Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening

Zaruhi Alaverdyan, Julien Jung, Romain Bouet, Carole Lartizien

Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening

Zaruhi Alaverdyan[a], Julien Jung[b], Romain Bouet[b], Carole Lartizien[a]

[a] *Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621, Lyon, France*
[b] *Lyon Neuroscience Research Center, CRNL, INSERM U1028, CNRS UMR5292, University Lyon 1, Lyon, France*

Abstract

In this study, we propose a novel anomaly detection model targeting subtle brain lesions in multiparametric MRI. To compensate for the lack of annotated data adequately sampling the heterogeneity of such pathologies, we cast this problem as an outlier detection problem and introduce a novel configuration of unsupervised deep siamese networks to learn normal brain representations using a series of non-pathological brain scans. The proposed siamese network, composed of stacked convolutional autoencoders as subnetworks is designed to map patches extracted from healthy control scans only and centered at the same spatial localization to 'close' representations with respect to the chosen metric in a latent space. It is based on a novel loss function combining a similarity term and a regularization term compensating for the lack of dissimilar pairs. These latent representations are then fed into oc-SVM models at voxel-level to produce anomaly score maps. We evaluate the performance of our brain anomaly detection model to detect subtle epilepsy lesions in multiparametric (T1-weighted, FLAIR) MRI exams considered as normal (MRI-negative). Our detection model trained on 75 healthy subjects and validated on 21 epilepsy patients (with 18 MRI-negatives) achieves a maximum sensitivity of 61% on the MRI-negative lesions, identified among the 5 most suspicious detections on average. It is shown to outperform detection models based on the same architecture but with stacked convolutional or Wasserstein autoencoders as unsupervised feature extraction mechanisms.

Keywords: Regularized Siamese network, Wasserstein autoencoder, Unsupervised representation learning, Brain lesions, Anomaly detection, Deep Learning

1. Introduction

A number of neurological pathologies are characterized by small lesions of various shapes, localizations and spatial patterns. This encompasses small vessel disease (SVD) (Wardlaw et al. (2013)), intracranial carotid artery calcification (ICAC) (Bos et al. (2014, 2012)) which both may lead to stroke as well as cognitive impairment or dementia, multiple sclerosis (MS) (Filippi et al. (2016)) and medically refractory epilepsy, which is often associated with malformations of cortical development (MCD), such as heterotopia or focal cortical dysplasia (FCD) (Guerrini et al. (2003)). For all these pathologies, the characterization of lesion profiles is crucial to perform an early diagnostic as well as define and monitor the optimal therapeutic strategy.

Neuroimaging techniques, especially those based on multiparametric magnetic resonance imaging (MRI), have been exploited to detect, characterize and monitor this type of pathologies in a non-invasive manner. The detection of small and subtle brain lesions during standard visual examinations, however, remains challenging. As an example, recent retrospective studies involving surgical epilepsy patients indicate that 30% to 80% of lesions, depending on their type, go undetected during routine MRI exams (Bernasconi and Bernasconi (2015)), dramatically impacting the success of resective surgery.

Considering the impact the detection of subtle brain lesions on MR imaging has on the diagnostic and therapeutic orientation, the development of automated neuroimaging analysis tools, providing probabilistic maps of the suspicious regions, may be of valuable help to neurologists during the diagnostic phase or therapeutic monitoring. The automatic detection of subtle brain pathologies, including MS, SVD and epilepsy lesions, on multiparametric MR imaging has been increasingly addressed over the past ten years. The advances in machine learning and, more recently, deep learning have further motivated new studies in this domain. A vast majority of the existing works was performed in the supervised learning framework, thus requiring large and accurately annotated data sets. However, acquiring a data set adequately representing the heterogeneity of the pathology at hand is a major issue, more so for the lesions that may be located anywhere in the organ of interest, have various shapes, size and texture, or even can not be visually identified on the images.

In this study, we tackle the problem of subtle brain lesion detection as an outlier detection problem, to both encompass the lack of annotated pathological cases and the class imbalance problem impacting the performance of supervised learning models. We consider this challenging detection problem from the perspective of unsupervised learning combining data-driven deep feature extraction and outlier detection in a single framework. We specifically target the detection of *cryptogenic* epilepsy lesions in multiparametric MRI, focusing on MRI-negative lesions, meaning that these lesions were not visually identified by clinicians on the MR scans, as illustrated on figure 1.

This work builds on the method that we proposed in El Azami et al. (2016)

Figure 1: Examples of three patients with confirmed epileptogenic lesions on FLAIR (left) and T1-weighted MRI scans, respectively. The ground truth is annotated in red circles. The first two lesions were detected during a visual analysis (MRI-positive); the third patient has an anomaly in the hippocampus that was not detected on MRI (*MRI-negative*).

to detect FCD and heterotopia lesions on T1-weighted (T1w) MR images. This pipeline combines a hand-crafted feature extraction mechanism targeting typical FCD lesions with an outlier detection model based on one-class support vector machine (oc-SVM). We generalize this framework to detect a wider range of subtle brain anomalies by exploiting unsupervised deep learning architectures as feature extraction mechanisms.
Our contributions in this work are summarized below:

1. We define a novel unsupervised representation model of normal brain patterns based on a siamese network composed of stacked convolutional autoencoders as subnetworks and a loss function tailored to the context of outlier detection.
2. This unsupervised deep representation is coupled to a one class SVM model to derive a brain anomaly detection model highlighting suspicious regions in 3D brain MRI at voxel level.
3. This pipeline is applied to the detection of MRI negative lesions in multiparametric MRI (T1w and FLAIR) with competitive performances with regards to state-of-the art methods.

A proof of concept of our anomaly detection pipeline was presented in Alaverdyan et al. (2018a,b). In this work, we build on this preliminary study, first, by providing a detailed description of the model and performing an extensive performance analysis based on a clinical dataset including 75 healthy subjects and 21 epilepsy cases (including 18 MRI negative patients). We then compare our unsupervised deep latent representation model based on the early fusion of T1w and FLAIR MRI with other unsupervised deep models including Wasserstein autoencoders as well as with intermediate fusion strategies based on multiple kernel learning. Finally, we put our results in perspective with those achieved by the most recent state of the art methods.

The paper is organized as follows. We start by reviewing the existing research in related domains before describing the components of our automated pipeline

in Section 3. This includes the presentation of our unsupervised representation learning model based on a regularized siamese network. Section 4 describes the different experiments conducted to evaluate the performance of our brain anomaly detection models and compare it with alternate deep unsupervised architectures. Finally, Section 5 illustrates the obtained results, followed by a discussion in Section 6.

2. Related work

2.1. Automated subtle brain lesion detection with deep supervised or weakly supervised architectures

The recent advances in deep learning and promising performances achieved in the domain of medical image segmentation (Litjens et al. (2017); Kamnitsas et al. (2017)) motivated new studies, tackling the detection of subtle brain pathologies as a supervised segmentation task, e.g. for MS lesion segmentation (Hashemi et al. (2019), Valverde et al. (2017), Brosch et al. (2016), Havaei et al. (2016)), for ICAC (Bortsova et al. (2017)). Some recent studies introduced specific losses accounting for the class imbalance (Hashemi et al. (2019); Sudre et al. (2017)). Although these deep architectures achieve impressive results compared to the state-of-the art performance, even for the segmentation of small lesions, they still require large voxel-wise annotated data sets for training. Moreover, as far as we know, such supervised architectures are tailored to segmentation of lesions that can be visually detected on the image; their performance on barely visible lesions has not been reported yet.

Some authors recently proposed to formulate segmentation tasks in semi- or weakly-supervised deep settings which allows to reduce or even bypass the need of voxel-level annotated data sets (Cheplygina et al. (2018)). A few attempts have been made to apply these methods to the problem of subtle lesion detection. Baur et al. (2017) introduced a framework for MS lesion segmentation in multi-parametric MRI coupling a standard U-Net architecture (Ronneberger et al. (2015)), trained on labeled data, with manifold embedding accounting for unlabeled data. This model shows promising performance, except for the detection of very small lesions. In Dubost et al. (2017), the authors exploit weak labels (the number of lesions in a scan) in a U-Net like architecture to segment SVD lesions in the basal ganglia based on PD-weighted MRI scans. This framework allows achieving a sensitivity 20% higher than the state-of-the art methods. The method, however, requires to input the number of lesions, thus assuming that the detection task can be performed easily. Such methods are indeed very promising to perform fastidious segmentation tasks with weak supervision. The existing methods, however, have not been designed or optimally tuned to perform challenging detection tasks.

2.2. Deep unsupervised anomaly detection problem

Another recent tendency specifically casts the lesion detection problem as an anomaly detection task. Anomaly detection, also referred to as outlier de-

tection, consists in identifying the observations that seem to be drawn from a different distribution than the one generating the normal examples. Over the recent years, the challenging topic of outlier detection has been studied extensively in different application domains Chandola et al. (2009) and has also been recently addressed from the perspective of deep learning (Kiran et al. (2018)).

One category of deep anomaly detection methods is based on projecting the data samples to a low dimensional manifold and then mapping them back to the original space through a *reconstruction* of the original data points. The reconstruction error is later used to distinguish anomalies. In different computer vision studies, the reconstruction was obtained via various deep architectures such as autoencoders (AE) in Munawar et al. (2017a), variational autoencoder (VAE) in An and Cho (2015), long short memory networks (LSTM) in Munawar et al. (2017b) or generative adversarial networks (GAN) in Zenati et al. (2018). In the medical imaging domain, Schlegl et al. (2017) were the first to propose a GAN-based architecture, trained on normal samples only, associated to an anomalous score function composed of the GAN reconstruction and discrimination losses. This architecture was applied to the detection of fluid in high resolution clinical optical coherence tomography of the retina. Other architectures such as autoencoders (Pawlowski et al. (2018)) or adversarial autoencoders (AAE) (Chen and Konukoglu (2018)) were adapted to the brain tumor segmentation problem.

Another category of anomaly detection methods seeks to learn a discriminative boundary around the *normal* training instances (Chandola et al. (2009)). Any test instance that does not fall within the learnt boundary is declared anomalous. From the deep learning perspective, this amounts to first learning latent representations of normal samples with a deep unsupervised network, similar tot the first category of anomaly detection methods cited above, and then feeding the learned representations to a one-class classification algorithm in order to estimate the boundaries of the normal examples, as proposed by Erfani et al. (2016). As far as we know, we are the first to build on this approach in the medical image analysis domain and for the challenging task of subtle epilepsy detection (Alaverdyan et al. (2018a,b)).

