Unsupervised representation learning for anomaly detection on neuroimaging. Application to epilepsy lesion detection on brain MRI

Zaruhi Alaverdyan

N°d'ordre NNT : 2019LYSEI005

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé (CREATIS)

Ecole Doctorale N° 160
(Electronique, Electrotechnique, Automatique)

Spécialité de doctorat : Traitement du Signal et de l'image

Soutenue publiquement le 18/01/2019, par :
Zaruhi ALAVERDYAN

Unsupervised representation learning for anomaly detection on neuroimaging. Application to epilepsy lesion detection on brain MRI

Devant le jury composé de :

Cardoso, Jorge M.	Professeur des Universités	King's College London	Rapporteur
Mateus, Diana	Professeur des Universités	Ecole Centrale Nantes	Rapporteure
Bonnet-Loosli, Gaëlle	Maître de Conférences	Université Clermont Auvergne	Examinatrice
Fromont, Elisa	Professeur des Universités	Université de Rennes 1	Examinatrice
Jung, Julien	Praticien Hospitalier	CHU de Lyon	Examinateur
Lartizien, Carole	Directrice de recherche	CNRS	Directrice de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	M. Stéphane DANIELE Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	M. Gérard SCORLETTI École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	M. Philippe NORMAND UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 Avenue Jean CAPELLE INSA de Lyon 69 621 Villeurbanne Tél : 04.72.68.49.09 Fax : 04.72.68.49.16 **emmanuelle.canet@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 Fax : 04.72.43.16.87 **infomaths@univ-lyon1.fr**	M. Luca ZAMBONI Bât. Braconnier 43 Boulevard du 11 novembre 1918 69 622 Villeurbanne CEDEX Tél : 04.26.23.45.52 **zamboni@maths.univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Marion COMBE Tél : 04.72.43.71.70 Fax : 04.72.43.87.12 Bât. Direction **ed.materiaux@insa-lyon.fr**	M. Jean-Yves BUFFIÈRE INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Marion COMBE Tél : 04.72.43.71.70 Fax : 04.72.43.87.12 Bât. Direction **mega@insa-lyon.fr**	M. Jocelyn BONJOUR INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Viviane POLSINELLI Brigitte DUBOIS INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **viviane.polsinelli@univ-lyon2.fr**	M. Christian MONTES Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Մերոնց.

Acknowledgements

I would first like to thank my PhD supervisor, Carole Lartizien, for starting this project and allowing me to become a part of it. It has been an ambitious endeavour and this work reflects only a part of the efforts we have invested.

I would also like to thank all the people who, voluntarily or not, shaped my experience of the past 3 years. Mo, Meriem and Younes, thanks for welcoming me, showing me around and tailoring my expectations of what was going to come next. Thanks to Robert, for introducing many important ideas into my life and Tiago, for all the music, dances and nice moments we shared. Aneline, Sami, Yuemeng, Maxime and Vincent, I thank you all for the countless jokes and foolishness which made my last days in Lyon so memorable. Special thanks to Emeline, for all the support you gave me and for the fun we had. Nina and Kenny, well, you know ... I am deeply thankful to my dear mexicans, Pavel, for lifting my spirits and putting up with all the whining, and Manuel, simply for being the best.

Finally, I would like to thank all those who, despite being miles away, are of the utmost importance to me. I thank my parents and my dear sister, for their blind faith in me and the unconditional support. My friends, Anka, Arpy and Hrant, for reminding me of the good times every time we talked.

Having all of you in my life is perhaps the best luck of all.

Abstract

Epilepsy affects around 50 million people worldwide, a third of those diagnosed with medically refractory epilepsy where seizures cannot be controlled by pharmacotherapy. For such patients, surgical resection of the epileptogenic zone may offer a seizure-free life. The success of such surgeries largely depends on the accuracy of the epileptogenic zone localization. Neuroimaging, including magnetic resonance imaging (MRI) and positron emission tomography (PET), has been increasingly considered in the pre-surgical examination routine.
This work represents one attempt to develop a computer aided diagnosis system for epileptogenic lesion detection based on neuroimaging data, in particular T1-weighted and FLAIR MR sequences. Given the complexity of the task and the lack of a representative voxel-level labeled data set, the adopted approach, first introduced in [El Azami et al., 2016], consists in casting the lesion detection task as a per-voxel outlier detection problem. The system is based on training a one-class SVM model for each voxel in the brain on a set of healthy controls, so as to model the normality of the voxel. For an unseen patient, each voxel is assessed against the corresponding one-class SVM model which yields a signed score of its anomalousness. Anomalous lesions can hence be found as local neighborhoods of voxels with low scores.
The main focus of this work is to design representation learning mechanisms, capturing the most discriminant information from multimodality imaging. Manual features, designed to mimic the characteristics of certain epilepsy lesions, such as focal cortical dysplasia (FCD), on neuroimaging data, are tailored to individual pathologies and cannot discriminate a large range of epilepsy lesions. Such features reflect the known characteristics of lesion appearance; however, they might not be the most optimal ones for the task at hand. Our first contribution consists in proposing various unsupervised neural architectures as potential feature extracting mechanisms and, eventually, introducing a novel configuration of siamese networks, to be plugged into the outlier detection context. The proposed system, evaluated on a set of T1-weighted MRI of epilepsy patients, showed a promising performance but a room for improvement as well. To this end, we considered extending the CAD system so as to accommodate multimodality data which offers complementary information on the problem at hand. Our second contribution, therefore, consists in proposing strategies to combine representations of different imaging modalities into a single framework for anomaly detection. The extended system showed a significant improvement on the task of epilepsy lesion detection on T1-weighted and FLAIR images. Our last contribution focuses on the integration of PET data into the system. An obstacle encountered often in medical applications is the small number of subjects with the full set of imaging modalities. This limits the performance of a system when the subjects with missing data are discarded. We therefore delve into strategies of synthesizing PET data from the corresponding MRI acquisitions and show an improved performance of the system when synthesized images are used in addition to the real ones.

Résumé étendu

Environ 50 million personnes souffrent d'une épilepsie partielle, dont un tiers atteint d'une épilepsie réfractaire à tous les médicaments. La chirurgie, qui constitue aujourd'hui le meilleur recours therapeutique, nécessite un bilan préopératoire complexe. L'analyse de données d'imagerie telles que l'imagerie par résonance magnétique (IRM) anatomique et la tomographie d'émission de positons (TEP) au FDG (fluorodésoxyglucose) tend à prendre une place croissante dans ce protocole, et pourrait à terme limiter les recours à l'électroencéphalographie intracérébrale (SEEG), procédure très invasive mais qui constitue encore la technique de référence.

Cette étude vise à développer un système d'aide au diagnostic (CAD) pour le détection de lésions épileptogènes, reposant sur l'analyse de données de neuroimagerie, notamment, l'IRM T1 et FLAIR. Etant donné la complexité du problème et le manque d'une base de données, annotée à l'échelle de voxel, représentative de la pathologie, l'approche adoptée, introduite précédemment par [El Azami et al., 2016], consiste à placer la tâche de détection dans le cadre de la détection du changement à l'échelle du voxel. Le système est basé sur l'apprentissage d'un modèle one-class SVM pour chaque voxel dans le cerveau, en utilisant un ensemble de sujets sains, dont le but est de modéliser la normalité du voxel. Pour un patient donné, chaque voxel est évalué par le modèle oc-SVM, correspondant à sa localization spatiale, et ce dernier produit une valeur numérique signée, représentant l'anormalité du voxel. Les lésions anormales peuvent ensuite s'identifier comme des voisinages de voxels, ayant des valeurs très négatives.

L'objectif principal de ce travail est de développer des mécanismes d'apprentissage de représentations, qui capturent les informations les plus discriminantes à partir de l'imagerie multimodale. Les caractéristiques manuelles, conçues pour imiter les caractéristiques de certaines lésions épileptogènes sur la neuroimagerie, notamment les dysplasies corticales focales (FCD), sont spécifiques aux pathologies individuelles et n'ont pas la capacité de discriminer un ensemble varié de lésions épileptogènes. Ce type de caractéristiques reflète la connaissance existante sur l'apparence de lésions. Par contre, elles ne sont pas forcément les plus pertinentes pour la tâche visée. Notre première contribution porte sur l'intégration de différents réseaux profonds non-supervisés, en tant que mécanismes d'extraction de caractéristiques, dans le cadre du probleme de detection de changement. Eventuellement, nous introduisons une nouvelle configuration des réseaux siamois, mieux adapté à ce contexte. Le système CAD proposé a été évalué sur l'ensemble d'images T1 IRM des patients atteint d'épilepsie. Nous avons démontré une performance importante qui reste, tout de même, à améliorer. Pour cela, nous avons considéré d'étendre le système pour intégrer des données multimodales qui possèdent des informations complémentaires sur la pathologie en question. Notre deuxième contribution, donc, consiste à proposer des stratégies de combinaison des différentes modalités d'imagerie dans un système pour la détection des changements. Ce système

multimodal a montré une amélioration importante sur la tâche de détection de lésions épileptogènes sur les IRM T1 et FLAIR.
Notre dernière contribution se focalise sur l'intégration des données PET dans le système proposé. Très souvent, dans les applications médicales, le nombre de sujets ayant les acquisitions de toutes les modalités envisagées, est assez limité. La performance des systèmes, où l'on ne considère que les sujets ayant toutes les acquisitions, est souvent faible. Pour cette raison, nous envisageons de synthétiser les données manquantes à partir des images des autres modalités présentes. Nous essayons, donc, de générer des images TEP en se servant des images IRM disponibles. Nous démontrons que le système entraîné sur les données réelles et synthétiques présente une amélioration importante par rapport au système entraîné sur les images réelles uniquement.

Contents

General introduction

Epilepsy is a common neurological disorder affecting around 50 million people worldwide according to the World Health Organization (WHO). It is characterized by an enduring predisposition to generate unprovoked brain seizures [Fisher et al., 2014]. Epilepsy treatment involves consistent intake of antiepileptic drugs on a long-term basis which allows to control the seizures in up to 70% of focal epilepsy patients. The remaining 30% are referred to as intractable epilepsy patients [Kwan and Brodie, 2000]. For such patients, surgical resection of the epileptogenic zone may offer a seizure-free life. The success of such surgeries largely depends on the accuracy of the epileptogenic zone localization. Neuroimaging, including magnetic resonance imaging (MRI) and positron emission tomography (PET), has been increasingly considered in the pre-surgical examination routine. On neuroimaging data, epilepsy lesions, however, have very subtle characteristics and highly variable profiles which results in clinicians frequently considering the scans normal (MRI-negative patients). For such patients with visually unconfirmed lesions, the success rate of surgery is 2-3 times lower than when the lesion is detected over a routine visual examination [Téllez-Zenteno et al., 2010]. Neurologists would greatly benefit from a computer aided diagnosis (CAD) system automatically processing the data so as to provide probability maps highlighting abnormal regions in the image. The clinical benefit of such an automated image analysis tool during the pre-surgical planning is to optimally select candidates for the epilepsy lesion resection surgery and to guide the placement depth of EEG electrodes when an invasive EEG is required for an accurate delineation of the epileptogenic zone.

Over the recent years, many attempts have been made in order to propose automated solutions for epilepsy detection on neuroimaging data. Most of those studies are based on the extraction of different descriptors from the images, reflecting the clinical knowledge on the appearance of specific epilepsy lesions. The descriptors are then exploited either in statistical analysis based approaches [Chen et al., 2008, Focke et al., 2008, Riney et al., 2012], or, more recently, in machine learning based frameworks [El Azami et al., 2016, Hong et al., 2014, Ahmed et al., 2015, Ahmed et al., 2016].

This work represents one attempt to develop a computer aided diagnosis system for epileptogenic lesion detection based on neuroimaging data, in particular T1-weighted and FLAIR MR sequences. Given the complexity of the task and the lack of a representative voxel-level labeled data set (the annotations are much more difficult to obtain for MRI negative patients), the adopted approach, first introduced in [El Azami et al., 2016], consists in casting the lesion detection task as a per-voxel outlier detection problem. The system is based on training a one-class SVM model for each voxel in the brain on a set of healthy controls, so as to model the normality of the voxel. For an unseen patient, each voxel is assessed against the corresponding one-class SVM model which yields a signed score of its anomalousness. Anomalous lesions can hence be found as local neighborhoods of voxels with low scores. This approach bypasses the need of labeled training data set and therefore, offers an alternative to supervised learning.

The main focus of this work is to design representation learning mechanisms, capturing the most discriminant information from multimodality imaging. Manual features, designed to mimic the characteristics of certain epilepsy lesions, such as focal cortical dysplasia (FCD), on neuroimaging data, are tailored to individual pathologies and cannot discriminate a large range of epilepsy lesions. Such features reflect the known characteristics of lesion appearance; however, they might not be the optimal ones for the task at hand.

