., "The multimodal brain tumor image segmentation benchmark (brats)," IEEE transactions on medical imaging, vol. 34, no. 10, pp. 1993–2024, 2014.
[41] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Ameli, J.-C. Ferré et al., "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," Scientific reports, vol. 8, no. 1, pp. 1–17, 2018.
[42] S. Vukusic, R. Casey, F. Rollot, B. Brochet, J. Pelletier, D.-A. Laplaud, J. De Sèze, F. Cotton, T. Moreau, B. Stankoff et al., "Observatoire français de la sclérose en plaques (ofsep): A unique multimodal nation-wide ms registry in france," Multiple Sclerosis Journal, vol. 26, no. 1, pp. 118–122, 2020.
[43] C. Confavreux, D. Compston, O. Hommes, W. McDonald, and A. Thompson, "Edmus, a european database for multiple sclerosis." Journal of Neurology, Neurosurgery & Psychiatry, vol. 55, no. 8, pp. 671–676, 1992.
[44] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," Medical image analysis, vol. 5, no. 2, pp. 143–156, 2001.
[45] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," Neuroimage, vol. 17, no. 2, pp. 825–841, 2002.
[46] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick et al., "Automated brain extraction of multisequence mri using artificial neural networks," Human brain mapping, vol. 40, no. 17, pp. 4952–4964, 2019.
[47] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," IEEE transactions on medical imaging, vol. 29, no. 6, pp. 1310–1320, 2010.
[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
[49] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar, "Adaptive methods for nonconvex optimization," Advances in neural information processing systems, vol. 31, 2018.
[50] T. Tan, S. Yin, K. Liu, and M. Wan, "On the convergence speed of amsgrad and beyond," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019, pp. 464–470.
[51] T. A. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons," Biol. Skar., vol. 5, pp. 1–34, 1948.
[52] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection," International Journal of Computer Vision, vol. 129, no. 4, pp. 1038–1059, 2021.