2.3. State-of-the-art automated detection systems for epilepsy lesion detection

Medically refractory epilepsy is often associated with malformations of cortical development (MCD) encompassing a large diversity of lesion types (Barkovich et al. (2012)). MCD include focal cortical dysplasia (FCD) categorized as histological subtypes I, II and III as well as focal, band-shaped or subependymal subcortical heterotopia characterized by the presence of neurons located deep in the white matter. These lesions may also be associated to gliosis and hippocampus sclerosis. Recent retrospective studies involving surgical epilepsy patients indicate that up to 33% with typical FCD type II lesions and 87% with FCD type I lesions go undetected during routine MRI exams (Bernasconi and

Bernasconi (2015)). Similarly, subtle heterotopia may only become apparent after MRI post-processing (Huppertz et al. (2005)). The success rate of surgery is around 70% (Wiebe et al. (2001); Keller et al. (2007); Bien et al. (2012)) but is significantly lower in MRI-negative patients than for MRI-positive patients (Alarcon et al. (2006); Bell et al. (2009); Bien et al. (2009)).

Over the recent years, various automated detection systems have been proposed for the challenging task of epilepsy lesion detection (Kini et al. (2016)). The vast majority of those systems specifically target FCD type II lesions and address this problem from a supervised machine learning perspective combining standard classification algorithms (e;g SVM, decision trees) with manually engineered features that have been shown to correlate with the appearance of these lesions on MR scans, including features derived from surface based morphometry (SBM) (Hong et al. (2014a); Ahmed et al. (2015); Gill et al. (2017)). More recently, Gill et al. (2018) proposed to harness deep learning by designing an architecture composed of two convolutional neural networks trained on T1w and FLAIR patches extracted from the gray matter area in the brain. They demonstrated an improved detection performance compared to the method proposed in Gill et al. (2017) combining a number of SBM, intensity and gradient features with an ensemble of RUSBoosted decision trees. All these approaches require a delineation of lesional zones which is not accurate in MRI-negative patients. For these patients, the ground truth comes from the post-surgical scans containing the resected zone which also includes non-lesional voxels. In Ahmed et al. (2015), the authors showed that manually reducing the resection masks for MRI-negative patients to correct the label noise resulted in a detection rate of 58% while more "generous" annotations achieved only 12%.

To bypass the difficulty of obtaining ground truth annotations, other approaches propose to compare patients to a cohort of normal subjects in order to discriminate the lesions (Thesen et al. (2011); Srivastava et al. (2005); Bruggemann et al. (2007)). Voxel-based morphometry (VBM) analysis builds on this approach by computing mass univariate general linear models (GLM) at the voxel level based on a feature map encoding typical FCD lesion patterns such as cortical thickening and blurring of the GM/WM interface Srivastava et al. (2005); Bruggemann et al. (2007). In another unsupervised setting, Ahmed et al. (2016) evaluated the performance of hierarchical conditional random fields on one among four SBM features or their combination, focusing on cryptogenic epilepsy. In El Azami et al. (2016), we extracted textural features (Huppertz et al. (2005); Wagner et al. (2011)) to learn a one class SVM (oc-SVM) model per voxel and later identify the lesions as the clusters of voxels with the most negative scores output by oc-SVMs. The reasoning behind the choice of building location specific (e.g. voxel based) classification models is that it should enable detecting subtle pattern variations induced by the presence of the lesion, unlike global models pooling all voxels.

Limitations of the current computational models for epilepsy detection have been recently pointed out in survey by Kini et al. (2016). In particular, the au-

thors emphasize the need to acknowledge all types of epilepsy lesions, and recommand to circumvent the lack of massive databases representing the variability of all pathological cases by considering the development of neurocomputational models, first, from the perspective of unsupervised outlier detection and, second, by combining complementary information provided by multimodality imaging to gain in specificity.

3. Method

In this work, we build on the anomaly detection frameworks proposed by Erfani et al. (2016) and El Azami et al. (2016) by combining unsupervised feature learning and robust one class classification with oc-SVM for anomaly detection. In the following, we present the general pipeline of our anomaly detection system, then detail the different elements of this pipeline, especially emphasizing the novel architecture of unsupervised representation model based on siamese networks.

3.1. ***General pipeline***

In the proposed general anomaly detection system, each voxel is characterized by a latent representation vector extracted from an unsupervised deep network. For each voxel in the brain, its normality is modeled with a oc-SVM classifier in the latent representation space. Eventually, for an unseen patient, abnormalities can be found as local neighborhoods of voxels found anomalous by the corresponding oc-SVM models. The general pipeline is illustrated on figure 2. It consists of two major steps

- In the first step, we extract image patches of all the available volumes of the healthy control MRI dataset and learn a latent representation of each patch with a deep unsupervised network. Once this step is performed, the central voxel of each patch extracted from a brain volume will be associated to a latent representation yielded by this deep network.
- In the second step, we build one oc-SVM model per voxel in the latent representation space learned at the previous step. Each voxel is associated with a oc-SVM classifier, hence the number of classifiers is equal to the number of voxels in the volume of interest. For a given voxel v_i, the associated oc-SVM classifier C_i is trained on the matrix composed of the representations of the patches of all the subjects from the healthy control dataset centered at v_i.

For a new patient, each voxel v_i is first assigned a latent representation based on the deep unsupervised model; this latent representation is then matched against the corresponding classifier C_i and is assigned the signed score output by the classifier. This yields a *distance map* D_p for the given patient.
Note that the pipeline represented on fig. 2 takes one imaging modality, 3D MRI T1-W, at input, for clarity purpose. Our model actually combines T1-w an FLAIR images as input channels of the deep unsupervised architecture. Each step of the system will be explained in details in the following sections.

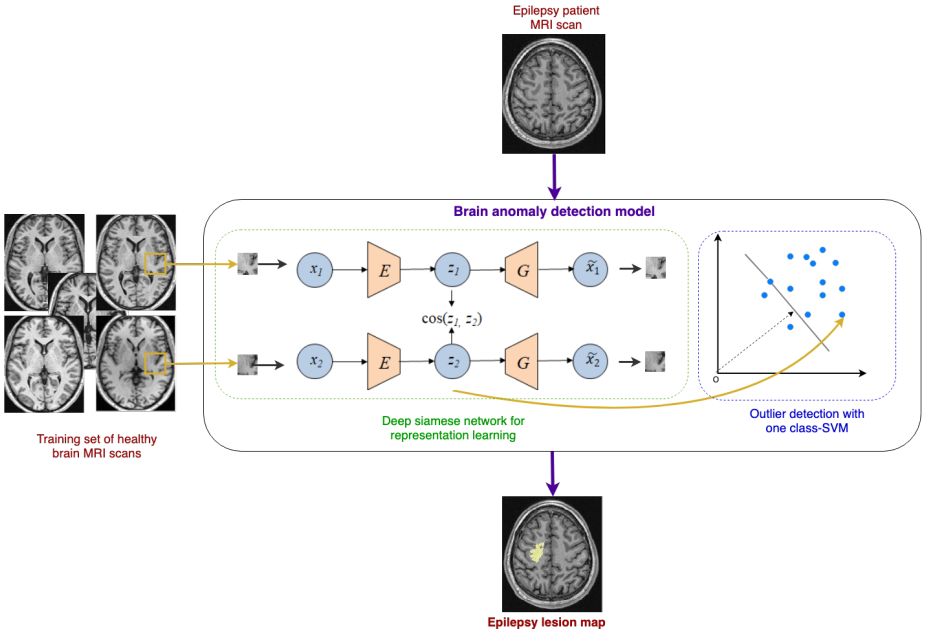

Figure 2: General pipeline of our brain anomaly detection model. The training is shown in a yellow path, testing in a purple path. This pipeline illustrates the processing of monomodal T1-w MRI patches for clarity purpose. In the multimodal setting, patches of the T1-w and FLAIR imaging modalities extracted at the same spatial location in the brain are combined as channels at the input of the deep unsupervised representation model.

3.2. ***Regularized siamese neural network for representation learning***

3.2.1. Rationale

Our objective is to map the original patches to a representation space that captures the global semantic information from the image whatever the brain region and where the patches belonging to different *healthy* subjects and centered at the same spatial localization are *close* with respect to a chosen metric. Autoencoders (Hinton and Zemel (1994)) and their variations have been shown efficient for feature extraction in various contexts and applications, while siamese networks (Bromley et al. (1993); Chopra et al. (2005)) perform well in learning representation space where the distance between similar and dissimilar pairs of instances can be controlled. In this study, we leverage both autoencoders and siamese network architectures in a unified framework to match our objective.

We hypothesize that such a siamese network composed of identical autoencoders as subnetworks should better capture the shared fine patterns of each patch by encoding the similarity constraint in the loss function and thus reinforce the compacity of the patch distribution in the latent space with regards to standard autoencoders. The latent representations of healthy patches centered at the same spatial location will thus be driven closer while that of an anomalous

patch centered at the same location in the brain, which has never been processed by the network, will lay further in the representation space.

We also propose a novel architecture that bypasses dissimilar pairs unlike standard siamese architecture and is thus trained on similar pairs only. This choice is motivated by the fact that the notion of *dissimilar* pairs can not be objectively defined in our context. The inclusion of dissimilar pairs, however, is not an absolute prerequisite of the good performance of siamese architecture. Zheng et al. (2015), indeed, showed that siamese networks learn better latent representations when trained on similar pairs only for an application to face verification. The intuition behind this result is that learning dissimilarities is challenging and not well conditioned as opposed to learning similarities. Thus, depending on the data, the two terms of the objective function related to dissimilarity and similarity may be partly contradictory and inhibit the learning of similarities during the training process.

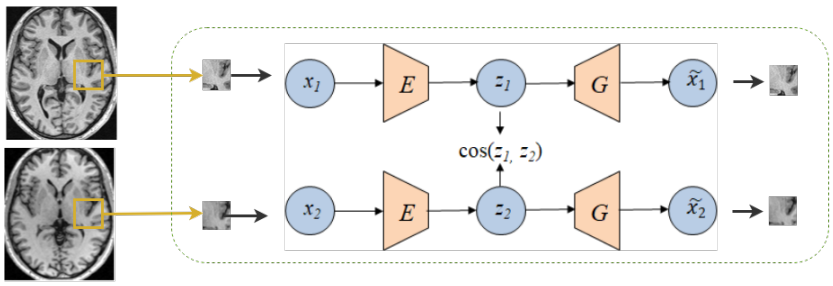

Figure 3: Siamese neural network composed of stacked convolutional autoencoders as subnetworks. The input consists of a pair of patches of 2 different subjects centered at the same spatial localization in the brain. The middle-layer representation is denoted by z.