Our first contribution consists in proposing various unsupervised neural architectures as potential feature extracting mechanisms and, eventually, introducing a novel configuration of siamese networks, to be plugged into the outlier detection context. The proposed system, evaluated on a set of T1-weighted MRI of epilepsy patients, showed a promising performance but a room for improvement as well.
To this end, we considered extending the CAD system so as to accommodate multimodality data which offers complementary information on the problem at hand. Our second contribution, therefore, consists in proposing strategies to combine representations of different imaging modalities into a single framework for anomaly detection. The extended system showed a significant improvement on the task of epilepsy lesion detection on T1-weighted and FLAIR images. Our last contribution focuses on the integration of PET data into the system. An obstacle encountered often in medical applications is the small number of subjects with the full set of imaging modalities. This limits the performance of a system when the subjects with missing data are discarded. We therefore delve into strategies of synthesizing PET data from the corresponding MRI acquisitions and show an improved performance of the system when synthesized images are used in addition to the real ones.

This work is divided into three main parts. Part I starts with a detailed description of modern CAD systems in chapter 1. Chapter 2 presents an overview of currently popular applications of deep learning methods in medical imaging. In chapter 3, we describe and review the existing methods for epilepsy lesion detection on neuroimaging. Chapter 4

presents our analysis of the challenges and constraints of the problem at hand and formalizes our approach.

In part II, we introduce the main contribution of this work, by first giving an overview of the proposed CAD system and a detailed description of the available data set in chapter 5. Chapter 6 presents the existing unsupervised deep architectures and concludes with a novel configuration tailored to the task of outlier detection. Chapter 7 presents the performance obtained with the proposed system, using the representations learnt with deep architectures, on the detection of epilepsy lesions on T1-weighted MRI.

Part III comprises a review on the possible strategies of multimodality data fusion and our own choices in chapter 8. Chapter 9 presents the results obtained with the proposed CAD system on the combination of T1-weighted and FLAIR MRI data with two data fusion strategies. In chapter 10, we review the current approaches of cross-modality image generation, and apply one for MRI to PET synthesis. Finally, we present the performance of the CAD system using the synthesized PET images as well as the real acquisitions. The manuscript ends with our overall conclusion and perspectives on the future work.

I Medical and scientific context

Chapter 1

Image-based computer aided diagnosis systems

The clinical environment often encounters emerging challenges or the existing ones turning more complicated with the growing amount of information and limited resources to exploit. In particular, the amount of data generated through various medical protocols, including different medical imaging techniques, becomes overwhelming and difficult to analyze with no automated solutions. As such, computer aided diagnosis (CAD) systems are tools designed to make the analysis of medical data more efficient and less time-consuming, in order to eventually assist clinicians in their tasks. Many image-based CAD systems have been investigated for various tasks, ranging from organ / tissue segmentation to detection of various cardiac, brain or cancerous pathologies in patients.

Over the recent years, CAD systems have been enriched with all the more powerful machine learning algorithms as the core decision making mechanism, outputting relevant information to a clinician. In many applications such systems have achieved impressive performance rates, sometimes higher than those of humans. Such CAD systems are typically trained on the characteristics relevant for the problem at hand (observed in the clinical practice) that are formalized and computed using mathematical expressions. Recently, the explicit translation of clinical characteristics to computable features has been replaced with deep learning architectures, operating directly on the raw medical data (such as images). Deep learning based CAD systems have shown outstanding results in many problems which has turned them into the method of choice in research on many medical problems.

In this chapter we introduce and describe the main components of image-based CAD systems. We present the methods frequently used in CADs, their advantages and limitations. Eventually, we present the common performance evaluation strategies allowing to assess the quality of a CAD system.

1.1 General CAD description

Various medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT) and positron emission tomography (PET), have played a major role in medical diagnosis as they provide an internal view on the anatomical and functional state of a patient and, hence, guide the radiologists' medical decisions. With the ubiquitous use of such imaging techniques, the past two decades have marked a fast evolution of computer-aided diagnosis (CAD) systems. Image-based computer-aided diagnosis systems are tools that assist clinicians in the interpretation and analysis of medical images for various tasks. Some typical applications of CAD systems include organ and lesion segmentation, abnormality detection and many others (an example is shown on fig. 1.2). Over the recent years many CAD systems have been proposed for breast [Dheeba et al., 2014], lung [Hua et al., 2015] and prostate cancer detection [Niaf et al., 2014, Litjens et al., 2014]; for brain pathologies, various CAD systems tackled such problems as Alzheimer's disease diagnosis, Multiple Sclerosis lesion segmentation, detection of enlarged perivascular spaces in the basal ganglia, etc. An efficient CAD system can improve the decisions of radiologists who, due to various reasons, may miss or overlook a piece of information in the high load of data [Doi, 2007].

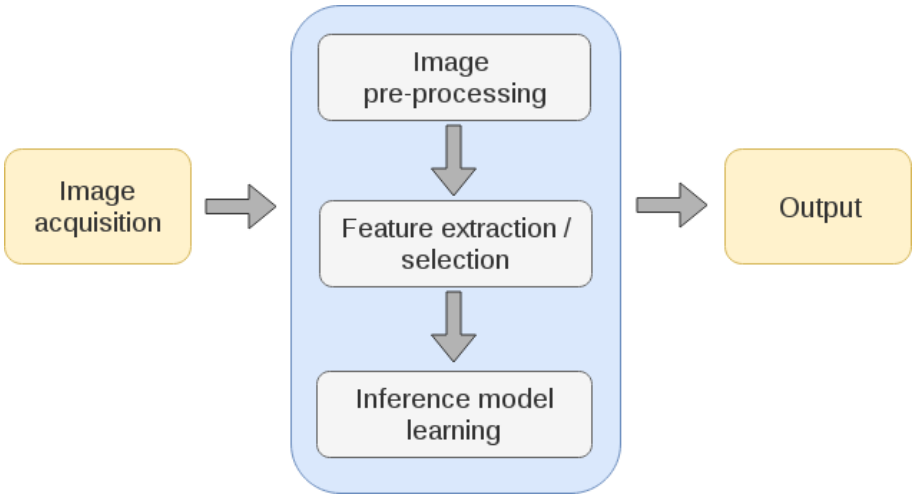


Figure 1.1: A typical image-based CAD system pipeline.

Image-based CAD systems are designed to apply an automated model on the given input images, so as to produce an output corresponding to the problem [van Ginneken et al., 2011]. As illustrated on fig. 1.1, a typical CAD system entails the following steps - image pre-processing, feature extraction/selection and inference model learning (typically using a machine learning algorithm). Fig. 1.2 illustrates a CAD system taking at input multimodality neuroimaging data and outputting a probabilistic map highlighting suspicious areas found by the system.

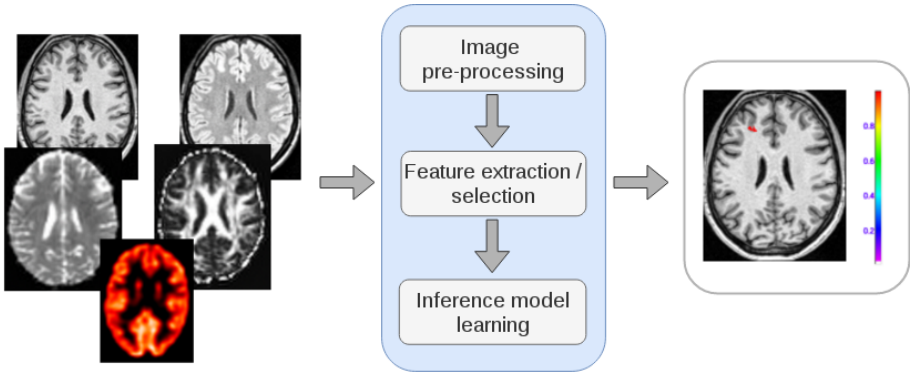


Figure 1.2: A CAD system for neuroimaging. Input: multimodality neuroimaging data. Output: Probabilistic map highlighting suspicious areas detected by the CAD system.

Input-output granularity

In a CAD system, the input may correspond to an image voxel, an image patch, a region of interest (ROI) or the entire image. The first three cases have the advantage of a smaller data size in comparison to the full image approach. The output of a CAD system may be at voxel-level, ROI-level or image (subject) level. For voxel-level CAD systems, each voxel is assigned a value produced by the system, eventually constituting an output map. For subject-level CAD systems, the output corresponds to the given image as a whole.

Image pre-processing

Typically, the CAD input images are first processed in order to enhance their quality. Common pre-processing steps include reduction of artifacts, noise reduction, image normalization, etc.

Feature extraction/selection

This step associates the input images with a set of characteristics (features) measured on the image following some definition/formula, relevant to the task at hand. For example, clinical knowledge on the pathology in question may be translated to a set of mathematical formulae and produce a set of descriptors of the pathology. Feature selection can further be performed in order to select the most relevant descriptors and discard the redundant ones. The features may also be learnt in an automatic data-driven fashion. In many recent applications this step is incorporated into deep architectures tailored to the task.

Inference model learning

In this step an inference model is built either based on the explicit descriptors acquired in the previous step or on the raw data. The choice of the model depends on the particularity of the task: it can be supervised (labels are available), unsupervised (labels are not available) and semi-supervised (labels are somewhat available).

Below we present some more details on these steps.

1.1.1 Input-output granularity

The granularity of the system depends on the desired outcome and the clinical need. CAD systems may be broadly categorized in two types. CADe systems seek to identify abnormal regions of a given image with respect to the pathology of interest. CADx systems aim at characterizing the pathology, its type/category, stage and severity [Petrick et al., 2013]. Image voxels, patches, ROIs and entire images may serve as input to CADs. The main categories of CADs with respect to the output granularity are:

Subject-level CADs

In this case the desired outcome of the CAD is usually to classify an image of a given subject into some category. Commonly such CADs perform binary classification by discriminating healthy versus pathological cases. The output may be expressed as a probability or a label. While it is possible to locate approximately the abnormal zone resulting in the binary output, its detection is not the main objective and usually is not performed. Some CADs are used to classify a given image into one of the categories of the pathology.

ROI-level CADs

In this case the model is based on a part of the input image corresponding to the region of interest (ROI), delineated by a radiologist; the remaining part is either irrelevant to the task at hand or is not significant and, hence, is not considered by the CAD system. Especially when the feature extraction step is applied explicitly, reducing the focus to a ROI instead of the full image allows acquiring relevant features over the ROI and reduce the dimensionality of the input data. The output of ROI-level CADs is a ROI-level score or label.

Voxel-level CADs

Some very popular CADs are designed to discriminate each voxel and produce either a probabilistic score map or a binary/n-ary map where each voxel is assigned a probability of being pathological or a binary/n-ary label representing the category it has been classified into.

1.1.2 Feature extraction and selection

An explicit extraction of features from the raw data has been a common choice for a very long time. In some medical applications there may be an accumulated clinical knowledge on the characteristics of the pathology of interest which can be translated into features by using appropriate formulae (table 3.1 lists such characteristics and their corresponding features for epilepsy). This certainly helps the discrimination of the pathology. However, in many contexts such knowledge is either not available or is insufficient and, hence, other

methods are exploited to extract features. For instance, one common approach is to extract generic image descriptors including

1. textural features describing textural patterns, frequently represented by statistical measures computed over a neighborhood (mean, standard deviation, etc) or derived from the grey-level co-occurrence matrix as described in [Haralick et al., 1973] (contrast, entropy, energy, etc). The latter matrix models the joint probability density of the occurrence of grey levels for two pixels with a spatial relationship defined by the chosen relative direction and the distance between the two pixels.

2. local descriptors (filters) to detect edges and shapes such as Gabor filters [Manjunath and Ma, 1996] that can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope.

3. robust image descriptors such as HOG [Dalal and Triggs, 2005], SIFT [Lowe, 1999] and SURF [Bay et al., 2006]. SIFT (scale invariant feature transform) combines a scale invariant region (key point) detector and a descriptor represented by the histogram of the gradient distribution in the detected regions.

By varying the parameters of such image descriptors (such as the relative direction and the voxel distance in Haralick features) a very large number of features can be obtained. Typically, such a choice of features is accompanied by a *feature selection* strategy which consists in keeping only the most relevant feature subset and discard the rest. Since it is computationally exhaustive to evaluate all possible subsets and keep the most discriminative one, other practical strategies have been proposed for feature selection. As such, forward-stepwise (backward-stepwise) feature selection consists in greedily adding (eliminating) the most (least) informative feature starting from the empty (full) set of features. The informativeness of the candidate features in each step is measured by some quantity, for instance Akaike Information Criterion (AIC) [Akaike, 1974]. Recursive feature elimination [Guyon et al., 2002] is another feature selection method that starts by fitting a model, that assigns weights to features (such as the coefficients of a linear model), to the entire feature set and later eliminates the features with the smallest weights. The procedure is performed recursively on the current feature subset. Feature selection may improve the performance of the system by eliminating irrelevant or redundant features and enhance the interpretability of the system, especially when the number of features is greater than the number of examples [Guyon and Elisseeff, 2003].

Features obtained in such way are referred to as *handcrafted*/*manual* features. The major disadvantage of such descriptors is their limited capacity in modeling complex phenomena as they are constrained with the kind of transformation they are designed to perform. Moreover, these features are not data-dependent and, hence, do not leverage the particular patterns that may be present in the data. Mainly for this reason there has been a major

shift over the last years from handcrafted features to data-driven features automatically learnt from the data [Litjens et al., 2017]. Neural networks are one way of accomplishing data-driven feature learning (more details will follow in chapter 2).

1.1.3 Inference model learning

The vast majority of the state-of-the-art CAD systems for various applications employ machine learning algorithms in order to build automated models that perform the task at hand. Such models take into account the nature of the application, the available data and the particularity of the problem.