3.2.2. Network Architecture and loss function

The proposed architecture is illustrated on figure 3. Our regularized siamese neural network (rSN) consists of two identical (same architecture, shared parameters) subnetworks - stacked convolutional autoencoders (CAE) with K hidden layers and a cost module. The input $\mathbf{x}$ of a CAE is first encoded to a middle-layer representation by a series of convolutional and max-pooling operations and later decoded with a series of deconvolutions and up-poolings to produce a reconstruction $\hat{\mathbf{x}}$ of the input. A convolutional layer l is composed of N_l kernels and biases and can be expressed as

$$\mathbf{H}_l^m = f(\mathbf{W}_{l-1}^m * \mathbf{H}_{l-1} + b_{l-1}^m)$$

where $\mathbf{H}_l^m$ is the m-th feature map of layer l, $\mathbf{W}_l^m$ is the kernel matrix associated with $\mathbf{H}_l^m$ and b_l^m is its bias, f is an activation function (usually non-linear). $*$ denotes the convolution operation.
The siamese network receives a pair of patches $(\mathbf{x_1}, \mathbf{x_2})$ at input, then each

patch is propagated through the corresponding subnetwork yielding representations $\mathbf{z_t} \in \mathcal{Z}, t = (1, 2)$ in the middle layer which are then passed to the loss function L_{rSN} below.
Our loss function is designed to maximize the cosine similarity between $\mathbf{z_1}$ and $\mathbf{z_2}$. To compensate for the lack of dissimilar pairs, we propose to add a regularizing term consisting of the mean squared error between the input patches and their reconstructions output by the subnetworks. Without a proper regularization term, the loss function could be driven to 0 by mapping all the patches to a constant value. The proposed loss function for a single pair hence is:

$$L_{rSN}(\mathbf{x_1}, \mathbf{x_2}; \Theta) = \sum_{t=1}^{2} ||\mathbf{x_t} - \hat{\mathbf{x}}_\mathbf{t}||_2^2 - \alpha \cdot cos(\mathbf{z_1}, \mathbf{z_2}) \tag{1}$$

where $\hat{\mathbf{x}}_\mathbf{t}$ is the reconstructed output of subnetwork t of the patch $\mathbf{x_t}$ while $\mathbf{z_t}$ is its (vectorized) representation in the middle layer and α is an hyperparameter that controls the tradeoff between the two terms. Θ represents the parameter set.

3.3. ***Voxel-level outlier detection with oc-SVM***

A one class support vector machine (oc-SVM) (Schölkopf et al. (2001)) is an outlier detection method based on the SVM algorithm assigning labels $y_i \in \{-1, 1\}$ to two distinct classes of objects, based on n training samples $(\mathbf{z_i}, y_i) \in X$ from the *negative class only*. The training examples are first mapped to a higher dimensional space via a feature map ϕ associated with a kernel K such that $K(\mathbf{z}_i, \mathbf{z}_j) = \langle \phi(\mathbf{z}_i), \phi(\mathbf{z}_j) \rangle$. The corresponding optimization problem is the following:

$$\begin{aligned} \min_{\mathbf{w}, \rho, \xi_i} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \mathbf{w} \cdot \phi(\mathbf{z_i}) \geq \rho - \xi_i, \xi_i \geq 0, i \in [1, n] \end{aligned} \tag{2}$$

where ξ_i-s are slack variables relaxing the inequality constraints so as to account for the non-separable classes, $\mathbf{w}$ and ρ define the separating hyperplane, ν is a parameter that sets a boundary to the fraction of allowed outliers. The decision function, for an example $\mathbf{z}$, is $\mathbf{w} \cdot \phi(\mathbf{z}) - \rho$. This decision function contributes to the signed score output by a oc-SVM model. In a typical scenario, examples with negatives scores would be considered outliers.

In the scenario depicted on figure 2, a given voxel v_i is associated with a oc-SVM classifier C_i trained on the matrix $M_i = [\mathbf{z_{i1}}, ..., \mathbf{z_{in}}]$ where $\mathbf{z_{ij}}$ is the feature vector corresponding to the patch centered at v_i of subject j and n is the number of subjects. During the test phase, each voxel v_i of a given patient p is matched against the corresponding classifier C_i and is assigned the signed score output by the classifier. This yields a *distance map* D_p for the given patient.

3.4. ***Post-processing***

For a given patient p, the distance map D_p outputted at the previous step is then post-processed to obtain the final anomaly detection map. A 3-step post-processing is proposed as follows.

The first step consists in normalizing the distance maps with respect to the intra-subject spatial variability. For that purpose, the distance maps of the control subjects are computed by performing a k-fold evaluation of the training set (i.e. for each fold of normal subjects, the distance maps are obtained with oc-SVMs trained on the remaining subjects). These maps are used to estimate the standard deviation of the *normal subjects' distance* distribution at voxel-level. For a given patient p, a new map $\acute{D}_p$ is computed by a voxel-wise division of the output distance map D_p over the estimated standard deviations. The final distance map F_p is then derived by averaging D_p and $\acute{D}_p$ i.e.

$$F_p = \frac{1}{2}\left(\frac{D_p}{max(abs(D_p))} + \frac{\acute{D}_p}{max(abs(\acute{D}_p))}\right). \tag{3}$$

The reason behind the additional term is that some zones in the brain have more intra-subject variability than others and therefore are more likely to be considered as anomalies. By weighing them by the standard deviation, the score maps account for this effect.

The second step consists in thresholding the F_p map to produce a *cluster map*. To this end, all the voxel score values of F_p are pooled together into a histogram which is then approximated by a non-parametric distribution using a kernel density estimator (Bowman and Azzalini (1997)). The approximated patient distance score distribution is then thresholded at some pre-chosen value and a 26-connectivity rule is applied to identify the connected components. These components are referred to as *clusters*. By varying the threshold, the number of clusters can be controlled according to a clinician's needs. An example of F_p output score map, as well as the cluster maps obtained with different thresholds are shown on fig. 4. Clusters smaller than a certain size may also be discarded.

The third step consists in ranking the detected clusters. We use the following *ranking criterion* to assign a rank to a cluster c_i, inspired from Ahmed et al. (2016)

$$rank(\mathbf{c_i}) \sim \omega * \frac{score(\mathbf{c_i})}{min_j score(\mathbf{c_j})} + (1-\omega) * \frac{size(\mathbf{c_i})}{max_j size(\mathbf{c_j})} \tag{4}$$

where $score(c_i)$ is the average of the voxel scores in the cluster and $size(c_i)$ is the number of voxels in the cluster and ω is a parameter weighing the relative contribution of the cluster size and score. Such a ranking favors large clusters with the most negative average score. Using this ranking, we keep the top n detections and discard the rest.

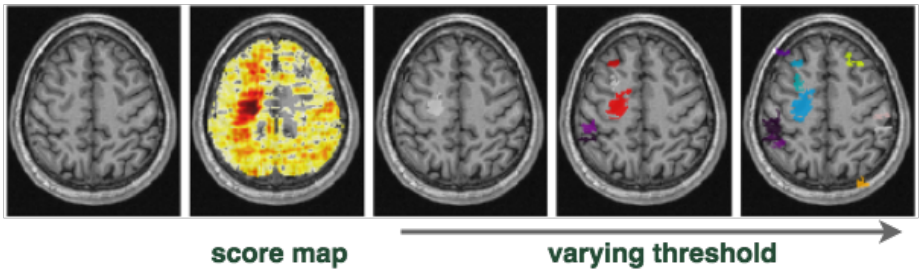

Figure 4: An example of post-processing on a patient. The first column shows the original slice centered at the lesion, the second column corresponds to the normalized distance map F_p and the last three columns are obtained by thresholding F_p at three different values and identifying the connected components as detected clusters.

4. Experimental settings

4.1. Dataset description and pre-processing

The study was approved by our institutional review board with approval numbers 2012-A00516-37 and 2014-019 B and a written consent was obtained for all participants.

4.1.1. Study group

The data set considered in this study consists in a training set of healthy subjects and a test set of epilepsy patients.

Patient group: The test group consists of 21 patients who had been admitted to the Neurological Hospital of Lyon and diagnosed with medically intractable epilepsy. The age of the patients varies between 17 and 47 years, with a median of 29. As a part of the pre-surgical evaluation, they all had T1-weighted (T1w) and FLAIR MRI sequences as detailed below. Additionally, the patients underwent intracranial EEG exams in order to localize the origin of seizures.

Healthy control group: The training data set consists of 75 healthy individuals aged between 20 and 66 years. All the subjects had similar T1w and FLAIR MRI sequences as the patient group. Sixty-nine of these control subjects were used as the training datasets while the remaining 6 subjects were used as a validation dataset.

4.1.2. MRI acquisition and pre-processing

All the healthy controls and patients had 3D anatomical T1w brain MRI sequences (TR/TE 2400/3.55; 160 sagittal slices of 192 x 192 1.2mm cubic voxels) and FLAIR MRI sequences (176 slices of 196 x 256 1.2mm cubic voxels) on a 1.5 T Sonata scanner (Siemens Healthcare, Erlangen, Germany).
The pre-processing was done using the SPM8 software (Ashburner (2009)). The T1-weighted MRI volumes were first processed with the unified segmentation algorithm (UniSeg) (Ashburner and Friston (2005)) implemented in SPM (using the default parameter values) that performs tissue segmentation (white/grey

matter, cerebrospinal fluid), correction for magnetic field inhomogeneities and spatial normalization. All the 3D MR volumes were normalized to the standard brain template of the Montreal Neurological Institute (MNI) (Mazziotta et al. (2001)) with a voxel size of 1 x 1 x 1 mm. Further, FLAIR images were rigidly co-registered to the individual T1w MR images. Next, the transformation from the subjects' native space to the MNI space, produced by the UniSeg algorithm, was applied on the co-registered FLAIR images in order to normalize them to the MNI space as well.

We excluded the brain regions (the cerebellum and brain stem) that are not susceptible to epilepsy using a masking image in the MNI space derived from the Hammersmith maximum probability atlas described in Hammers et al. (2003). After the elimination of the corresponding voxels, the number of remaining voxels adds up to around 1.5 million per volume. We removed top 1% intensities and scaled the images between 0 and 1 at image level before feeding the patches to the deep architectures.

4.1.3. Patient lesion annotations

The patient group was selected in collaboration with the expert neurologists of the Lyon Neurological University Hospital to sample highly difficult detection cases so as to challenge the performance of the proposed automated detection system. Most of them had a resective surgery and became seizure free at most 6 months after the surgery. A few patients had successful thermocoagulation instead (patients A^-, G^+). The positive outcome of resective surgery or thermocoagulation is considered as the ground truth to establish the epileptogenic lesion localization. The ground truth annotations used in the performance evaluation were thus obtained by outlining the visible zones of the MRI-positive patients (D^+, G^+ and R^+), and by combining the information from the intracranial EEG, post-surgical or post-thermocoagulation MR images and the resected zones for *MRI-negative* patients. Note that the MRI-negative patients considered in this study remained MRI-negative after a retrospective inspection of the T1w and FLAIR scans. For patients undergoing surgery, the histopathological analysis of the resected tissue indicated FCD type II in 3 patients (D^+, G^+ and Q^-) and FCD type III in 3 patients (C^-, P^- and S^-). The histopathological analysis was inconclusive for the remaining patients, which is in accordance with the statistics reported in Bernasconi et al. (2011) and underlines our objective to detect lesions with unknown signatures. A full description of the lesion types and localizations is provided in Table 1.

4.2. Implementation details of the brain anomaly detection model

The general architecture of the brain anomaly detection model depicted on figure 2 was adapted to the specific application of epilepsy lesion detection in multiparametric T1-w and FLAIR MRI. Hereafter, we provide the detailed description of the different elements of the pipeline. The entire pipeline was implemented in Python, with Theano/Keras libraries for the representation learning stage and a LibSVM wrapper in scikit-learn (Pedregosa et al. (2011)) for one class SVM.

4.2.1. Unsupervised representation learning with rSN.

The proposed rSN consists of two identical stacked convolutional autoencoders as depicted on figure 3 and on the top left part of figure 6. The specific architectures of the encoder (E) and decoder (or generator) (G) elements are shown on figure 5. The input of the encoder consists of the patches of each modality combined as channels. For each couple of patches of a given subject, a random 'similar pair' is selected among the identically located patches of other subjects yielding in total around 3.5 million pairs of patches for training. The input patch size was set to 15x15 after a number of tested configurations and is justified by the subtle nature and size of epilepsy lesions. Indeed, larger patch sizes (e.g. 30x30) were not successful at detecting subtle lesions. The encoder and decoder are composed of two convolutional layers with kernel size of 3x3 and a stride of 1. A max-pooling operation is performed only at the first and last layer of the encoding and decoding part respectively. As shown on figure 5, the middle layer has 16 feature maps of 2x2 which, when flattened, yields a 64-dimensional vector. This dimension compromises the need to capture subtle patterns at the patch level while conforming to the size of the training database (69 subjects), to avoid over-fitting of the oc-SVM models. We used ReLU activation function in all the layers except the last one where the sigmoid function is used. The α parameter in loss (1) was set to 0 during the first 10 epochs, then grew linearly for 15 epochs until it reached some α_{max} value and then plateaued for 5 more epochs. The Adam optimizer was used with the learning rate set to 0.001.

4.2.2. outlier detection with oc-SVM

We used oc-SVM classifiers with RBF kernel. The γ kernel parameter was derived for each voxel v_i individually by setting it to the median of the standardized euclidean pairwise distances of the corresponding matrix M_i (see section 3.3) as in Caputo et al. (2002). Varying the parameter ν corresponding to the upper bound on the fraction of permitted outliers (see eq. 2) did not significantly impact the results; the fraction of the outliers is indeed controlled in the post-processing step with the threshold applied on the distance map F_p. Thus, ν was set to 0.03 for all voxels and all scenarios.

4.2.3. Post-processing

In the first post-processing step, we performed a 10-fold cross validation to compute the output maps of the normal population that served to derive the output map $\acute{D}$ in eq. 3. In the second step of the post-processing, we empirically set the threshold of the output score maps to the value that resulted in at most 10 clusters. We indeed observed among the output maps of the normal population that larger values typically produce a small number of very large clusters. The minimal cluster size was set to 82 voxels corresponding to the expected cluster size calculated with the SPM analysis of the T1w MRI data (see supplementary file). The size of the majority of the detected clusters varies between 500 and 3000 voxels, this threshold therefore does not affect the performance in any significant way. In the third post-processing step, the parameter ω weighing

the relative contribution of the cluster size and score was set to 0.5 to equally account for the cluster average score and size.

4.3. Experiments

4.3.1. Performance of our brain anomaly detection model

The brain anomaly detection model of figure was trained on the training dataset of the 69 healthy subjects as described in section 4.1. The validation dataset was used to define convergence of the rSN model and fit the parameter set Θ of the loss term in eq. 1. The hyperparameter α that controls the tradeoff between the two terms of the rSN loss term was varied among the values : 0.25 and 0.5. Detection performance of the model was evaluated on the series of 21 patients based on the metric defined in section 4.4.

4.3.2. Comparison with intermediate fusion strategy of T1-w and FLAIR modalities

We explore an alternative strategy to the early fusion of T1w and FLAIR MRI modalities as input channels of the siamese network. This consists in performing intermediate fusion by training individual networks for each modality and combining the learned representations with a multiple kernel algorithm. We propose to use the slimSimpleMKL algorithm (Loosli and Aboubacar (2017)) to leverage the original formulation of the simpleMKL algorithm (Rakotomamonjy et al. (2008)) that extends the oc-SVM formalism to fit the multiple kernel paradigm (Bach et al. (2004); Sonnenburg et al. (2006)). By controlling the number of support vectors with a tradeoff parameter λ, tight normality bounds can be achieved with SlimSimpleMKL which, in turn, can lead to an improved performance of the outlier detection model. Two separate architectures referred to as *T1 rSN* and *FLAIR rSN* were thus trained with the T1-w and FLAIR healthy subject dataset, respectively. Both architectures used similar encoder and decoder architectures as those depicted on figure 5 except that the number of input channel was set to 1. The intermediate fusion with slimSimpleMKL was performed in Matlab, using the implementation provided by Loosli and Aboubacar (2017).

4.3.3. Comparison with other deep unsupervised representation models

We compare the proposed siamese model with two alternative unsupervised representation models, stacked convolutional autoencoders (CAE) and Wasserstein autoencoders (WAE) (Tolstikhin et al. (2017)). Figure 6 illustrates the general setup. In our comparison, the encoder E and decoder G of CAE and WAE have the same architecture as the ones of the rSNN subnetworks shown on figure 5.

The stacked convolutional autoencoder (CAE) serves as a baseline performance model to evaluate the potential added value of the proposed rSN model and its ability to capture a finer representation of the normal brain based on the proposed loss L_{rSN}.

Wasserstein autoencoders have been recently introduced as generative models combining the best properties of Wasserstein GANs and Variational autoencoders (Tolstikhin et al. (2017)). They consist of three components: an encoder E, a decoder G and an adversary network D that tries to distinguish the prior distribution of the latent code P_Z from the latent distribution Q_Z produced by the encoder. The resulting loss function can be expressed as

$$L_{WAE}(X;\Theta_{WAE}) = \frac{1}{N}\sum_{i=1}^{N} c(\mathbf{x_i},\hat{\mathbf{x}}_\mathbf{i}) + \beta \cdot D_Z(P_z,Q_z) \tag{5}$$

where D_Z measures the discrepancy between a given distribution P_z and Q_z for the dataset $X = \{\mathbf{x_i}\}_{1,..,N}$ and c measures the reconstruction error. β is a coefficient that controls the tradeoff between the two terms and Θ_{WAE} denotes the parameter set. The generic form of the WAE loss allows different reconstruction error functions and regularizers. We used the standard reconstruction error $c(\mathbf{x_i},\hat{\mathbf{x}}_\mathbf{i}) = ||\mathbf{x} - \hat{\mathbf{x}}_\mathbf{i}||_2^2$ and the Jenssen-Shanon divergence as D_Z, estimated with a discriminator.
As for the proposed rSN model, we hypothesize that this architecture should capture a finer representation than the standard CAE by enforcing the learned representations to follow the prior distribution. In our comparison, the WAE discriminator D consists of four fully connected layers of dimension 128, 128, 64 and 1 with LeakyReLU as activation (with scale 0.02 for negative input values). We varied the parameter β in the L_{WAE} loss (eq. 5) among the following values - 1,5,10,20 and 100.

4.4. ***Performance evaluation***

We chose typical metrics applied in lesion detection tasks to evaluate the performance of our brain anomaly detection model. We define the number of *true positive (TP)* and *false positive (FP)* detections at cluster-level i.e. we consider that a cluster is true positive when it overlaps with the ground truth annotation and false positive (FP) otherwise. As such, we quantify the *sensitivity* of the system as the percentage of the patients whose lesions were correctly detected. We couple the system sensitivity with the average number of FPs per patient to plot a fROC curve (Bunch et al. (1978)). We also perform a qualitative visual analysis of the detected cluster maps and provide a quantitative evaluation at the patient level.

5. Results

We first evaluate the performance of the proposed brain anomaly detection model depicted on figure 2. We then compare its performance to those achieved with variants of the same global architecture. The brain anomaly detection model is trained on 69 healthy subjects and is evaluated on 21 patients with confirmed epilepsy lesions described in section 4.1.

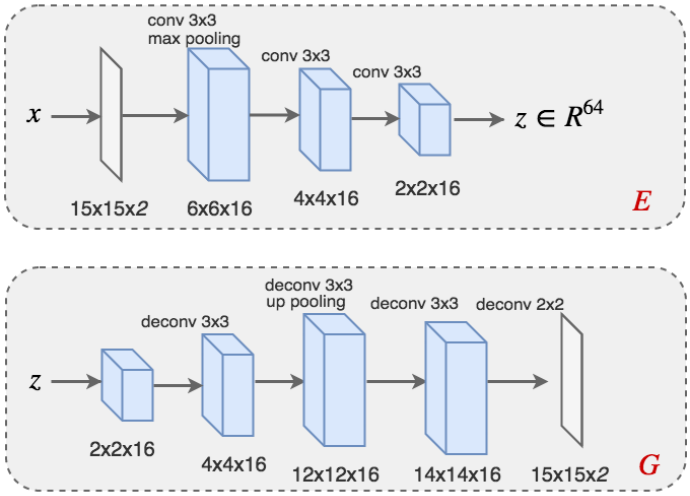

Figure 5: Encoder E and generator G used in the convolutional autoencoder, Wasserstein autoencoder and regularized siamese network multichannel architectures.