A typical learning problem has the following setup. Given a data set of n observations $X = \{x_i | i = 1, ..., n\}$ generated by a fixed but unknown distribution $P(x)$, a machine learning algorithm seeks to model a function $f(x)$ that, depending on the nature of the algorithm, outputs a relevant value for a (previously unseen) observation. Machine learning algorithms can be broadly categorized into two major groups - supervised learning methods and unsupervised learning methods.

In supervised learning, a set of output values $Y = \{y_i | i = 1, ..., n\}$, associated with the examples in X, following a fixed but unknown conditional distribution $P(y|x)$, is available. The learning algorithm seeks to find a function $f(x; \theta)$, characterized by a set of parameters θ, that approximates Y as accurately as possible. When Y is composed of continuous numerical values, the problem is referred to as *regression* problem; when Y takes values from a finite set of discrete values (*labels*), the problem is called $classification$ problem. The classification problems may further be divided into *binary* ($K = 2$) and *multiclass* ($K > 2$) classification.

In unsupervised learning, no labels are associated with the observations and an unsupervised learning algorithm seeks to discover the properties of the data set, usually based on the distribution $P(x)$.

Below we present the most common supervised and unsupervised methods used in modern CAD systems. Since most CAD systems are designed to solve classification tasks, we will not detail on the regression methods. A particular case of the classification problems relates to the contexts where all observations have the same label (one-class classification) and will be presented separately due to its specificity.

► *Supervised learning*

As stated above, supervised learning methods seek to model a function $f(x; \theta)$, characterized by a set of parameters θ, that predicts Y as accurately as possible. In order to pick the best function f from a set of candidate functions $\mathcal{F}$ parameterized with θ, a *loss function* $L(f(x; \theta), y)$ quantifying the discrepancy between the predicted value and the actual output is necessary. Finally, the function $f(x; \theta^*)$ corresponding to the optimal parameter

set θ^* is the one that minimizes the expectation of the loss function, referred to as *true error*

$$\theta^* = \underset{\theta}{\operatorname{argmin}} R(\theta)$$

where $R(\theta)$ is

$$R(\theta) = \mathbb{E}\left[L(f(x,\theta),y)\right] = \int L(f(x,\theta),y)dP(x,y)$$

As $P(x,y)$ is unknown, $R(\theta)$ is replaced by the *empirical error* $R_{emp}(\theta)$ estimated on the given training data set $\mathcal{X}$

$$R_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i,\theta),y_i)$$

Typically, the evaluation of a model is performed on a set of observations not seen by the algorithm during training (referred to as *test set*) but generated by the same distribution as the training set. The algorithm is expected to perform well on the unseen data. In this case the model will be said to generalize well; otherwise it *overfits* the training set. The generalization ability of a model depends largely on the assumptions and choices made during the training. In particular,

1. The success of the model depends on how representative the training set is for the data distribution of the given task. If the training data set represents adequately the distribution generating it, it will be reasonable to expect the algorithm to perform well on previously unseen data. When the training data set is not a representative subset of all possible observations, the model trained on it will learn to approximate the given observations, only to perform poorly on observations different from the training set i.e. it overfits.

2. Different families of candidate functions may be selected when designing an algorithm. Each family provides functions with different properties. The choice of the candidate functions should therefore be in line with the known properties of the data set at hand.

3. Most families of functions come with hyper-parameters that need to be tuned to achieve the best performance. The hyperparameters are usually chosen among a wide range of values by optimizing the performance on a set of observations, different from both training and test sets, called *validation set*. Improper hyperparameter values may lead to overfitting.

It is common to constraint the considered family of functions by adding a *regularization* term constraining its structure in order to improve the generalization properties of the learning method. In this case, the empirical error is enhanced with an additional term:

$$R_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i,\theta),y_i) + \gamma \cdot \Omega(f,\theta) \tag{1.1}$$

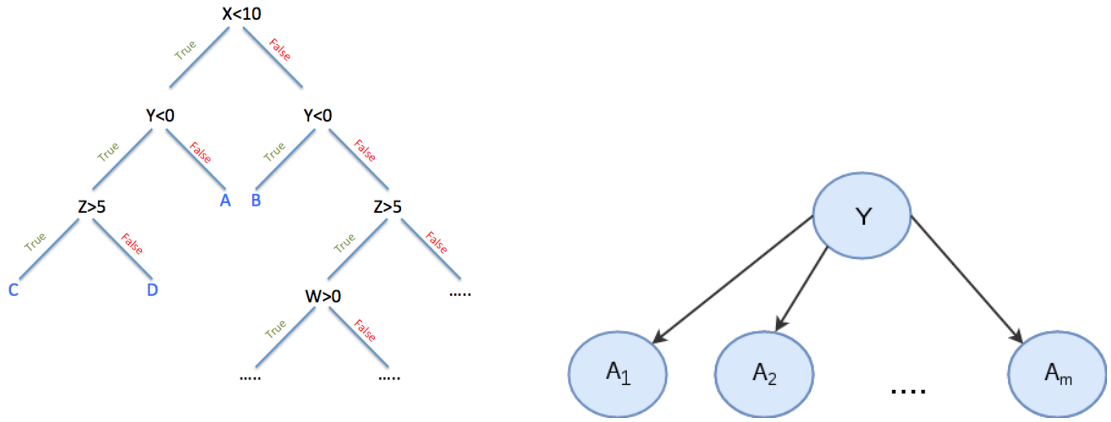


Figure 1.3: *(a)* A decision tree for a data set of 4 features - X, Y, Z and W, and 4 classes - A, B, C and D. *(b)* An illustration of Naive Bayes DAG for the variables $A_1, ..., A_m$ and class variable Y.

Below we briefly present some common supervised classification methods. A more complete description of the methods is included in the review by [Kotsiantis et al., 2007].

- **Decision trees**

Decision trees [Murthy, 1998] are structures that classify instances by consecutively checking the value of each of their features. Each node in a decision tree corresponds to a feature and each outgoing branch represents a possible value the feature can take on. The leaves of the tree correspond to the classes an instance is supposed to be categorized into. An optimal decision tree would contain the most discriminative feature at its root and each consecutive node would be assigned the most discriminative feature given its parent nodes. Building an optimal decision tree is a NP-complete problem and therefore heuristics are used in practice. An important component of building a decision tree is the choice of the metric with respect to which the features are chosen at each node. The most common of such metrics are information gain [Kent, 1983] and gini index [Breiman, 2017]. In order to prevent overfitting in decision trees, a strategy called *pruning* may be used that consists in disregarding the bottommost sub-trees in the built decision tree. Another practice is to employ an ensemble learning method, *random forest*, composed of multiple decision trees, that outputs a decision based on the outputs of all the trees in the forest e.g. with the majority voting. An example of a decision tree is shown on fig. 1.3a.

- **Bayesian networks**

A Bayesian network [Jensen, 1996] is a graphical model describing the relationships between the features/variables. The structure of a Bayesian network is given by a directed acyclic graph (DAG) where each node corresponds to a variable in the given data set. The (conditional) dependence/independence relationships between the variables are modeled with particular structures in the graph. The second component of a Bayesian network

is the conditional probability tables quantifying the relationships between each node and its parents. Learning a Bayesian network assumes learning the structure of the DAG and estimating the conditional probabilities (parameters). Learning the exact DAG structure requires an exhaustive search among a number of candidates, exponential to the number of variables. Methods based on greedy search have been proposed for practical uses [Chickering, 2002, Tsamardinos et al., 2006]. When the DAG structure is known (usually given by the experts), only the parameter estimation is necessary. The latter is usually achieved by maximizing the joint probability of the network. Naive Bayes is the simplest Bayesian network with a very primitive DAG composed of a single root (the class variable to predict) and its child nodes, conditionally independent of each other given the class variable, as illustrated on fig. 1.3b. This simple structure makes very strong assumptions on the relationships between the variables which is almost never true; however, it results in a simple expression of the joint probability and the estimation of the parameters becomes straightforward with Maximum Likelihood Estimation. The joint probability in Naive Bayes is given by

$$p(Y, A_1, A_2, ..., A_m) = p(Y) \prod_{i=1}^{m} p(A_i|Y)$$

where Y is the class variable, A_i is the i-th variable. The decision $\hat{y}$ for an example x is then given by

$$\hat{y} = \underset{k \in \{1,..,K\}}{\operatorname{argmax}}\, p(Y = k) \prod_{i=1}^{m} p(x_i|Y = k)$$

- **Instance-based methods**

Instance-based methods are a category of approaches that delay the model inference to the point of classification of the test data. This means that the heavy computation phase is performed not during the training (like for other approaches above) but over the test time. *k-Nearest Neighbour* algorithm [Cover and Hart, 1967] is the most popular method of this category. The main assumption is that data points located in close vicinity have common properties, such as belonging to the same class to predict. Therefore, the class of a test point can be determined by the known classes of the points surrounding it (typically with a majority voting). It is necessary to provide the number of points to consult for the decision - the parameter k of the model (hence, the name). It is also necessary to choose a proper distance metric in order to determine the k-nearest neighborhood of an observation (Euclidean instance being a common choice). The fact that to classify an instance it is necessary to evaluate its distance to all the other points (which should be stored for testing) makes the method less practical for large-scale problems.

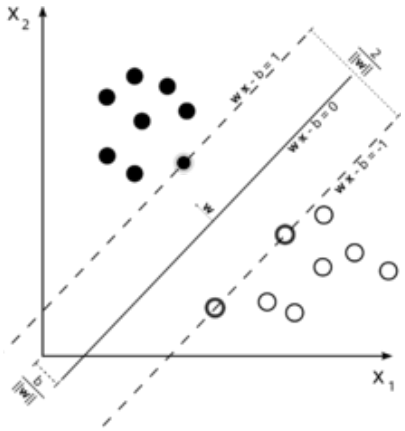


Figure 1.4: An illustration of SVM. The examples of the two classes are marked with black/white circles.

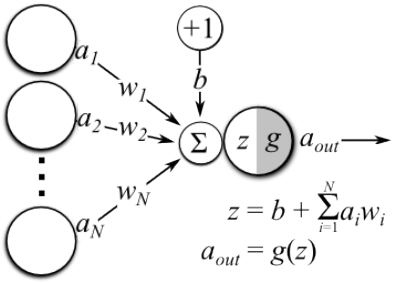
- **Support Vector Machines**

Support Vector Machines (SVM) are a very common supervised learning method introduced in [Vapnik, 1995]. SVMs seek to find a hyperplane separating the points of the two classes ($y_i \in \{1, -1\}$) in a way that maximizes the distance between the hyperplane and the closest point at either side of it (the distance is referred to as *margin*), as shown on fig. 1.4. The decision for an unseen example depends only on the linear combination of the points lying on the margin, called support vectors. To avoid the problem of misclassified training examples present in the data which do not allow the algorithm to find an optimal hyperplane, a *soft margin* formulation allows the misclassification of some examples, at a cost added to the term maximizing the margin. In this case using SVMs boils down to minimizing the following cost

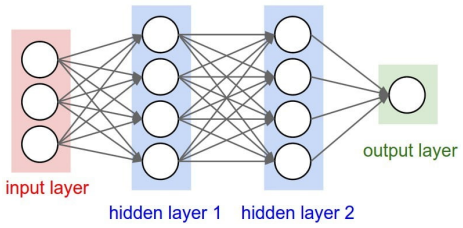
$$\underset{w,b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(w \cdot x_i - b)) + \lambda ||w||^2$$

where x_i is the i-th example, y_i is its corresponding class label, n is the number of examples, w and b define the hyperplane and λ is the tradeoff coefficient controlling the number of misclassified examples. The first term is the *hinge loss* and one can recognize the explicit form of the empirical error R_{emp} given in 1.1.

As the points may not be linearly separable in the original feature space, it is common to use the so called kernel trick to project the points to a higher dimensional space where the points are better separated. To use the kernel trick it is necessary to select the kernel function, the most common being radial basis function (RBF) kernel and polynomial kernel, each coming with hyperparameters to tune (more details on the kernel trick will be given in section 5.1.3). The SVM optimization problem eventually reaches a global minimum which has made it a very popular method. Some strategies have been proposed to extend the binary SVM for a multiclass classification, for example, through building SVMs for each class versus all the rest combined.



(a) A neuron, the computational unit in neural networks. w and b are the weights and bias associated with it, a is the input and g is a (usually non-linear) activation function.



(b) A simple artificial neural network with 2 hidden layers.

Figure 1.5: Artificial neural networks.

- **Artificial neural networks**

Artificial neural networks (ANN) are a category of methods vaguely resembling the neural network of animal brains. ANNs [Rosenblatt, 1958, LeCun et al., 1989] consist of computational units called neurons that together form a layer. A neuron is shown on fig. 1.5a. Layers may be stacked to form more complex structures as shown on fig. 1.5b. The neurons of one layer are connected to the neurons of the previous layer. The last layer in a typical neural network corresponds to the classification (less frequently, regression) task at hand. The structure particular to ANNs allows learning representations describing the input at different levels. So, the first layers usually capture more primitive patterns present in the input while the topmost layers model abstract representations. The main advantage of ANNs is that there is no need to gather relevant feature vectors to perform training; the relevant features are being learnt while training the network for the task at hand. Formally speaking, the connections between the units in the network are modeled with a (usually) non-linear function on top of a linear transformation of the incoming neurons. ANNs are the core of the so-called deep learning methods which will be presented in more details in the next chapters.