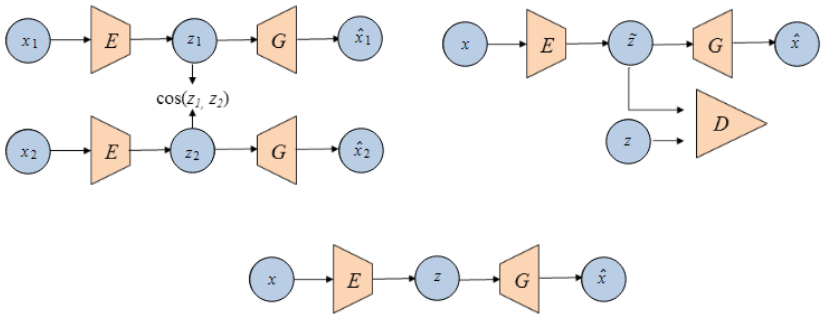

Figure 6: Left : Siamese neural network composed of stacked convolutional autoencoders as subnetworks (CAE). Center: Convolutional Autoencoder (CAE) composed of an encoder E and decoder D. Right: Wasserstein autoencoder (WAE) composed of an encoder E, a decoder G and an adversary discriminator D. For all three models, the encoder and decoder have the same architecture described on figure 5.

5.1. Performance of our brain anomaly detection model

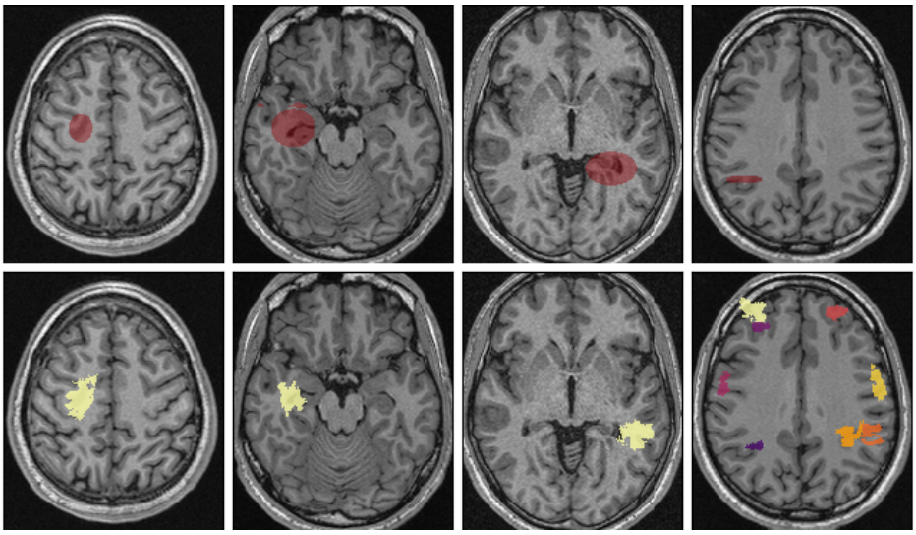

Figure 7: Output detection maps for patients D^+, N^-, B^- and E^- respectively ($^+$ stands for MRI-positive patients, $^-$ for MRI-negative patients). Top row: Transverse slices centered at the lesion locations (highlighted in purple). Bottom row: Maximum intensity projections (MIP) of the cluster maps overlaid on the MRI transverse slices. The maps show the top 1, top 1, top 1 and top 8 clusters, respectively, corresponding to the rank of the true positive detections. The color scale indicates the ranking of the detected clusters, ranging from light yellow (for the most suspicious ones) to dark brown (for the lowest ranked ones).

In this section, we report the performance of our brain anomaly detection model based on deep unsupervised representation extracted from a multichannel T1/FLAIR regularized siamese network and associated to a oc-SVM classification and ad-hoc post-processing.

Figure 8a reports the fROC curves computed on the series of 21 patients considering two values of α[1] in loss L_{rSN} (see eq (1)) - 0.25 and 0.5. Table 1 reports performance at patient level for the best configuration corresponding to $\alpha = 0.5$. The system achieves an overall sensitivity of 62% (13/21) with a mean rank of the TP detections of 3.6 on the 21 patients. The corresponding performance on the 18 MRI^- patients is 61% (11/18) with a mean TP rank of 4.1. The system fails to detect the lesions of unknown type in 5 patients and 2 patients with hippocampal sclerosis.

Figure 7 visualizes example cluster maps output by the brain anomaly detection model, while table 1 reports its performance at patient level. All the detected clusters, with a rank higher or equal to that of the true positive detection, are shown on the figure. The highest ranked cluster, corresponding to the most suspicious detection, is shown in light yellow while the lowest ranked clusters are

[1]Hereafter, a rSN with $\alpha = m$ is a shorthand for $\alpha_{max} = m$.

depicted in darker colors ranging from light orange to dark brown. The symbol ✓ in table 1 indicates that the lesion was correctly detected by the system while the rank of the true detection is given inside parentheses.
The first column on figure 7 corresponds to the MRI-positive patient D^+ where the highest ranked detected cluster matches the lesion localization. Table 1 confirms that the two MRI-positive patients (D^+ and G^+) with confirmed FCD type II lesions were detected with the highest rank. Our system allows detecting very small and subtle lesions with unknown signatures in patients N^- and B^-. These lesions are detected with the highest rank as indicated in table 1. Patient E^- illustrates a very difficult case of a small lesion of an unknown type. In such difficult cases, the anomalies accompanying the epilepsy lesions are so subtle that they appear together with other anatomical outliers. For this patient, the lesion was ranked $8th$.

For comparison, we also report the performance of the two anomaly detection models based on rSN representations learned separately on T1w and FLAIR MR images, respectively. These two models are referred to as *T1 rSN* and *FLAIR rSN* models in the following. Fig. 8a shows the corresponding fROC curves for two values of the α coefficient. Table 1 summarizes their performance at patient level. Results reported on figure 8a and table 1 clearly demonstrate that our model outperforms those based on features learned from one imaging modality. The best configuration of *T1 rSN* model (corresponding to $\alpha = 0.25$) allowed to achieve 39% sensitivity (7/18) with a mean TP rank of 3.2 ± 2.5, for MRI-negative lesions. The best performance of the *FLAIR rSN* model was also reached for $\alpha = 0.25$, leading to 50% sensitivity (9/18) on MRI-negative patients with a mean TP rank of 5.0 ± 3.2. These results speak of the complementary nature of the two modalities. The deep architectures are flexible with respect to accommodating multiple modalities; hence, other modalities could be easily integrated into the architecture.

5.2. Comparison with intermediate fusion strategy of T1-w and FLAIR modalities

Fig. 8b shows comparative performances of our brain anomaly detection model with a variant of this architecture consisting in applying multiple kernel learning with slimSimpleMKL on the features learnt with independent monomodal (T1w or FLAIR) rSN models. The parameter λ controlling the number of support vectors in slimSimpleMKL was varied among the values 0, 0.05, 0.1, 0.25, and 0.5. The intermediate fusion with $\lambda = 0.1, 0.25, 0.5$ results in identical performance and, therefore, the corresponding fROC curves are seen as a single one. From this comparison, it is apparent that even a slight regularization on the number of support vectors with $\lambda = 0.05$ offers a significant improvement over the original SimpleMKL formulation ($\lambda = 0$). The sensitivity jumps from 32% to 52% for the same false positive rate (9 FPs).
As it can be seen on figure 8b, the early fusion with our multichannel T1/FLAIR rSN model outperforms significantly the intermediate fusion strategy with simpleSimpleMKL. Combining modalities in the network training stage is thus likely

Patient	**Lesion location**	**Lesion type**	**T1 rSN** $\alpha = 0.25$	**FLAIR rSN** $\alpha = 0.25$	**T1/FLAIR rSN** $\alpha = 0.5$
Patient A^-	Insula R	Unknown	✓(8)	✗	✗
Patient B^-	Temporal Lobe L	Unknown	✓(1)	✓(2)	✓(1)
Patient C^-	Hippocampus R	FCD type III with HS	✗	✗	✗
Patient D^+	Superior frontal gyrus R	FCD type II	✓(2)	✓(3)	✓(1)
Patient E^-	Inferiolateral remainder of parietal lobe R	Unknown	✗	✓(10)	✓(8)
Patient F^-	Hippocampus L, parahippocampus L	Unknown	✗	✓(3)	✓(9)
Patient G^+	Middle frontal gyrus L	FCD type II	✓(4)	✓(1)	✓(1)
Patient H^-	Superior frontal gyrus R	Unknown	✓(1)	✓(8)	✓(3)
Patient I^-	Hippocampus L, parahippocampus L	Unknown	✗	✗	✗
Patient J^-	Precentral gyrus R	Unknown	✗	✗	✗
Patient K^-	Superior temporal gyrus R	Unknown	✗	✗	✗
Patient L^-	Middle frontal gyrus R	Unknown	✗	✗	✓(1)
Patient M^-	Anterior temporal lobe R	Unknown	✗	✗	✓(4)
Patient N^-	Anterior temporal lobe R Hippocampus R	Unknown	✓(9)	✓(1)	✓(1)
Patient O^-	Middle frontal gyrus L	Unknown	✓(1)	✓(6)	✓(2)
Patient P^-	Hippocampus R	FCD type III with HS	✗	✗	✗
Patient Q^-	Lateral remainder of occipital lobe L	FCD type II	✓(2)	✓(3)	✓(7)
Patient R^+	Orbital gyrus R	Ganglioglioma	✗	✓(6)	✗
Patient S^-	Anterior temporal lobe R Hippocampus R	FCD type IIIa	✗	✓(8)	✓(6)
Patient T^-	Posterior temporal lobe R	Unknown	✗	✗	✗
Patient U^-	Posterior temporal lobe L	Unknown	✓(1)	✓(4)	✓(3)
Overall # of detections			**9**	**12**	**13**

Table 1: Performance of our brain anomaly detection model (T1/FLAIR rSN) and comparison with a similar architecture trained on a single imaging modality (T1 rSN or FLAIR rSN). ✓ denotes a true detection followed by its rank inside parentheses. ✗ denotes no true positive detection meaning that the lesion was not detected among the 10 highest ranked clusters detected by the model.

to better leverage the complementary information contained in T1w and FLAIR MR images while the intermediate fusion, that operates on features learned separately on each modality, may skip those properties.