► ***Unsupervised learning***

The setup for unsupervised machine learning problems is similar to the one for supervised methods, the difference being that there is no output vector Y associated with the training examples. Unsupervised methods aim at learning some hidden structure in the data set that can be useful to discover and describe the tendencies and patterns for the given task and the data. It is more difficult to evaluate the quality of an unsupervised method, unlike in the supervised setting, where the true output is given and comparing it to the predictions is straightforward. The most common application of unsupervised learning is clustering.

Clustering

Clustering or cluster analysis algorithms seek to partition the given data points into cohesive groups of points similar or close to each other with respect to some measure, so called *clusters*. Clustering, therefore, may reveal interesting information on the structure of the data set. The most common clustering methods are listed below. A detailed review of clustering methods has been done in [Jain et al., 1999] and [Xu and C. Wunsch II, 2005].

1. *Hierarchical clustering*

   Hierarchical clustering methods aim at partitioning the data points based on their distance by building a dendrogram, a hierarchy of levels, each level yielding a set of clusters. The advantage of such methods is that cutting a dendrogram at different levels allows obtaining a certain number of clusters, without retraining the model. The construction of a dendrogram can proceed in either agglomerative or divisive approach. The former starts by considering each data point as a cluster and further recursively merges the current clusters by combining the closest pairs. The divisive approach starts with a single cluster containing all data points and recursively splits the current clusters into smaller ones by separating the farthest points. In either case a measure of distance between entities is necessary. Most common distance choices are implemented in *single linkage* clustering, *complete linkage* clustering, *average linkage* clustering and *Ward's* clustering [Ward Jr, 1963]. More hierarchical clustering methods are discussed in [Murtagh and Contreras, 2011].

2. k-means *clustering*

   k-means clustering [MacQueen et al., 1967] seeks to partition the data points into a set S of k clusters by minimizing the sum of squared error (SSE) criterion quantifying the within-cluster variance (the sum of the distances of the points to their cluster center) given by

   $$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - c_i||^2$$

   where c_i is the centroid of cluster S_i. It proceeds by randomly selecting k points as cluster centroids and further alternates between two steps - 1. assigning each data point to the cluster with the closest centroid and 2. updating the current cluster centers with the centroids of the newly found clusters. k-means is sensitive towards the initialization of the method, though multiple initialization approaches have been proposed as discussed in [Celebi et al., 2013]. The choice of the optimal cluster number k can be made by monitoring the changes in the SSE as the number k is varied through some range and pick the value resulting in minimal SSE or 1 standard deviation away from it (elbow method).

3. *Mixture model clustering*

   Mixture model based clustering assumes the cluster to be a random variable z of K possible values whose prior distribution $p(z)$ and the conditional probability function of an example x given the cluster variable $p(x|z)$ are known. Gaussian mixture models (GMM) is the most common configuration of such approaches where $p(z)$ is categorical/multinoulli distribution i.e. $p(z = k) = \pi_k, \sum_{k=1}^{K} \pi_k = 1$ and $p(x|z)$ is a Gaussian distribution with a mean vector μ_k and covariance matrix Σ_k i.e. $p(x|z = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. The GMM training consists in estimating the parameters of those distributions which is usually done with maximum likelihood estimation, yielding the parameters that maximize the joint probability over the entire data set, given by

$$L(\theta_1, ..., \theta_K; \pi_1, ..., \pi_K|X) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i; \theta_k)$$

   where $\theta_k = (\mu_k, \Sigma_k)$. The most commonly exploited method to do so is the Expectation - Maximization (EM) algorithm [McLachlan and Krishnan, 2007]. The algorithm proceeds in repeating two alternating steps - 1. e-step: compute the expectation of the complete data log-likelihood, 2. m-step: select new parameter estimates maximizing the previously computed function (k-means is a particular case of the EM algorithm). The posterior probability $p(z|x)$ of a cluster given an observation is then calculated with the Bayes' theorem using the estimated parameters.

For most clustering methods, it is required to decide upon the desired number of clusters K. The choice may be intuitive for some small-scale and relatively simple tasks but in the vast majority of real life problems, the choice of K is not trivial: small Ks may give little information on the structure of the data while greater values may reduce the interpretability. Several approaches exist to select an optimal K. In some cases a simple 2-dimensional visualization of the data points can give an idea of the order of K. For GMM, the number of clusters can be chosen among a range of values as the one minimizing the Akaike's information criterion (AIC) [Akaike, 1974] or maximizing the Bayesian inference criterion (BIC) [Schwarz et al., 1978]. The Gap statistic [Tibshirani et al., 2001], that compares the total within-cluster sum of squares for different values of K with their expected values under a null reference data distribution, can also be used, especially for k-means clustering.

► ***Outlier detection***

One-class classification problems are a particular case of classification problems where the training set contains examples of only one class and the aim is to later identify the representatives of that class. An important application of one-class classification problems is the so called outlier detection. Outlier detection methods, also known as anomaly detection, novelty detection, seek to distinguish outliers from the normal examples constituting the given data set. All the examples of the given training data set are *normal/positive* and

hence have the same label; the label however is not informative and does not appear anywhere in the problem formulation, which is why outlier detection problems are usually seen as unsupervised problems. The definition of an outlier may vary across different domain applications and data sets. Essentially, outliers are examples that do not quite fit to the characteristics of the *normal/inlier* observations. More precise definitions may apply. So, [Hawkins, 1980] defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. [Johnson and Wichern, 1992] defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. A more thorough overview is presented in [Ben-Gal, 2005].

Over the recent years the topic of outlier detection has been studied extensively in different application domains and many algorithms have been proposed for outlier detection depending on the nature of the data and the type of anomalies [Hodge and Austin, 2004, Chandola et al., 2009, Pimentel et al., 2014, Kiran et al., 2018]. Outlier detection methods have been applied in various domains and applications, including credit card fraud detection [Aleskerov et al., 1997], mobile phone fraud detection [Barson et al., 1996], network intrusion [Lazarevic et al., 2003], fault identification [Diaz and Hollmén, 2002], etc. We identify 5 major categories of the existing outlier detection methods: (1) probabilistic, (2) distance-based, (3) reconstruction-based, (4) domain-based and (5) information-theoretic. Below we present a brief summary of the assumptions and main characteristics of these methods. More comprehensive reviews have been conducted in [Pimentel et al., 2014, Chandola et al., 2009].

1. ***Probabilistic methods***

   Probabilistic outlier detection methods aim at estimating the underlying probability density function generating the data and declaring outliers as data points largely deviating from the estimated density function. The main assumption of such methods is that outliers correspond to low density areas whereas inliers are concentrated in the high density areas of the underlying distribution. Two major categories of probabilistic methods exist: *parametric* and *non-parametric*.

   Parametric methods assume that the data has been generated by a parametric distribution with a parameter set Θ; Θ is estimated using the available data. Naively, the inverse of the probability function for a test point can be interpreted as a measure of its anomalousness. Other approaches are based on statistical hypothesis tests that consider the null hypothesis as the event that a given point has been drawn from the estimated distribution. When the null hypothesis is rejected, the point is considered an outlier. For example, Grubb's test [Grubbs, 1969] assumes the training set was generated by a Gaussian distribution and treats the points whose distance from the estimated mean is larger than a certain threshold as outliers. It was designed

for univariate analysis; its alternative versions were proposed in [Laurikkala et al., 2000, Aggarwal and Yu, 2001]. A more sophisticated approach chooses to model the underlying data distribution as a mixture of parametric distributions. For example, Gaussian Mixture Models (GMM) are a very popular method based on modeling a mixture of Gaussians and declaring a test point an outlier when it does not seem to be generated by any of them. The most common technique to estimate the parameters for this approach is maximum likelihood estimation, in particular, the EM algorithm. The successful application of this category of methods depends on how well the basic assumptions on the underlying distribution correspond to the data at hand and how much data is available to approximate well the distribution in question.

Non-parametric methods do not assume any a priori form for the underlying density function; it is determined from the training set. One simple non-parametric approach is the histogram-based method that "learns" the shape of the density function by constructing a histogram. It is done by defining bins and counting the number of data points falling into each of them for a particular feature. Once the histogram is built, a test point is considered an outlier if it is not included in any of the bins. The key point here is to choose the bin size. The main drawback is that the histograms are built for each individual feature; discriminating outliers by taking into account the (possibly) complex interactions between the features may not be achieved. Kernel density estimator, also known as Parzen window method [Parzen, 1962], is another common method of density estimation. It places a Gaussian distribution centered at each of the training examples and computes their linear combination. The only free parameter of this technique is the kernel width which controls the smoothness of the learnt distribution. The outliers are then detected as the points located in the low-density areas of the estimated distribution.

2. ***Distance-based methods***

Distance-based outlier detection methods adopt the assumption that outliers are rather isolated data points and, therefore, their distance to their neighbors should be indicative of their anomalousness. k-nearest neighbor [Altman, 1992] is one of the approaches that considers a data point an outlier if its distance to its k-th neighbor is larger than a chosen threshold. A different category of methods consider the local density in the neighborhood of a point as an indicator of anomalousness. For instance, [Breunig et al., 2000] proposed a metric called Local Outlier Factor (LOF) defined as the ratio of the average density of the areas around the k nearest neighbors of a given point and the local density of the area around the point itself. In the LOF method, the local density, more precisely, the local reachability density, of a point is defined as the inverse of the average reachability distance of the point from its neighbors. Another approach called Local Outlier Probabilities (LoOP) combines the

main idea behind LOF with a probabilistic component to model the anomalousness of an example [Kriegel et al., 2009]. Unlike the nearest neighbor based methods that make assumptions on the distance of a given point from its neighbors, *cluster-based* outlier detection methods make assumptions about the distance of the point to its closest cluster center. Basically, such an approach involves clustering the training data using any clustering method (k-means, for example) and declare outliers based on their distance to the found clusters. This method would not suit to the cases where outliers themselves are numerous and close enough to form a cluster of their own.

3. ***Reconstruction-based methods***

   Reconstruction-based outlier detection methods assume a model able to reconstruct a given input after having it transformed to some intermediate representation. The deviation between the reconstruction and the original input is then considered indicative for the detection of outliers. Two major categories of such methods are subspace-based and neural network based algorithms. Principal Component Analysis (PCA) is a popular subspace-based method [Jolliffe, 2011] that seeks to project the original observations of correlated variables to a space defined by uncorrelated variables (principal components) that maximize the variance of the data. It has been widely used as a dimensionality reduction technique (convenient as the method seeks to project the original data points to a space where the desired amount of the information can be preserved by keeping a certain number of principal components). The transformation applied to the original input may be interpreted as encoding and may be used to obtain a *reconstruction* of the input: the deviation of the reconstruction from the input can be used to indicate outliers. Kernel PCA [Schölkopf et al., 1997] extends the original formulation by first performing a mapping through a non-linear kernel into a higher-dimensional space and then applying the PCA. PCA-based outlier detection was applied in [Huang et al., 2007, Brauckhoff et al., 2009, Choi et al., 2005]. Neural-network based anomaly detection methods will be reviewed in chapter 6.

4. ***Domain-based methods***

   Domain-based methods, unlike those that try to estimate the density of the given normal data points, seek to find their boundary. In this case the decision on the anomalousness of a given point depends on which side of the boundary the point is located. One class SVM (oc-SVM) [Schölkopf et al., 2001], a particular case of SVMs, is one of the most common methods of the category. It seeks to find a hyperplane with maximum margin separating the normal points from the origin. The kernel trick is usually applied and the separation is performed in a higher-dimensional space.

Support Vector Data Description (SVDD) method [Tax and Duin, 2004] is formulated similarly to the oc-SVM method and aims at finding the hypersphere with minimum volume containing the normal points. A point laying outside of it is considered an outlier. The oc-SVM algorithm will be described in details in section 5.1.3.

5. ***Information-theoretic methods***
   Information-theoretic methods are based on the idea of quantifying the information content of a data set as measured with entropy, conditional entropy, relative conditional entropy, information gain, information cost or other similar measures. The main assumption of these methods is that the presence of outliers changes dramatically the information content of a data set. Therefore, removing the outliers will have a significant impact on the measure of choice (e.g. entropy). Information theory-based methods seek to identify outliers as a subset of data points whose removal from the data set results in the largest change of the chosen measure as compared to the removal of the rest of the data points. [He et al., 2006] proposed a greedy algorithm (*Local Search Algorithm*) that iteratively labels the point with the largest entropy decrease among the points currently labeled as normal, as outlier, until the number of outliers reaches k, a preselected value. Choosing the value for the parameter k may not be intuitive in many applications and is a disadvantage of such an approach.