5.3. Comparison with alternate unsupervised representation learning models

We next compared our brain anomaly detection model with alternate models based on the same architecture but different multichannel unsupervised represention models, namely WAE and CAE, as described in section 4.3.3. Fig. 8c reports the fROC curves for these three models. As mentioned above, the WAE parameter β was varied among 1,5,10 and 20. Again, the performance achieved with the T1/FLAIR rSN features outperforms those achieved with the features extracted with the WAE and CAE models. WAE is shown to perform better than CAE for all but one value of β. The latter confirms our hypothesis that the reconstruction error, when enhanced with a regularization term, fits better to the anomaly detection context. The WAE performance is still inferior to that of rSN which might be due to a limitation of the model itself or the experimental choice of the hyper-parameters. We can see how the performance is affected by the choice of β; the value of 20 is less successful, probably since it prioritizes too much the adversarial term.

6. Discussion

This study presents a novel brain anomaly detection model combining unsupervised latent representation model with one-class classification at the voxel level. We have formulated a regularized siamese network architecture that learns normal brain representations using a set of non-pathological MR volumes. The features learnt with the network do not target a specific pathology but rather allow to capture normal variability from a cohort of healthy subjects. The framework allows integrating multiple modalities and we have shown the performance gain obtained by coupling T1w and FLAIR imaging for the task of detecting subtle epilepsy lesions in MRI-negative patients. To our knowledge, this is the first attempt to extract unsupervised deep latent representation for epilepsy lesion detection. The proposed approach achieves 61% detection rate on the 18 MRI-negative patients, meaning that it detects 61% of the lesions that were not visually detected by the clinicians in these challenging cases, at the same time outperforming Wasserstein and convolutional autoencoder architectures.

As stated in Section 2.3, most current studies target the detection of FCD lesions in T1-w MRI, mainly focusing on MRI^+ exams and do not report the average false positive detection rate per patient but rather compute specificity as the percentage of normal controls in whom no lesion was falsely identified. Table 2 summarizes the main performance of these different models. The vast majority uses manually designed features characterizing cortical malformations based on surface based morphometry (SBM) (Thesen et al. (2011); Hong et al. (2014b);

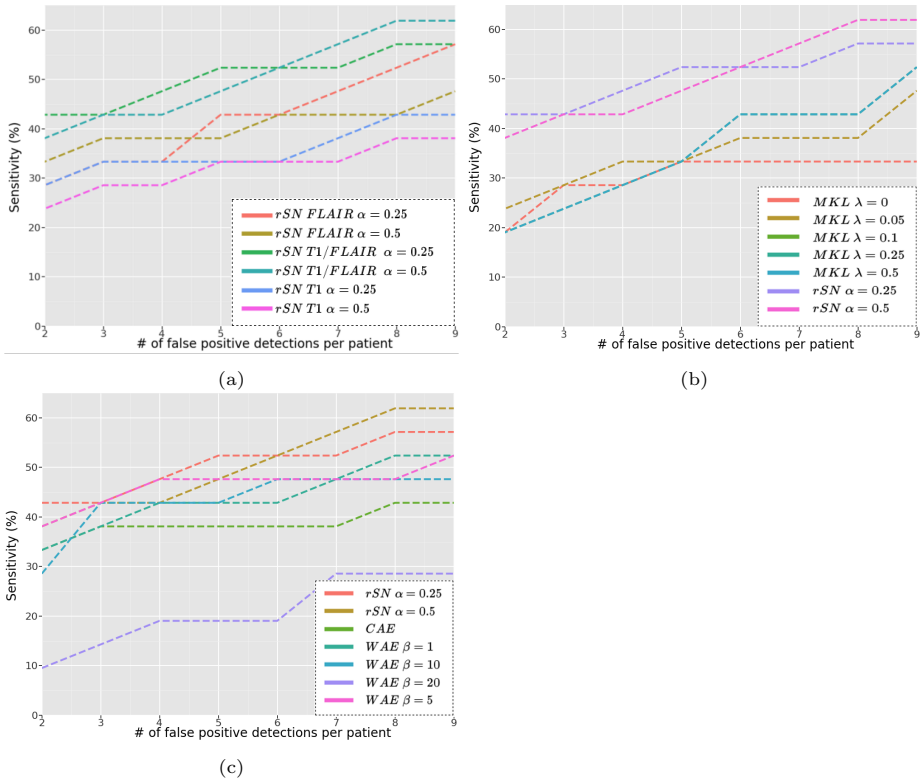

Figure 8: Comparative fROC curves estimated on the 21 patients of the test dataset. x-axis: number of false positive detections per patient, y-axis: sensitivity.
a) Performance of the proposed detection model based on T1/FLAIR rSN representation and comparison with similar architecture trained on a single imaging modality (T1 or FLAIR). T1, FLAIR and T1/FLAIR rSNs are shown for two values of α.
b) Comparison of our model with alternate architecture based on an intermediate fusion strategy of T1 and FLAIR modalities using slimSimpleMKL. Five values of the parameter λ were considered for slimSimpleMKL. fROC curves are superimposed for $\lambda \geq 0.1$.
c) Comparison of our model with alternate architectures based on two different unsupervised representation models, CAE and WAE.

Table 2: State-of-the art performances for the detection of FCD lesions in brain T1-w MRI. First column: Sensitivity and corresponding average FP rate inside parentheses. Note that the false positive rate is estimated on a normal control group. Second column: Size of the patient test dataset.

	Sensitivity (FP)	Nb of Patients
Gill et al. (2018)	0.87-0.90 (4)	102 MRI^+
Gill et al. (2017)	0.83 (4)	41 MRI^+
Tan et al. (2018)	0.82 (26)	28 MRI^+
Jin et al. (2018)	0.74 (90%)	44 MRI^+ and 17 MRI^-
Jin et al. (2018)	0.53 (90%)	17 MRI^-
Hong et al. (2014b)	0.74 (95-100%)	17 MRI^+

Table 3: Performance of our brain anomaly detection model compared with state-of-the art performances in similar experimental condition for the detection of MRI^- lesions. First column: Sensitivity and corresponding average false positive rate inside parentheses. Second column: Size of the patient test dataset. Third column: MRI sequences

	Sensitivity (FP)	Nb of MRI^- Patients	MRI
Ahmed et al. (2016)	0.70 (9)	20	T1
El Azami et al. (2016)	0.70 (4)	10	T1
our implementation of El Azami et al. (2016)	0.52 (9)	18	T1
our model	0.62 (9)	18	T1/FLAIR
our model with T1 rSN	0.39 (9)	18	T1
our model with FLAIR rSN	0.50 (9)	18	FLAIR

Ahmed et al. (2016); Gill et al. (2017); Jin et al. (2018); Tan et al. (2018)), thus restricting their analysis to the gray matter area. Recent studies by Ahmed et al. (2016) and El Azami et al. (2016) both restricted their evaluation to MRI^- patients and use the same performance metrics as ours thus allowing a more straightforward comparison as reported in table 3. The system proposed in Ahmed et al. (2016) based on SBM features coupled with semi-supervised hierarchical conditional random fields achieves 70% sensitivity among the top 10 detections per scan, while in El Azami et al. (2016), our model based on morphometric and intensity features coupled with a oc-SVM classifier allows achieving the same 70% sensitivity with an average of 4 false positives per scan.The slightly superior performance achieved by these two methods compared to our model is likely to be explained by the use of handcrafted features targeting FCD lesions. To emphasize this point, we implemented the method proposed in El Azami et al. (2016) and reported its performance on our patient dataset. As shown in table 3, a performance drop is observed when the evaluation is performed on our cohort of MRI^- patients including lesions with unknown signature. Our model, on the other hand, achieves good performance for the detection of FCD Type II lesions, as reported in table 1.

This comparative analysis confirms the competitive performance achieved by our brain anomaly detection model with 61% detection rate and a mean rank of 4.1 $\pm$ 2.9 on the 18 MRI^- lesions. The slightly inferior performance with regards to some reported results in the literature must be considered from the perspective of the challenging evaluation conditions considered in our study, where 1) the patient cohort consists of 18 purely MRI^- patients and lesions of unknown type, 2) the false positive rate is reported as an average number of FP detections estimated on the patient cohort and not on a control group.

Our brain anomaly detection model fails to identify the lesions of 8 patients. A visual analysis of the system's output for those cases seems to reveal two major reasons. For some of those patients, the raw output of the system highlighted some anomaly; however, after all the post-processing steps, those clusters were not ranked among top 10 detections. This is likely to mean that other anomalies present in the original images are considered 'anomalous' to a greater extent than the subtle epileptogenic lesions. The second category involves patients whose output score maps came out without any indication of

anomaly in the zone of interest. Our future work will be aimed at analyzing more thoroughly the cases when the system fails and investigate the reasons which may lay in the approach or the input images carrying no distinct marker for the lesion at all.

One potential limitation of our method is the need to register all the image volumes to perform a voxel-based analysis. El Azami et al. (2016) used a similar framework and showed that applying two different registration methods has no significant impact on the global performance of their oc-SVM models. In our configuration, the siamese network, receiving a pair of patches a priori centered at the same location in the brain but with possible registration inaccuracies, is likely to learn a latent representation that smooths these differences, given that most patches are registered adequately. Moreover, the raw output score maps are normalized in the post-processing step by dividing them by the estimated standard deviation among the normal population, which also penalizes the highly variable zones in the images including areas that are likely to carry registration inaccuracies. We are thus confident that our system is robust to such potential registration errors.

In this study, we chose to report the top 10 detected clusters which might seem quite a high number. The output maps, however, also provide the ranking of the detected clusters, thus allowing the radiologist to adjust the number of suspicious anomalies to visualize. As reported in the result section, the mean ranking of the true positive lesion ranges around 4, meaning that retaining the top 5 lesions may not result in a significant drop in sensitivity. As discussed previously by Ahmed et al. (2016), we think that this ranking approach provides a natural way to focus the radiologists' attention.

There are different options to improve the diagnostic performance of the proposed system. First, some pathology-specific information could be introduced in the post-processing step, by discarding some of the detected clusters based on shape and/or localization criteria. However, automated post-processing or cascaded classification systems should be cautiously evaluated since they may result in a significant sensitivity drop as observed in Tan et al. (2018). As shown on figure 7, some of the detected false positive clusters are indeed irregularities that can be easily discarded by a trained radiologist, especially those with a low rank. An alternative option is to move towards a semi-supervised setting by enhancing the neural network with a few 'pathological' patches that could be extracted from MRI-positive cases or after a careful analysis of retrospective MRI-negative patients, following, for instance, some ideas recently proposed in Shah et al. (2018). More improvement could be achieved by accounting for the complementary information provided by different imaging modalities. T1w and FLAIR modalities, introduced as channels to our network, allowed a significant diagnostic performance gain as shown on figure 8a. We expect a further performance gain by exploiting PET imaging as recently demonstrated in Tan et al. (2018).

Finally, the proposed method is quite straightforward to implement and to apply in daily practice as the output of the system can be obtained under a couple of minutes. Moreover, the system can be applied to detect other subtle pathologies which may serve as an advantage in the clinical routine. As encouraged by Kini et al. (2016), we also wish to make our computational pipeline available to clinicians. Ongoing work aims at transferring the different components (image pre-processing, deep feature extraction, etc.) of our brain anomaly detection model to the Neuroimaging toolbox of the VIP platform (https://vip.creatis.insa-lyon.fr) dedicated to the simulation and processing of massive data in medical imaging (Glatard et al. (2013)). This web portal allows users to access the CAD system as a service and significant computing resources and storage with no required technical skills beyond the use of a web browser.

7. Conclusion

In this study, we presented a novel configuration of unsupervised deep architectures for anomaly detection. The proposed regularized siamese network was leveraged as a representation learning mechanism, coupled with per voxel oc-SVM models. The clinical application of the proposed framework consists of automated detection of subtle epilepsy lesions in MRI-negative patients on T1-weighted and FLAIR MRI sequences. The approach achieved a sensitivity of 61% for 9 false detections per patient on pure MRI-negative patients. Epilepsy lesions were on average ranked among the top 4 most suspicious detected clusters, thus demonstrating a promising performance for this very challenging detection task.

8. Acknowledgements

This work received funds from the French Foundation for Research in Epilepsy (FFRE) and the Region Auvergne-Rhône-Alpes through the TADALOT project. It was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). The authors sincerely thank Jiazheng Chai for his valuable contribution to the implementation of the WAE model. The authors would also like to thank the Institute of Informatics, Slovak Academy of Sciences for access to GPU resources through the VIP portal and technical support.

References

References

Ahmed, B., Brodley, C.E., Blackmon, K.E., Kuzniecky, R., Barash, G., Carlson, C., Quinn, B.T., Doyle, W.K., French, J., Devinsky, O., Thesen, T., 2015.

Cortical feature analysis and machine learning improves detection of "mri-negative" focal cortical dysplasia. Epilepsy and behavior 48, 21–8.

Ahmed, B., Thesen, T., Blackmon, K.E., Kuzniekcy, R., Devinsky, O., Brodley, C.E., 2016. Decrypting "cryptogenic" epilepsy: Semi-supervised hierarchical conditional random fields for detecting cortical lesions in mri-negative patients. Journal of Machine Learning Research 17, 1–30.

Alarcon, G., Valentin, A., Watt, C., Selway, R., Lacruz, M., Elwes, R., Jarosz, J., Honavar, M., Brunhuber, F., Mullatti, N., et al., 2006. Is it worth pursuing surgery for epilepsy in patients with normal neuroimaging? Journal of Neurology, Neurosurgery & Psychiatry 77, 474–480.

Alaverdyan, Z., Chai, J., Lartizien, C., 2018a. Unsupervised feature learning for outlier detection with stacked convolutional autoencoders, siamese networks and wasserstein autoencoders: Application to epilepsy detection, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 210–217.

Alaverdyan, Z., Jung, J., Bouet, R., Lartizien, C., 2018b. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening, in: First conference on medical imaging with deep learning (MIDL 2018).

An, J., Cho, S., 2015. Variational autoencoder based anomaly detection using reconstruction probability. SNU Data Mining Center, Tech. Rep. .

Ashburner, J., 2009. Computational anatomy with the spm software. Magnetic resonance imaging 27, 1163–1174.

Ashburner, J., Friston, K., 2005. Unified segmentation. Neuroimage 26, 839–851.

Bach, F.R., Lanckriet, G.R., Jordan, M.I., 2004. Multiple kernel learning, conic duality, and the smo algorithm, in: Proceedings of the twenty-first international conference on Machine learning, ACM. p. 6.

Barkovich, A.J., Guerrini, R., Kuzniecky, R.I., Jackson, G.D., Dobyns, W.B., 2012. A developmental and genetic classification for malformations of cortical development: update 2012. Brain 135, 1348–1369.

Baur, C., Albarqouni, S., Navab, N., 2017. Semi-supervised deep learning for fully convolutional networks, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention (MICCAI 2017), Springer International Publishing, Cham. pp. 311–319.

Bell, M.L., Rao, S., So, E.L., Trenerry, M., Kazemi, N., Matt Stead, S., Cascino, G., Marsh, R., Meyer, F.B., Watson, R.E., et al., 2009. Epilepsy surgery outcomes in temporal lobe epilepsy with a normal mri. Epilepsia 50, 2053–2060.

Bernasconi, A., Bernasconi, N., Bernhardt, B.C., Schrader, D., 2011. Advances in mri for "cryptogenic" epilepsies. Nature Reviews Neurology 7, 99. URL: http://dx.doi.org/10.1038/nrneurol.2010.199, doi:10.1038/nrneurol.2010.199.

Bernasconi, N., Bernasconi, A., 2015. MRI-negative epilepsy: evaluation and surgical management. Cambridge University Press. pp. 16–27.

Bien, C.G., Raabe, A.L., Schramm, J., Becker, A., Urbach, H., Elger, C.E., 2012. Trends in presurgical evaluation and surgical treatment of epilepsy at one centre from 1988–2009. J Neurol Neurosurg Psychiatry , jnnp–2011.

Bien, C.G., Szinay, M., Wagner, J., Clusmann, H., Becker, A.J., Urbach, H., 2009. Characteristics and surgical outcomes of patients with refractory magnetic resonance imaging–negative epilepsies. Archives of Neurology 66, 1491–1499.

Bortsova, G., van Tulder, G., Dubost, F., Peng, T., Navab, N., van der Lugt, A., Bos, D., De Bruijne, M., 2017. Segmentation of intracranial arterial calcification with deeply supervised residual dropout networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 356–364.

Bos, D., Portegies, M.L., van der Lugt, A., Bos, M.J., Koudstaal, P.J., Hofman, A., Krestin, G.P., Franco, O.H., Vernooij, M.W., Ikram, M.A., 2014. Intracranial carotid artery atherosclerosis and the risk of stroke in whites: the rotterdam study. JAMA neurology 71, 405–411.

Bos, D., Vernooij, M.W., Elias-Smale, S.E., Verhaaren, B.F., Vrooman, H.A., Hofman, A., Niessen, W.J., Witteman, J.C., van der Lugt, A., Ikram, M.A., 2012. Atherosclerotic calcification relates to cognitive function and to brain changes on magnetic resonance imaging. Alzheimer's & Dementia 8, S104–S111.

Bowman, A.W., Azzalini, A., 1997. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. volume 18. OUP Oxford.

Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R., 1993. Signature verification using a "siamese" time delay neural network. IJPRAI 7, 669–688.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE transactions on medical imaging 35, 1229–1239.

Bruggemann, J.M., Wilke, M., Som, S.S., Bye, A.M., Bleasel, A., Lawson, J.A., 2007. Voxel-based morphometry in the detection of dysplasia and neoplasia in

childhood epilepsy: combined grey/white matter analysis augments detection. Epilepsy research 77, 93–101.

Bunch, P.C., Hamilton, J.F., Sanderson, G.K., Simmons, A.H., 1978. A free-response approach to the measurement and characterization of radiographic-observer performance. J. Appl. Photogr. Eng 4, 166–171.

Caputo, B., Sim, K., Furesjo, F., Smola, A., 2002. Appearance-based object recognition using svms: which kernel should i use?, in: Proc of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision, Whistler.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 15:1–15:58.

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 .

Cheplygina, V., de Bruijne, M., Pluim, J.P.W., 2018. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. CoRR abs/1804.06353. URL: http://arxiv.org/abs/1804.06353, arXiv:1804.06353.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 539–546.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W.J., Vernooij, M., De Bruijne, M., 2017. Gp-unet: Lesion detection from weak labels with a 3d regression network, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention (MICCAI 2017), Springer International Publishing, Cham. pp. 214–221.

El Azami, M., Hammers, A., Jung, J., Costes, N., Bouet, R., Lartizien, C., 2016. Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem. PloS one 11, e0161498.

Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recognition 58, 121–134.

Filippi, M., Rocca, M.A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., Rovira, A., Sastre-Garriga, J., Tintore, M., Frederiksen, J.L., Gasperini, C., Palace, J., Reich, D.S., Banwell, B., Montalban, X., Barkhof, F., 2016. Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. Lancet Neurol 15, 292–303. doi:10.1016/s1474-4422(15)00393-2.

Gill, R.S., Hong, S.J., Fadaie, F., Caldairou, B., Bernhardt, B., Bernasconi, N., Bernasconi, A., 2017. Automated detection of epileptogenic cortical malformations using multimodal mri, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 349–356.

Gill, R.S., Hong, S.J., Fadaie, F., Caldairou, B., Bernhardt, B.C., Barba, C., Brandt, A., Coelho, V.C., d'Incerti, L., Lenge, M., Semmelroch, M., Bartolomei, F., Cendes, F., Deleo, F., Guerrini, R., Guye, M., Jackson, G., Schulze-Bonhage, A., Mansi, T., Bernasconi, N., Bernasconi, A., 2018. Deep convolutional networks for automated detection of epileptogenic brain malformations, in: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing. pp. 490–497.

Glatard, T., Lartizien, C., Gibaud, B., Ferreira da Silva, R., Forestier, G., Cervenansky, F., Alessandrini, M., Benoit-Cattin, H., Bernard, O., Camarasu-Pop, S., Cerezo, N., Clarysse, P., Gaignard, A., Hugonnard, P., Liebgott, H., Marache, S., Marion, A., Montagnat, J., Tabary, J., Friboulet, D., 2013. A virtual imaging platform for multi-modality medical image simulation. IEEE Transactions on Medical Imaging 32, 110–18.