1.2 Performance evaluation of CAD systems

One crucial step of any CAD system is the performance evaluation. Evaluating a CAD system means assessing its *generalization* ability i.e. the capacity to perform the task of interest on previously unseen data. This amounts to estimating the expected prediction error (*test* or *generalization* error) at some new point. While annotated data may or may not be used during the model training phase, a typical CAD system evaluation involves comparing the output of the system to a reference. We will therefore focus on the evaluation protocol where the reference or ground-truth annotations are given for the evaluation. Formally, for a given training data set $D = \{(\mathbf{x}_i, t_i)\}$ for $i = 1, .., N$, a trained model predicts an output y via a function f i.e. $y = f(\mathbf{x})$. Assuming the optimal prediction is given by y^* and $L(t, y)$ is a loss measuring the discrepancy between t and y, the expected prediction error can be shown to have the following decomposition for some learning algorithms [Domingos, 2000]

$$\mathbb{E}[L(t, y)] = c_1\mathbb{E}[L(t, y^*)] + L(y^*, y^m) + c_2\mathbb{E}[L(y^m, y)]$$

where y^m is some central tendency of the learnt model. The test error decomposition was first derived for regression, with L being the squared loss and $c_1 = c_2 = 1$ (full derivation in [Hastie et al., 2009]). For a binary classification task, when L is the 0-1 loss, the expression above holds for certain values of c_1 and c_2. A closed-form expression exists in particular

for k-nearest neighbor algorithm. The main objective of the decomposition above is to separate the three terms:

1. the first term is the irreducible error, independent of the learning algorithm and hence beyond our control

2. the second term is the *bias* of the learning algorithm, the difference of its predictions and the optimal predictions. The bias is 0 for a model always making optimal predictions

3. the third term is the *variance* of the learning algorithm quantifying the differences around the central tendency. A high difference between the average predictions on the training and test sets is indicative of high variance.

Good learning algorithms therefore are characterized with low bias and low variance. Finding the bias-variance tradeoff is an important aspect of selecting a model.
In the clinical settings, estimating the generalization error usually amounts to testing the CAD system on a new set of patients, not used for the training of the CAD system. The first aspect to consider is therefore a strategy to split the data into sets that would be used for training and later for evaluation. Eventually, an appropriate metric should be chosen to quantify the performance on the unseen data set. Below we discuss these two aspects in details.

1.2.1 Data splitting strategies

An essential part of building an automated system is to decide upon the *training set*, *validation set* and *test set*. The training set, as the name suggests, is composed of the observations which are used for the training. The validation set comprises the observations not used for training but essential for the so-called *model selection*. Model selection consists in retaining a configuration, among a set of possibilities, that gives an acceptable performance on the validation set when trained on the training set. Frequently model selection boils down to selecting hyperparameters for an algorithm. Moreover, evaluating the system on the validation set creates a feedback that can be used to improve the current settings. Once the model is chosen and trained, it can be evaluated on a previously unused test set. The evaluation should be final and should not be used to modify the model afterwards. There exist several strategies of splitting the given data set into training, validation and test sets, depending on the availability of the data. When the overall data set contains a substantial number of observations, the most straightforward approach is a basic split into three distinct subsets, the training set typically being the largest. In many problems, especially in the medical domain and therefore in CAD development, the available data sets are scarce in the number of examples. For such cases, alternative strategies have been developed.

Cross validation is a common approach, especially applied when the data set is small. The data set is divided into k disjoint subsets of (roughly) equal size, hence the name of the strategy - *k-fold cross validation*. For each of the fold, the remaining $k-1$ folds together are used as a training set while the current fold itself is used as a test set. The overall performance of the model is estimated as the average across all the folds. It is also common to tune model parameters by computing the average cross validation error for all the parameter configurations and pick the one with the minimum error (or the one at a certain distance from the minimum). The extreme case of k-fold cross validation is N-fold cross validation (also known as *Leave-One-Out* LOO validation) which will naturally result in the best trained model (as a maximum amount of data is assigned for training) with approximately unbiased estimation of the test error. It is, however, the slowest of all and has a large variance [Hastie et al., 2009] because of the significant overlap between the training sets in each fold. $k = 5$ and $k = 10$ are the most common choices [Breiman and Spector, 1992]. In CAD evaluations, it is common to perform *Leave-One-Patient-Out* LOPO evaluation i.e. testing on a single subject using the model trained on all the remaining subjects. Nested cross validation [Varma and Simon, 2006] is a more advanced cross-validation strategy. It consists of performing two loops - an outer and an inner. In the outer loop the data is split into a model selection set and a test set while the inner loop further splits the model selection set into training and validation sets.

Bootstrap [Efron and Tibshirani, 1997] is another method for estimating the error of a model. For a data set of N examples, the bootstrap method constructs B data sets each of which is built by randomly drawing the original examples from the data set with replacement. Models trained on each of B bootstrapped data sets can later be evaluated on the original data set and their average error could be used as an estimate of the model performance. The estimate would not be accurate since the test set and the training set are not disjoint and may have a significant overlap. Another option is to evaluate the performance of a bootstrapped data set B_i on the examples of the original data set, not included in B_i. This modification provides a more accurate estimate.

The performance of a CAD system can be measured with different metrics depending on the problem it is designed for. Below we consider the appropriate measures for CADs for classification.

1.2.2 Performance metrics

The most typical method of quantifying the results of a CAD system for binary classification, distinguishing positive examples (P) from negatives (N), is the *confusion matrix*. A confusion matrix is shown in 1.1. The diagonal of the matrix corresponds to the number of positive and negative examples classified as positive and negative, respectively. FN is the number of examples of class P (positive examples) classified as N (negative examples)

and FP is the number of examples of class N classified as P. A confusion matrix can be constructed for an arbitrary number of classes in a similar fashion. Various performance metrics can be computed on a confusion matrix. Below we present the most common ones.

		Predicted class	
		P	N
True class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Table 1.1: A confusion matrix for a classification system for two classes P and N.

1. **Accuracy**: The proportion of the observations correctly classified by the classifier. Accuracy is not always indicative of the quality of the classifier since it is influenced significantly by the majority class when the classes are largely imbalanced.

$$Acc = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

2. **Sensitivity** or **True positive rate** or **Recall**: The proportion of the positive examples correctly labeled positive.

$$TPR = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

3. **Specificity** or **True negative rate**: The proportion of negatives classified correctly.

$$TNR = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

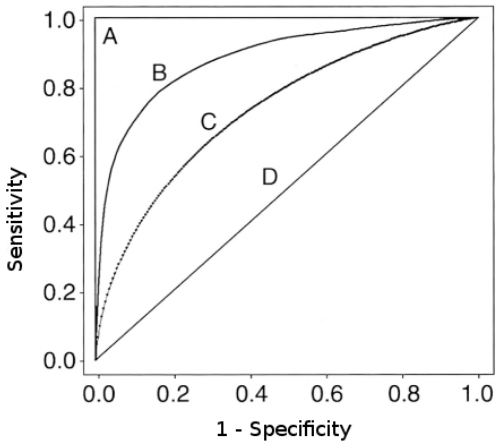
4. **Precision**: The proportion of examples labeled positive, actually being positive.

$$Precision = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

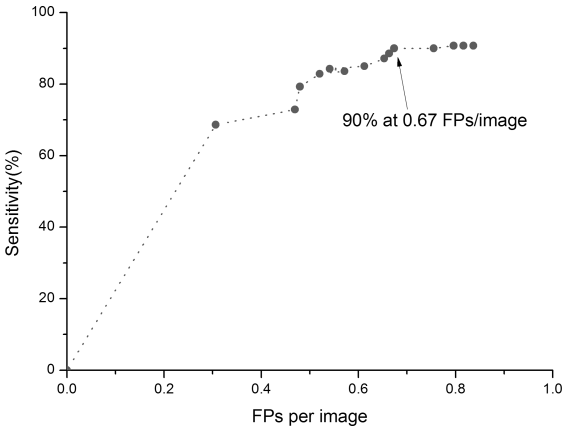
5. F_1**-score** or **Dice similarity coefficient (DSC)**: The harmonic mean of precision and recall. Provides a more balanced performance metric than the accuracy.

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

As mentioned above, when the problem at hand is characterized by a large class imbalance (common in the medical classification problems where healthy examples outnumber pathological cases), accuracy may not be an informative measure of performance. Alternatively, a tradeoff between the sensitivity and specificity is sought as a more appropriate



(a) Examples of ROC curves. Each point on the curve corresponds to a pair (sensitivity, 1-specificity). The curves A and D showcase the best and the worst possible performance, respectively. The CAD system corresponding to the curve B outperforms that of C.



(b) An example of fROC curve. Each point on the curve corresponds to a pair (sensitivity, average number of false positive detections per image).

indicator of the model performance. Such tradeoff is frequently found through the so-called **receiver operating characteristic** (ROC) curve [Metz, 1986]. Assuming the classifier outputs a numerical value before converting it to a label, a ROC curve can be obtained by varying the threshold upon which the prediction labels would be assigned, and tracing the sensitivity-specificity curves for all the thresholds. Each point on a ROC curve corresponds to a couple (Sensitivity, 1-Specificity) obtained for a given threshold and is called an operating point of the system. A ROC curve is illustrated on fig. 1.6a. The curve A corresponds to the perfect scenario where both sensitivity and specificity are optimal. D is the worst case as it has approximately the same performance as a classifier assigning random labels to observations. The 2 other curves show reasonable performance levels. ROC curves are usually quantified with a single value - the *Area Under the Curve (AUC)* which gives an indication of how good of a curve a system has. Two curves, however, may have identical AUCs but different sensitivity/specificity values. Comparing different systems based solely on their AUC values may be misleading and it is therefore important to analyze properly the ROC curves themselves [Park et al., 2004].

In medical applications, in particular those seeking to detect lesions in given images and considering a subject pathological when at least one lesion is found, it is common to use another characteristic curve quantifying the performance of the method. Free receiver operating curve (fROC) illustrates the relationship between the sensitivity of the method and the number of false positive findings per image [Bunch et al., 1978]. Fig. 1.6b shows an example of a fROC curve where a given point on the curve stands for the sensitivity of the y-axis for the corresponding average number of false positive detections per image on the x-axis. Such analysis is especially useful when, for one reason or another, the specificity of the method cannot be measured. One disadvantage of fROCs is that the x-axis does

not have an explicit upper bound which makes it impossible to compute such quantities as AUC for ROCs.

In order to calculate the performance metrics mentioned above, it is necessary to define the notion of TPs and FPs. The definition may vary between applications, depending on the choices made by the CAD developers and the problem at hand. For instance, one may define TPs and FPs at voxel level, when voxels correctly labeled by the system are considered TPs while the misclassified voxels constitute the FPs. This definition is more suitable for segmentation problems. On the other hand, in detection problems it is common to define TPs/FPs at detection/cluster level (clusters refer to neighborhoods of voxels identified by the system with respect to the problem). In this case, when a detected cluster coincides with the ground truth following some rule [Petrick et al., 2013], it is considered a TP and a FP otherwise. The choice of such a rule, naturally, affects the CAD performance measure and should be done accordingly. For patient-level CADs, the TPs and FPs may refer to patients / healthy controls identified by the system as pathological, respectively.

Chapter 2

Deep learning in medical applications

The concept of artificial neural networks, briefly introduced in section 1.1.3, can be traced back to the 40s and 50s when the first learning algorithms for rather shallow networks were proposed [Rosenblatt, 1958]. Inspired by an earlier idea of [Fukushima and Miyake, 1982], [LeCun et al., 1989] developed the first convolutional neural networks (CNN). These networks, however, remained rather unpopular over the next decade. The various techniques developed over the next years for the training of deep architectures [Vincent et al., 2010, Nair and Hinton, 2010, Srivastava et al., 2014, Ioffe and Szegedy, 2015], together with the advances in computing power, including the exploitation of graphics processing units (GPUs), paved the way for the successful application of deep architectures in large-scale real-life problems. AlexNet in [Krizhevsky et al., 2012] was a milestone contribution that allowed for deep architectures to rapidly make their way into the computer vision domain and later, the medical imaging domain. AlexNet outperformed the rest of the approaches (using handcrafted features) of the ImageNet challenge by a very large margin. Recently, more advanced architectures have been proposed for the same challenge [Simonyan and Zisserman, 2014, Russakovsky et al., 2015, Szegedy et al., 2015, Szegedy et al., 2017], making the Convolutional Neural Networks the method of choice in the computer vision community. The latter architectures and their derivatives have been applied in various applications (not necessarily aligned with the one they were originally designed for), including object-detection [Girshick et al., 2014], semantic segmentation [Long et al., 2015], video classification [Karpathy et al., 2014] and super-resolution [Dong et al., 2014].

The success of the neural networks therefore increasingly ignited an interest in the medical imaging community where many problems leveraged the potential of deep architectures. An important advantage of deep learning methods lays in the fact that the model training

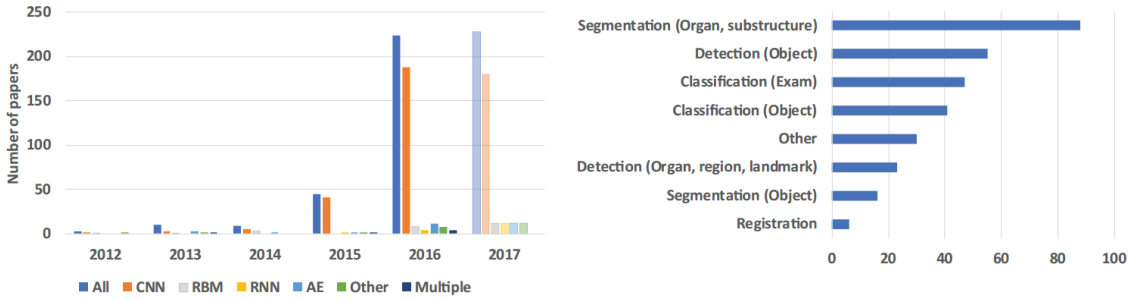


Figure 2.1: Left: Overlook on the publications in medical imaging using deep learning up to 2017. Right: The most common medical applications of deep learning methods. *Illustrations from [Litjens et al., 2017].*

is accompanied with an implicit representation learning in a data-driven manner. This means that the features learnt from the data are immediately relevant to the task the architecture is designed for. This offers an effective alternative to the feature extraction routine described in section 1.1.2. [Shen et al., 2017] and [Litjens et al., 2017] presented comprehensive reviews of deep learning in the medical domain.

In this chapter we review the medical applications that have benefited from the introduction of deep learning methods. We further focus on the deep approaches developed for neuroimaging data which is the main focus of this work.

2.1 Deep learning in general medical applications

As reviewed by [Litjens et al., 2017] and shown on fig. 2.1, the number of publications in the medical imaging domain exploiting different types of deep learning architectures has grown significantly over the last few years. Fig. 2.1 also shows the various medical application contexts where deep learning methods have been integrated successfully. In most scenarios, such methods outperformed the conventional ones used previously and for many medical tasks, they have become the main method of choice.

The medical problems tackled with deep architectures include various pathologies and tasks. For instance, lung nodule detection is one of the problems where the methodological shift towards deep architectures is remarkable. [Kumar et al., 2015] first learned features with an autoencoder and then applied a binary decision tree on the feature vectors for lung nodule classification on CT scans. [Hua et al., 2015] used both Deep Belief Networks and CNNs for the same task. [Shen et al., 2015] performed lung nodule classification on CT scans by combining the features output by three CNNs with patches of different scales at input. [Dou et al., 2017] proposed a multiscale 3D convolutional network enhanced with contextual information that is especially designed to reduce the false positive rate in pulmonary nodule classification.

Breast cancer diagnosis is another important application where deep learning based methods have been used recently. [Wang et al., 2016] exploited several pretrained architectures

on breast images to classify patches as normal versus cancerous. [Kooi et al., 2017] applied a CNN enhanced with various additional features for the same problem. Similarly to other works for the same medical task [Fotin et al., 2016], the method first performs a candidate patch selection and then proceeds to the architecture training. [Kisilev et al., 2016] proposed a CNN-based architecture trained to generate regions of interests (ROI) surrounding suspicious areas which where then passed to the next layers, by so avoiding the explicit candidate selection step.

Several recently proposed approaches targeted cardiac segmentation. So, [Poudel et al., 2016] proposed a U-Net-like recurrent fully-convolutional network on 2D slices, that leverages inter-slice spatial dependencies through internal memory units, to segment the left ventricle. [Kong et al., 2016] combined a 2D CNN and an LSTM to perform temporal regression in order to identify specific frames and a cardiac sequence.

[Zhu et al., 2017b] proposed an elaborate version of the U-Net architecture [Ronneberger et al., 2015] for prostate segmentation. Another method based on a modified version of U-Net enhanced with residual connections was proposed for prostate segmentation in [Yu et al., 2017].

Organ detection and landmark localization are yet another area where substantial efforts have been made with deep networks. For organ localization, [de Vos et al., 2016] used the pretrained AlexNet to classify 3 regions of interests (ROI), i.e. heart, aortic arch and descending arch, on three axes separately and predicted the 3D rectangular boxes containing the ROIs. [Cai et al., 2016] employed Convolutional Restricted Boltzmann Machines on multiple modalities for vertebrae recognition. [Payer et al., 2016] proposed to regress heatmaps for landmarks instead of their absolute coordinates with an end-to-end training of a CNN on hand radiographs.

All the applications above showcase the potential of deep learning in various medical contexts. Below we will focus on one of the most popular set of medical tasks consisting of the detection of various brain pathologies.

2.2 Deep learning for pathology detection on neuroimaging

In neuroimaging, many different medical problems, targeted with deep learning methods, can be identified. The most common of those include registration, segmentation and detection tasks. In brain segmentation problems, the objective is to classify the voxels into a number of categories for which, typically, a voxel-level annotated data set is available for training. As such, the brain tumor segmentation has become especially popular with the MICCAI Brain Tumor Segmentation (BRATS) challenge [Menze et al., 2015]. In the scope of this challenge the clear tendency of exploiting convolutional neural networks has emerged by outperforming the previous successful methods such as random forests as in

[Goetz et al., 2014, Kleesiek et al., 2014]. The rather first attempts of using fully convolutional networks for brain tumor segmentation by [Zikic et al., 2014, Pereira et al., 2015] have gradually evolved into more elaborate models such as cascaded multi-path convolutional networks in [Havaei et al., 2017] and dual pathway deep 3D convolutional network in [Kamnitsas et al., 2017] with Dice similarity coefficient reaching 89.8%.

This work, however, is focused on detection problems in neuroimaging. While segmentation tasks operate in contexts where precise discrimination of tissue types is sought, detection problems emerge in scenarios where pathologies are usually subtle and may not be easily identified and contoured by a human expert. Therefore, the evaluation metrics for segmentation and detection problems are different. While segmentation methods may leverage a large range of metrics suitable for classification, as described in section 1.2, evaluating a detection system depends on the difficulty of the task and the desired level of granularity (voxel-level detections, subject-level detections, etc). Below we review some important detection problems in neuroimaging.

2.2.1 Supervised brain pathology detection

A number of studies tackled the detection of brain pathologies as a segmentation task. In this case, a voxel-level annotated training data set is available and the problem is cast to a classification problem with an appropriate performance evaluation. Several neurological pathologies are characterized by small lesions of various shapes, localizations and spatial patterns. Small vessel disease (SVD), for instance, usually is a result of small vessel abnormalities and has various imaging biomarkers such as lacunes, white matter hyperintensities, microbleeds, perivascular spaces and brain atrophy [Wardlaw, 2008]. Such abnormal lesions, having a diameter inferior to 2mm for the smallest perivascular space lesions and up to 20mm for the largest observed lacunes, may lead to stroke, cognitive impairment or dementia [Wardlaw et al., 2013]. Intracranial carotid artery calcification (ICAC) is another example of small entities that can be identified on plain head computed tomography (CT). ICAC is a marker of Intracranial arteriosclerosis which represents a major cause of stroke [Bos et al., 2014] and might contribute to the risk of cognitive impairment and dementia [Bos et al., 2012]. Multiple sclerosis (MS) lesions constitute another type of brain lesions of varying size and time evolutive pattern. The characterization of lesion profiles, including brain lesion load, as well as the temporal detection of appearance of new lesions are crucial to perform an early diagnostic, define and monitor the optimal therapeutic strategy [Filippi et al., 2016].

Recently, many automated detection methods for such pathologies have been proposed. For MS lesion segmentation, [Valverde et al., 2017] proposed a cascaded 3D convolutional neural network approach that consists in producing an intermediary lesion probability map with a convolutional neural network (CNN) and then feeding it into a second network that reduces the number of false positive detections. When T2-w, T1-w and FLAIR

were combined at input, the approach was ranked first among 60 candidate methods on the MICCAI2008 challenge, including the 3D convolutional encoder network with shortcut connections and two interconnected pathways proposed by [Brosch et al., 2016]. [Havaei et al., 2016] applied a CNN-based framework compensating the effect of missing modalities to the MS segmentation context. [Ghafoorian et al., 2017a] proposed a CAD system for the detection of lacunes of presumed vascular origin, consisting of two components - a first fully convolutional network detecting candidates and a second 3D convolutional network trained to discriminate true detections versus false positives. The CAD system achieves a sensitivity of 0.974 with 0.13 false positives per slice. [Dou et al., 2016] applied a 3D convolutional network for cerebral microbleed detection. Several approaches exist for white matter hyperintensity segmentation. For example, [Ghafoorian et al., 2016] proposed a framework where patches are selected in a non-uniform manner and later compared three convolutional architectures - 1. single-scaled, 2. multi-scaled with early fusion and 3. multi-scaled with late fusion. In [Ghafoorian et al., 2017b] this work has been extended to incorporate location information which resulted in a substantial performance gain. [Bortsova et al., 2017] proposed a deeply supervised dropout network for the segmentation of ICAC lesions and achieved a DICE score of 76.2% between the predicted ICAC lesions and the manual annotations (the FPR was not reported).

2.2.2 Unsupervised brain pathology detection methods

Although the deep architectures described above achieve impressive results compared to the state-of-the art methods, they require annotated data sets for training and do not necessarily solve the problem of false positive detections. The nature of brain pathologies, however, is highly variable and well-annotated data sets with an adequate representation of different cases are not always available.

In order to bypass the need of voxel-level annotated data sets, some authors recently proposed to formulate subtle lesion detection tasks in *semi-supervised* or entirely *unsupervised* settings. The number of such works, in general, and for brain pathologies, in particular, is by far inferior to that of supervised methods. However, the interest towards this category of methods has been growing over the recent years and the first works show significant potential. [Baur et al., 2017] introduced a framework accounting for both labeled and unlabeled data in a deep architecture for MS lesion segmentation. In [Dubost et al., 2017] the authors exploit weak labels (the number of lesions in a scan) in an architecture called GP-Unet to detect enlarged perivascular spaces in the basal ganglia. The network solves a regression problem and outputs a number of detected lesions via a U-Net-like convolutional pathway, as shown on fig. 2.2. [Shah et al., 2018] proposed to exploit a deep autoencoder whose middle layer representation is plugged into a supervised extension that discriminates inliers vs. outliers using a small amount of annotated data. The method was evaluated,

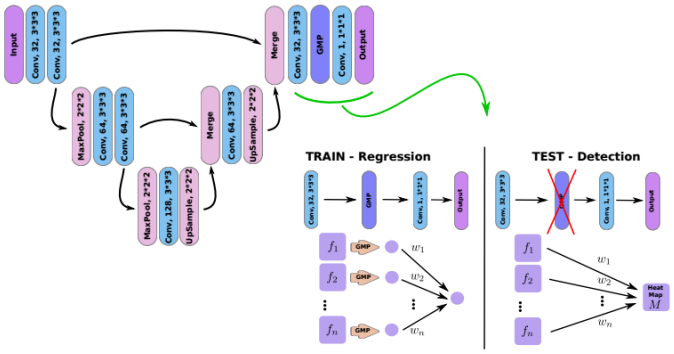


Figure 2.2: GP-Unet, a deep architecture with weak labels (the number of lesions in a scan) to detect enlarged perivascular spaces in the basal ganglia. *Illustration from [Dubost et al., 2017].*

among others, on three biomedical data sets, including the BRATS17 data set.

Some works go even further and propose to treat lesion detection tasks as outlier detection problems in fully *unsupervised* contexts. [Chen and Konukoglu, 2018] recently proposed a constrained adversarial autoencoder trained on a set of non-pathological brain images. The approach extends the original formulation of the Adversarial Autoencoders (AAE) with an additional term imposing consistency in the learnt representation space. The evaluation, however, was performed on some test cases from the BRATS data set, which is not a typical detection problem.

The deep learning architectures have become common methods in various medical applications and, particularly, in neuroimaging, as can be seen from the studies mentioned above. A significant number of the existing works tackle segmentation problems with the majority of the approaches developed in supervised settings on voxel-level annotated data sets. The limited access to such labeled data sets, with a sufficient number of representative examples, has motivated an increasing interest in weakly-supervised and unsupervised methods. The studies developed in such settings present interesting methodological solutions and show a promising performance.

Chapter 3

CAD systems for epilepsy detection in neuroimaging

3.1 Epilepsy description

Epilepsy is one of the most common neurological disorders affecting around 50 million people worldwide according to the World Health Organization (WHO). It is characterized by an enduring predisposition to generate unprovoked brain seizures [Fisher et al., 2014]. Epilepsy treatment involves consistent intake of antiepileptic drugs on a long-term basis which allows to control the seizures for up to 70% of focal epilepsy patients; the remaining 30%, however, do not respond to pharmacotherapy and are referred to as *medically refractory/drug-resistant/intractable epilepsy* patients [Kwan and Brodie, 2000]. The two most common medically refractory epilepsy types are temporal lobe epilepsy and focal cortical dysplasia (FCD) [Lerner et al., 2009]. In both cases, surgical removal of the epileptogenic lesions is the most effective treatment that may offer a seizure-free life.

Temporal Lobe Epilepsy

TLE is the most common form of epilepsy with focal seizures originating in the temporal lobe. TLE is commonly a result of mesial temporal sclerosis (MTS). Temporal lobectomy is one of the main surgical approaches for TLE. Hippocampal sclerosis (HS) is a common pathology encountered in mesial temporal lobe epilepsy, characterized with severe neuronal cell loss and gliosis in hippocampus. According to a recent consensus classification system [Blümcke et al., 2013], 3 types of HS are distinguished, further categorizing types 2 and 3 as *atypical* and type 1 as *classical* HS. The new classification is based on the patterns of neuronal loss and gliosis as measures of sclerosis.