Guerrini, R., Sicca, F., Parmeggiani, L., 2003. Epilepsy and malformations of the cerebral cortex. Epileptic Disorders 5, 9–26.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Human brain mapping 19, 224–247.

Hashemi, S.R., Mohseni Salehi, S.S., Erdogmus, D., Prabhu, S.P., Warfield, S.K., Gholipour, A., 2019. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access 7, 1721–1735. doi:10.1109/ACCESS.2018.2886371.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. Hemis: Hetero-modal image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 469–477.

Hinton, G.E., Zemel, R.S., 1994. Autoencoders, minimum description length and helmholtz free energy, in: Cowan, J.D., Tesauro, G., Alspector, J. (Eds.), Advances in Neural Information Processing Systems 6. Morgan-Kaufmann, pp. 3–10. URL: http://papers.nips.cc/paper/798-autoencoders-minimum-description-length-and-helmholtz-free-energy.pdf.

Hong, S.J., Kim, H., Schrader, D., Bernasconi, N., Bernhardt, B.C., Bernasconi, A., 2014a. Automated detection of cortical dysplasia type ii in mri-negative epilepsy. Neurology 83, 48–55.

Hong, S.J., Kim, H., Schrader, D., Bernasconi, N., Bernhardt, B.C., Bernasconi, A., 2014b. Automated detection of cortical dysplasia type ii in mri-negative epilepsy. Neurology 83, 48–55.

Huppertz, H.J., Grimm, C., Fauser, S., Kassubek, J., Mader, I., Hochmuth, A., Spreer, J., Schulze-Bonhage, A., 2005. Enhanced visualization of blurred gray–white matter junctions in focal cortical dysplasia by voxel-based 3d mri analysis. Epilepsy research 67, 35–50.

Jin, B., Krishnan, B., Adler, S., Wagstyl, K., Hu, W., Jones, S., Najm, I., Alexopoulos, A., Zhang, K., Zhang, J., Ding, M., Wang, S., Wang, Z.I., 2018. Automated detection of focal cortical dysplasia type ii with surface-based magnetic resonance imaging postprocessing and machine learning. Epilepsia 59, 982–992. doi:10.1111/epi.14064.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis 36, 61 – 78. URL: http://www.sciencedirect.com/science/article/pii/S1361841516301839, doi:https://doi.org/10.1016/j.media.2016.10.004.

Keller, S.S., Cresswell, P., Denby, C., Wieshmann, U., Eldridge, P., Baker, G., Roberts, N., 2007. Persistent seizures following left temporal lobe surgery are associated with posterior and bilateral structural and functional brain abnormalities. Epilepsy research 74, 131–139.

Kini, L.G., Gee, J.C., Litt, B., 2016. Computational analysis in epilepsy neuroimaging: A survey of features and methods. Neuroimage 11, 515–529.

Kiran, B.R., Thomas, D.M., Parakkal, R., 2018. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging 4.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sanchez, C.I., 2017. A survey on deep learning in medical image analysis. Med Image Anal 42, 60–88. doi:10.1016/j.media.2017.07.005.

Loosli, G., Aboubacar, H., 2017. Using svdd in simplemkl for 3d-shapes filtering. arXiv preprint arXiv:1712.02658 .

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al., 2001. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). Philosophical Transactions of the Royal Society of London B: Biological Sciences 356, 1293–1322.

Munawar, A., Vinayavekhin, P., De Magistris, G., 2017a. Limiting the reconstruction capability of generative neural network using negative learning. arXiv preprint arXiv:1708.08985 .

Munawar, A., Vinayavekhin, P., De Magistris, G., 2017b. Spatio-temporal anomaly detection for industrial robots through prediction in unsupervised feature space, in: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE. pp. 1017–1025.

Pawlowski, N., Matthew C.H. Lee, M.R., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., Zeiler, F., Digby, R., Coles, J.P., Rueckert, D., Menon, D.K., Newcombe, V.F., Glocker, B., 2018. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders, in: First conference on medical imaging with deep learning (MIDL 2018).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y., 2008. Simplemkl. Journal of Machine Learning Research 9, 2491–2521.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, Springer International Publishing. pp. 146–157.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural computation 13, 1443–1471.

Shah, M.P., Merchant, S.N., Awate, S.P., 2018. Abnormality detection using deep neural networks with robust quasi-norm autoencoding and semi-supervised learning, in: IEEE International Symposium on Biomedical Imaging (ISBI 2018), pp. 568–572.

Sonnenburg, S., Rätsch, G., Schäfer, C., 2006. A general and efficient multiple kernel learning algorithm, in: Advances in neural information processing systems, pp. 1273–1280.

Srivastava, S., Maes, F., Vandermeulen, D., Van Paesschen, W., Dupont, P., Suetens, P., 2005. Feature-based statistical analysis of structural mr data for automatic detection of focal cortical dysplastic lesions. NeuroImage 27, 253–266.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham. pp. 240–248.

Tan, Y.L., Kim, H., Lee, S., Tihan, T., Ver Hoef, L., Mueller, S.G., Barkovich, A.J., Xu, D., Knowlton, R., 2018. Quantitative surface analysis of combined mri and pet enhances detection of focal cortical dysplasias. Neuroimage 166, 10–18.

Thesen, T., Quinn, B.T., Carlson, C., Devinsky, O., DuBois, J., McDonald, C.R., French, J., Leventer, R., Felsovalyi, O., Wang, X., et al., 2011. Detection of epileptogenic cortical malformations with surface-based mri morphometry. PloS one 6, e16430.

Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B., 2017. Wasserstein Auto-Encoders. ArXiv e-prints arXiv:1711.01558.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. NeuroImage 155, 159–168.

Wagner, J., Weber, B., Urbach, H., Elger, C.E., Huppertz, H.J., 2011. Morphometric mri analysis improves detection of focal cortical dysplasia type ii. Brain 134, 2844–2854.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J.T., Barkhof, F., Benavente, O.R., Black, S.E., Brayne, C., Breteler, M., Chabriat, H., Decarli, C., de Leeuw, F.E., Doubal, F., Duering, M., Fox, N.C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., Oostenbrugge, R., Pantoni, L., Speck, O., Stephan, B.C., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P.B., Dichgans, M., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol 12, 822–38. doi:10.1016/s1474-4422(13)70124-8.

Wiebe, S., Blume, W.T., Girvin, J.P., Eliasziw, M., 2001. A randomized, controlled trial of surgery for temporal-lobe epilepsy. New England Journal of Medicine 345, 311–318.

Zenati, H., Foo, C.S., Lecouat, B., Manek, G., Chandrasekhar, V.R., 2018. Efficient gan-based anomaly detection, in: ICLR workshop. URL: https://openreview.net/forum?id=BkXADmJDM.

Zheng, L., Idrissi, K., Garcia, C., Duffner, S., Baskurt, A., 2015. Triangular similarity metric learning for face verification, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7. doi:10.1109/FG.2015.7163085.

Supplementary File for paper : Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening

Zaruhi Alaverdyan[a], Julien Jung[b], Romain Bouet[b], Carole Lartizien[a]

[a] *Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621, Lyon, France*
[b] *Lyon Neuroscience Research Center, CRNL, INSERM U1028, CNRS UMR5292, University Lyon 1, Lyon, France*

1. Comparison with unsupervised models based on handcrafted features

We compare the performance of our brain anomaly detection model to those obtained with two configurations based on handcrafted features.

The first configuration is a general linear model (GLM) (SPM analysis) learned on feature maps derived from T1w images using three settings - 1. junction contrast, 2. extension contrast and 3. the conjunction of both contrasts - for a statistical threshold of 0.001 as done in El Azami et al. (2016). These feature maps model the junction between the gray and white matters as described in Huppertz et al. (2005) and Wagner et al. (2011). For a fair comparison, the same clustering and ranking procedures as performed for our anomaly detection model (see section 3.4) were applied and the top 10 most suspicious clusters were considered. The results are summarized in table 1. While extension contrast detects one additional lesion compared with our monomodal architecture based on T1w MRI (T1 rSN), the combination of junction and extension contrasts results in an inferior performance. This shows that it is not trivial to perform a multivariate analysis within this approach, retaining the best performance obtained in the univariate case. We should also note that without applying the ranking method, the original SPM implementation produces much more false positive detections without any significant change in sensitivity.

The second configuration whose performance is also summarized in table 1 is the model proposed in El Azami et al. (2016) where a oc-SVM is built per voxel using the same junction and extension maps, as above. This comparison

Table 1: Comparative sensitivity of our CAD model and those achieved with GLM and oc-SVM based on handcrafted features. First column: overall sensitivity; the number of detected patients / total number of patients inside parentheses. Second column: sensitivity calculated on MRI-negative patients only. The reported sensitivity corresponds to an average number of 9 false positive (FP) detections per patient.

	Overall sensitivity	Sensitivity on MRI^-
T1/FLAIR rSN + oc-SVM	**0.62 (13/21)**	**0.61 (11/18)**
T1 rSN + oc-SVM	0.43 (9/21)	0.39 (7/18)
FLAIR rSN + oc-SVM	0.57 (12/21)	0.5 (9/18)
T1 rSN/FLAIR rSN + MKL oc-SVM	0.52 (11/21)	0.5 (9/18)
Junction-Extension + oc-SVM	0.38 (8/21)	0.39 (7/18)
Junction + GLM	0.28 (6/21)	0.27 (5/18)
Extension + GLM	0.43 (9/21)	0.44 (8/18)
Junction-Extension + GLM	0.24 (5/21)	0.22 (4/18)

reveals the advantage of the representations learnt with the rSN network versus the handcrafted features. As it can be seen, the maximum sensitivity achieved with the handcrafted features is 8/21 while 9/21 with our system on T1-w MRI, for 9 FPs. The gap becomes even more remarkable when the FLAIR modality is considered, leading our system to achieve 62% sensitivity.

References

El Azami, M., Hammers, A., Jung, J., Costes, N., Bouet, R., Lartizien, C., 2016. Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem. PloS one 11, e0161498.

Huppertz, H.J., Grimm, C., Fauser, S., Kassubek, J., Mader, I., Hochmuth, A., Spreer, J., Schulze-Bonhage, A., 2005. Enhanced visualization of blurred gray–white matter junctions in focal cortical dysplasia by voxel-based 3d mri analysis. Epilepsy research 67, 35–50.

Wagner, J., Weber, B., Urbach, H., Elger, C.E., Huppertz, H.J., 2011. Morphometric mri analysis improves detection of focal cortical dysplasia type ii. Brain 134, 2844–2854.