Malformations of cortical development

Medically refractory epilepsy is often associated with malformations of cortical development (MCD), a variety of structural and metabolic abnormalities of brain arising during gestation, present in up to 40% of drug resistant epilepsy patients [Guerrini et al., 2003].

Focal cortical dysplasia (FCD) is one of the most common MCDs in epilepsy patients. It is the first/third most frequent cause of epilepsy in children and adults, respectively [Lerner et al., 2009]. ILEA (International League Against Epilepsy) consensus classification differentiates three types of FCDs [Blümcke and Spreafico, 2011]. Type I refers to FCDs with abnormal cortical lamination, further divided into subtypes Ia (radial cortical lamination, mostly located in temporal lobes), Ib (tangential 6-layer cortical lamination) and Ic (radial and tangential cortical lamination) [Blümcke et al., 2011]. Type II FCDs are characterized with a presence of dysmorphic neurons, with or without balloon cells (subtypes IIb and IIa, respectively), commonly found in frontal lobe. Type III FCDs represent architectural distortions of cortical layer adjacent to hippocampal atrophy (subtype IIIa), glial or glioneuronal tumor (subtype IIIb), vascular malformation (subtype IIIc) and other lesions acquired in early childhood (subtype IIId).
On MRI, the FCD lesions may be characterized with [Kabat and Król, 2012]

1. increased thickness of the cortical gray matter
2. blurring of the gray-white matter junction
3. increased T2 and fluid attenuated inversion recovery (FLAIR) signal intensity in the subcortical white and gray matter
4. abnormal sulcal or gyral pattern

However, these findings may be very subtle and not easy to detect when visually inspecting MR images which eventually results in a high rate of normal MRI screenings among epilepsy patients [Lerner et al., 2009]. The three FCD types have different extent of visibility on MRI scans. Type II FCDs are significantly more visible on MR imaging than type I FCDs [Lerner et al., 2009]. Moreover, FCDs IIb are detected visually more frequently than FCD IIa [Colombo et al., 2012]. Fig. 3.1 gives an example of a subtle FCD lesion.

Heterotopia is another category of MCDs characterized by cortical cells (grey matter) encountered in inappropriate locations in the brain, as a result of interruption in their migration to the correct location in the cerebral cortex. Grey matter heterotopia may be unilateral or bilateral. Its most common form is bilateral periventricular nodular heterotopia (grey matter heterotopia lining the lateral ventricles). It can also occur in subcortical white matter (subcortical nodular heterotopia).

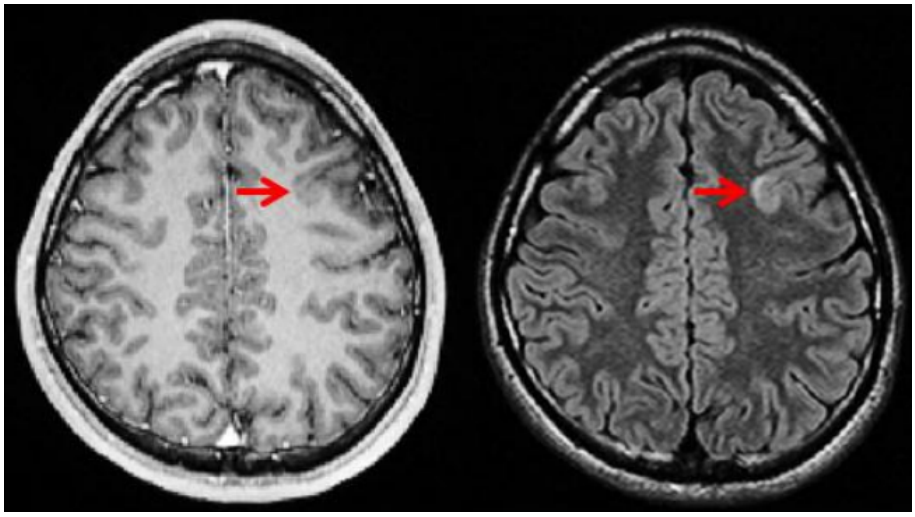


Figure 3.1: T1-weighted and T2-weighted axial MR images of an epilepsy patient. The focal cortical dysplasia (red arrows) present as loss of gray-white contrast on T1-weighted imaging and a hyperintensity on T2-weighted imaging. *Illustration from [Kini et al., 2016].*

3.2 Pre-surgical evaluation of intractable epilepsy

For patients diagnosed with medically refractory epilepsy the surgical removal of the lesions may offer a seizure-free life. The success rate of surgery, however, is only around 70% [Wiebe et al., 2001, Keller et al., 2007, Bien et al., 2012]. Moreover, the success rate of surgery in patients with normal MRI have been shown to be significantly lower than for MRI positive patients [Alarcon et al., 2006, Bell et al., 2009, Bien et al., 2009]. The resective epilepsy surgery consists of a complete disconnection of the epileptogenic zone while preserving the 'eloquent' cortex. Such a surgical intervention, therefore, depends heavily on the localization of the epileptogenic zone.

3.2.1 Clinical protocol for epileptogenic zone localization

The epileptogenic zone is the area of cortex indispensable for the generation of seizures. In practice, various techniques and tools are used to define the location and the extent of the epileptogenic zone, each of them having its proper definition and approximation of the zone in question. During the first stage of epilepsy diagnosis, neuroimaging techniques are used to infer the location of an *epilepsy lesion*, the cause of the seizures seen from the radiographic perspective. Since not all lesions found on neuroimaging data are actually responsible for the generation of seizures, electroencephalography (EEG) and video EEG are performed to target the most relevant. When a lesion found on neuroimaging data is coherent with the results of scalp video EEG telemetry, the patient can be recommended for surgery. Consequent techniques are considered to optimize the surgical procedures as well as to assess and minimize the related risks. When the neuroimaging data carries no visible relevant abnormality detected, more steps are required before a patient would be

referred to a possible surgery. In particular, the *irritative zone* (the region of cortex generating interictal electrographic spikes) and the *seizure-onset zone* (the area of the cortex where clinical seizures are *actually* generated) are inferred through EEG (scalp or invasive intracranial), magnetoencephalography (MEG) or functional MRI (fMRI) triggered by interictal spikes for the former and EEG (scalp or invasive intracranial) or ictal single photon emission computed tomography (SPECT) for the latter [Koepp and Woermann, 2005, Duncan et al., 2016]. These additional techniques provide necessary information on whether or not the surgery should be performed and if so, which zone should be targeted. The epileptogenic and seizure-onset zones may not coincide; one of them being larger than the other may lead to a surgical success or failure depending on if the sufficient part of the actual cause of seizures has been removed or not [Rosenow and Lüders, 2001].
This protocol of epileptogenic zone localization, however, is not carried out routinely at large scale. The reasons vary; some analysis techniques are not always available in medical centers (such as MEG or fMRI), others are skipped due to lack of skills to interpret them (e.g. PET) and in many cases, eventually, invasive exams (e.g. intracranial EEG) are performed without consulting less troublesome methods. Analyzing the information present on neuroimaging, therefore, may offer a chance to infer epileptogenic lesion localization with less discomfort for the patient or, when inconclusive, may guide the electrode placement depth in invasive exams, when they are an absolute necessity.

3.2.2 MRI and PET imaging in the lesion localization protocol

Neuroimaging has gradually become the technique of choice to gain a perspective on the structure and the functionality of the brain. High quality neuroimaging data has been exploited in the non-invasive diagnosis and therapeutic follow-up of various neuropathologies. Neuroimaging techniques, especially those based on multiparametric magnetic resonance imaging (MRI) and positron emission tomography (PET), have been exploited to detect epileptogenic lesions in a non-invasive manner.
The International League Against Epilepsy (ILAE) suggests an optimal protocol including T1-weighted, T2-weighted and FLAIR MRI sequences. Certain features, observed in different types of intractable epilepsy, can emerge in those MRI sequences. As such, volumetry, T2 relaxometry and FLAIR hyperintense signal are used to asses mesial temporal lobe epilepsies [Huppertz et al., 2011]. FCDs may appear on T1-w images as cortical thickening (50–90% of cases), abnormally deep sulci and blurring of the GM/WM interface (60–80% of cases), and may be associated with abnormalities of gyration [Barkovich and Kuzniecky, 1996, Besson et al., 2008]. FLAIR hypersignal is also often present (71-100% of cases) [Bernasconi and Bernasconi, 2015].
Diffusion tensor imaging (DTI) is another MRI sequence applied for intractable epilepsy detection [Lee et al., 2004, Thivard et al., 2006, Chen et al., 2008, Fonseca et al., 2012]. The sequence measures the diffusion of water molecules to create an anisotropy map. Using the

index of fractional anisotropy allows finding the orientation of the white matter tracts. In TLE, fractional anisotropy is consistently decreased and for FCD lesions, abnormalities in diffusion indices are present in the subcortical white matter, adjacent to the lesion [Fonseca et al., 2012, Bernasconi and Bernasconi, 2015]. Evidence also suggests that the appearance of DTI tracts can predict the surgical outcome, with displaced tracts recovering more favourably than those infiltrated by the target lesion [Bagadia et al., 2011].
The detection of small lesions, however, remains challenging. Subtle lesions are easily missed during standard visual inspections of the images. Recent retrospective studies involving surgical epilepsy patients indicate that up to 33% with typical FCD type II lesions and 87% with FCD type I (i.e. intracortical) 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 for the surgical resection when the lesion is visually detected (*MRI positive*) is 2-3 times higher than for the visually undetected lesions [Téllez-Zenteno et al., 2010]. Patients with lesions undetected during visual examination are referred to as *MRI negative* or *cryptogenic epilepsy* patients [Bernasconi et al., 2011]. Developing techniques capable of identifying subtle epileptogenic lesions on MRI data is therefore of a great importance for a possible surgery.

PET imaging, a nuclear medicine technique used to observe physiological processes in the body such as metabolism, has been receiving an increasing attention in the scope epilepsy lesion localization. Acquiring a PET scan involves injecting a tracer, labelled with a positron-emitting radionuclide, into a patient. For drug resistant epilepsy, the most widely available and clinically used PET tracer is ^{18}F fluoro-deoxyglucose (^{18}F-FDG) [Hammers, 2015]. This tracer allows assessing regional glucose metabolism. Areas of focal glucose hypometabolism are often larger than a lesion or the epileptogenic zone, but are generally correlated with seizure onset zones and/or areas of seizure spread [Juhász et al., 2000, Rathore et al., 2014].
The positive contribution of PET imaging was emphasized in a number of studies [Kim et al., 2011, Lamusuo et al., 2001, Rathore et al., 2014]. In [Kim et al., 2011], the diagnostic sensitivity reaches 83% while MRI results in 62%. Similar findings were reported in [Lamusuo et al., 2001], showing the significant hypometabolism in TLE patients with hippocampal damage, and in [Carne et al., 2004] where correct lateralization, coherent with the ictal EEG findings, was achieved for 26 out of 30 patients whose hippocampal sclerosis went undetected on MRI. The authors in [Salamon et al., 2008] went further by proposing to include PET images *co-registered* with MR scans in the presurgical evaluation of epilepsy patients which resulted in detecting one third of the lesions, not detected on MRI. Fig. 3.2 shows two examples of patients with normal MRIs while the corresponding PET images show significant hypometabolism in the concerned area. A meta-analysis of the studies evaluating the impact of PET imaging was carried out in [Willmann et al.,

2007] which reports the hypometabolism shown on PET imaging to have a predictive value of 80% in patients with normal MRI.
These works motivate an interest in PET imaging as a source of important complementary information in pre-surgical evaluation of epilepsy patients.

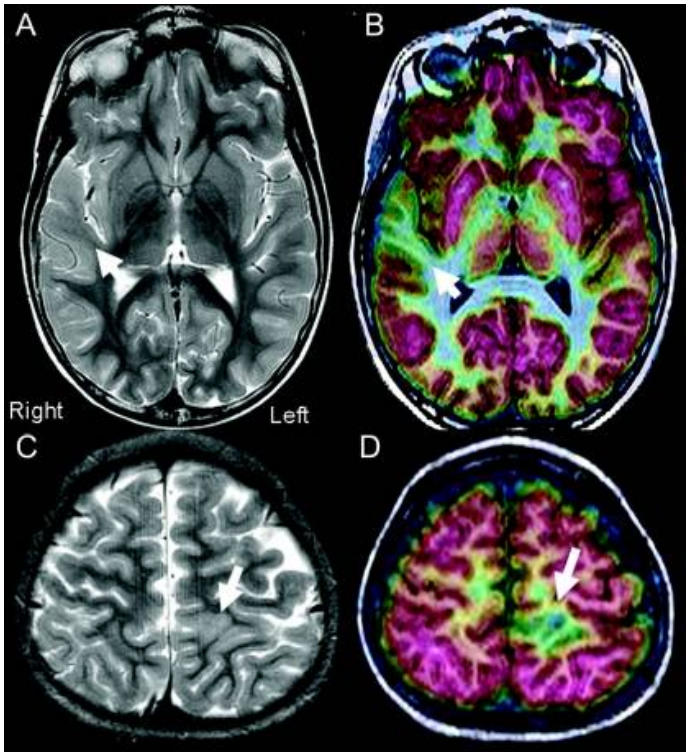


Figure 3.2: Examples of two patients with FCD type I. Both patients (shown row-wise) had normal MRI reading (left column). Right column corresponds to the MRI transverse slices overlaid with coregistered PET slices. For the first patient, FDG-PET/MRI coregistration indicates hypometabolism in the right superior temporal gyrus (B, arrow). For the second patient, FDG-PET/MRI coregistration indicated a focal area of hypometabolism in the left superior parietal region just behind the sensory cortex (D, arrow). *Illustration from [Salamon et al., 2008]*

3.3 State-of-the-art CAD systems for epilepsy

Over the recent years automated epilepsy detection has become the focus of many computer aided diagnosis systems. Some of them are based on EEG monitoring data. Those methods seek to identify epilepsy seizures on the EEG signals and more commonly perform a classification of normal versus abnormal signals, linked to epilepsy. This constitutes to patient-level systems as described in section 1.1.1. So, in [Subasi et al., 2005] the authors proposed to extract characterization of EEG signal with fast Fourier transform or autoregressive models and feed it into a neural network classifying a patient as epileptic or normal. A classification rate of 92.3% was achieved with this approach. Neural networks i.e. radial basis network and recurrent Elman Network, coupled with such features as spectral entropy, sample entropy and wavelet entropy, were used in [Kumar et al., 2010] to

classify healthy and unhealthy signals. Wavelet-based methods have been used in several studies including [Huafu and Hai, 2004, Zandi et al., 2010]. A comprehensive overview of the methods developed for the detection of epilepsy in EEG data is performed in [Saini and Dutta, 2017] and [Orosco et al., 2013].

The majority of works, however, develop CAD systems based on neuroimaging data. Among those, several contexts can be pinpointed. Some of them proposed subject-level CAD systems to discriminate healthy controls versus epilepsy patients. More commonly, the proposed CADs perform voxel-level analysis with explicit localization of the found lesion. Another common approach is based on surface-based morphometry [Dale et al., 1999], a method of constructing the surfaces of structural boundaries in the brain. Those boundaries (e.g. between white matter and gray matter) are established through a brain segmentation stage, and later the surface is reconstructed with a meshing algorithm. Within this category of approaches, the decisions are usually made at vertex-level where a vertex is where the corners of the triangles, constituting the mesh, meet.

3.3.1 Ground truth

Ground truth annotations provide the true labels for the observations in the data set. Depending on the granularity of a system, the nature of the ground truth annotations may vary. For subject-level CAD systems, subject-level labels are required i.e. each example is labeled as healthy control or epilepsy patient. For CADs aiming at localizing the lesion, as opposed to simply discriminating epilepsy patients versus healthy controls, more high-level labels are necessary. As such, the ground truth may be given by roughly outlined regions where the true zone of interest is located, or by its meticulous delineation at voxel-level. Naturally, the last case is the most time and resource consuming while image-level labels are relatively easy to obtain. Many epilepsy patients, as stated in section 3.2, have normal MRIs (MRI-negative). While it is possible to obtain voxel-level annotations for MRI-positive cases over a visual analysis, outlining a lesion in an MRI-negative patient requires external justification including post-surgical examination of the resected zone, histopathology or intracranial EEG analysis. For those patients, the most common reference is provided by a rough delineation of the supposed epilepsy lesion.
Evaluation of a CAD system, with respect to the available ground truth, may take place in several manners. For CADx systems discriminating healthy controls versus epilepsy patients the evaluation is rather straightforward; any metric described in 1.2.2 could be used to quantify the number of correct/incorrect decisions by the CAD. For CADe systems, the typical approach is to treat the detections overlapping with the true lesion delineation as true positives and those outside of it, as false positives. Naturally, when precise lesion contours are available, true positives/false positives may be defined as the detected voxels located inside/outside the ground truth.

3.3.2 Features in CAD systems for epilepsy detection

The current CAD systems for medically refractory epilepsy have considered almost all characteristics discovered by radiologists in search for markers of epilepsy on neuroimaging. Substantial efforts have then been invested in translating those clinical findings into automatically computable features. Commonly, the features are computed at voxel-level or at the level of vertices of the triangle mesh modeling the cortical surface (Surface Based Morphometry). The most frequent of those features associated with the appearance of FCDs on MRI data are:

- The distance between the gray/white matter boundary and the outermost surface of the gray matter quantifying the *cortical thickness*. Increased cortical thickness has been associated with FCDs [Blümcke and Spreafico, 2011].
- *Gray-white matter junction* computed by a convolution of the binarized image, quantifies the gray-white matter blurring. *Gray-white matter extension*, computed by applying a Gaussian smoothing filter over the segmented gray matter image, is a measure quantifying the extension of the gray matter into the white matter. Both characteristics are frequently used to identify FCDs [Huppertz et al., 2005].
- *Sulcal depth* is estimated by calculating the dot product of the movement vectors with the surface normal.
- *Curvature* is quantified as the inverse of the radius of an inscribed circle and mean curvature represents the average of two principal curvatures.

Table 3.1 lists the features (mostly relevant for FCDs) used frequently in the development of epilepsy CAD systems. Combinations of those features have been used in various works [Huppertz et al., 2005, Ahmed et al., 2016, Thesen et al., 2011, Hong et al., 2014].

3.3.3 Methods in CAD systems for epilepsy detection

As explained in the previous chapters, obtaining ground-truth annotations for epilepsy patients is a difficult and time consuming task for a number of reasons, more so for the *MRI-negative* patients. A significant number of the state-of-the-art epilepsy detection methods, however, perform supervised classification, mostly on data sets containing MRI-positive epilepsy patients. A few studies acknowledged the difficulty of ground truth annotations of epilepsy lesions and proposed semi-supervised or unsupervised approaches. It is important to mention that most of the works, exploiting the methods below for epilepsy CADs, targeted FCD lesions solely, using the features corresponding to the clinical knowledge on the appearance of FCD lesions on neuroimaging data.

Feature	**Computed with**
Image intensity	Voxel-based morphometry ([Ashburner and Friston, 2000]), difference maps, Laplacian intensity gradient, statistical measures (mean, median, variance, skewness, kurtosis, energy, entropy)
Cortical thickness	Diffeomorphic registration based cortical thickness, distance between gray/white and pial isocontour surfaces
Gray-white junction blurring	Gradient map using Gaussian smoothing, identify areas with highest cortical thickness, MAP, iterated local searches on neighborhood
Sulcal reconstruction	Graph matching, gyrification index, spherical wavelets
Lobar or volume atrophy/enlargement	Deformation based morphometry, Jacobian of heat equation vector field applied to spherical harmonics with a point distribution model
Curvature	Gaussian intrinsic curvature, integral measures of curvature, orientation fields from gradient structure tensors, area-minimizing flows to spherical registration
Asymmetry analysis	Asymmetry index, asymmetry analysis on cortical folding
Other cortical measures	Fractal analysis of the cortex, metric distortions on spherical registration
Texture Analysis	
3D Texture analysis	Drectional Riesz wavelets
Gray-level co-occurrence (contrast, homogeneity, inverse difference, energy, entropy)	Haralick et al algorithm [Haralick et al., 1973]
Gray-level run-length (short/long run emphasis, gray level distribution, run-length distribution)	Haralick et al algorithm [Haralick et al., 1973]

Table 3.1: Overview of features commonly used in CADs for epilepsy detection. Different combinations of these features were used to isolate and identify lesions (usually focal cortical dysplasias). *Table from [Kini et al., 2016].*

GLM-based statistical analysis is a common approach in neuroimaging based on a mass univariate analysis that fits a General(ised) Linear Model [McCullagh and Nelder, 1989] for each voxel. General Linear Models include a number of methods such as linear regression. Once the model for each voxel is fitted and its parameters are estimated, the latter are used to produce a statistic (e.g. t-statistic, F-statistic) allowing to accept or reject a corresponding hypothesis for each voxel. This statistical analysis, also known as voxel-based morphometry (VBM) in the neuroimaging community [Ashburner and Friston, 2000], has been used frequently in epilepsy studies as a tool to compare a patient with a cohort of healthy controls. The approach allows to assess the hypothesis of significant differences between a subject and a cohort of healthy controls and eventually yields a map of clusters where each cluster corresponds to a set of voxels where the differences are remarkable. Those clusters are reported as the localizations of the predicted epilepsy lesions. GLM-based analysis with various choices of statistics and GLM has been used in [Riney et al., 2012, Chassoux et al., 2010, Chen et al., 2008, Focke et al., 2008, Bruggemann et al., 2007, Thivard et al., 2006, Srivastava et al., 2005]. The features considered in those works include gray/white matter probability maps, cortical thickness, FLAIR intensities etc. A comprehensive GLM-based method is implemented in the Statistical Parametric Mapping (SPM) software, a common choice in many epilepsy studies.

Supervised learning methods have also been explored in CAD systems for epilepsy detection. Since it is extremely difficult to obtain accurate lesion delineations in MRI-negative patients, most supervised approaches consider mainly MRI-positive cases.
Voxel-based detection has been performed in several studies. So, [Antel et al., 2003] first trained a Bayesian classifier on different intensity features, including cortical thickness, relative intensity and intensity gradient magnitude, to classify voxels as lesional or normal. As a next step, Fisher's discriminant ratio using textural features, derived from gray-level co-occurrence matrices, was used to reclassify voxels classified as lesional in the first step. The method achieved 83% sensitivity for 100% specificity on a data set containing 7 MRI-negative cases. [Yang et al., 2011] used a Naive Bayes classifier to discriminate healthy versus lesional voxel cubes using statistical measures on cortical thickness and gradient vectors. Vertex-based approaches have been explored as well. [Besson et al., 2008] proposed a two-step system consisting of a neural network classifying vertices as lesional/normal based on several surface-based features (cortical thickness, curvature, sulcal depth, etc) and a false positive reduction step discriminating the previously found true detections and false positive clusters with fuzzy k-Nearest Neighbour classifier. The first step allowed to detect 18/19 lesions while the second step, after reducing significantly the number of false positives, reached 13/19 detection rate. More recent approaches such as [Hong et al., 2014] and [Ahmed et al., 2015] applied Linear Discriminant Analysis and Stratified Logistic regression on the vertices of the cortical surface and evaluated the methods on data sets,

containing among others, 19 and 24 MRI-negative patients, respectively. 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%. In [Hong et al., 2014], however, the lesions of the patients initially considered MRI-negative were later visually detected on MRI. [Adler et al., 2017] applied a simple neural network to classify healthy versus pathological vertices on 28 cortical features and each of them individually, achieving up to AUC = 0.87, depending on the feature. A similar approach was proposed in [Jin et al., 2018]. In [Gill et al., 2017], the authors first proposed to use vertex-level and cluster-level RUSBoost on 30 cortical features, achieving 83% (4 ± 5 FPs) sensitivity and 92% (0.08 ± 0.27 FPs) specificity. To the best of our knowledge, the only work on automated epilepsy lesion detection on data-driven features was proposed in [Gill et al., 2018], where 2 convolutional neural networks were trained to classify raw image voxels, reaching 91% (3 ± 2 FPs) sensitivity and 92% (1 ± 0 FPs) specificity.

Unsupervised or semi-supervised methods have also been explored in the development of CADs for epilepsy detection. So, a simple approach based on univariate z-score thresholding was proposed by [Thesen et al., 2011]. The considered features were cortical thickness, gray-white matter contrast, curvature, sulcal depth and Jacobian-distortion and the thresholding was done for each feature individually. Eventually, the method achieved at best 100% specificity for 84% sensitivity on thickness and 84% specificity for 61% sensitivity using gray-white matter contrast. Another simple approach was proposed in [Strumia et al., 2012] where a number of intensity, texture and form based features were used to model the normality of each voxel with a Gaussian distribution. Later, the voxels with low probability with respect to the estimated Gaussians of all features were identified as epileptogenic voxels. [Ahmed et al., 2014] formulated a semi-supervised extension of hierarchical conditional random fields (HCRF) and used it on cortical thickness to classify vertices as abnormal and eventually discriminate epileptogenic vertices. The supervision is added to compute node (patches at different scales) potentials in HCRF, with labels computed to represent if the node is different from the corresponding nodes of the healthy control population. In [Ahmed et al., 2016] the authors extend the previous study by adding more relevant features for the detection of FCD such as gray/white-matter contrast, sulcal depth and curvature. Moreover, the abnormality of each vertex was measured by the probabilistic output of the LoOP outlier detection method [Kriegel et al., 2009]. An important contribution of the latter study was that the evaluation was performed on a group of *MRI-negative* patients and achieved at best a detection rate of 70% for 9 false positive detections per patient.

Another approach that has initiated the scope of this work was proposed by a former PhD

student in [El Azami et al., 2016]. It consists in a CAD system using an entirely unsupervised method. For each voxel in the brain, a oc-SVM model was trained on gray-white matter junction and extension values and possible epilepsy lesions were predicted as the clusters of voxels that were classified as outliers by the corresponding oc-SVM models. The main advantage of this method is that no annotated training data is required. The CAD system achieved a detection rate of 77% for a false positive rate of 3.2 per patient on a data set composed of 13 epilepsy patients with 10 MRI-negatives.

3.3.4 CAD systems for TLE and FCD

The current CAD systems for TLE are summarized in table 3.2. The main objective of these works is the TLE diagnosis, therefore precise detection of lesions is not sought. Obtaining labeled data set being easier in this context, supervised methods have been applied frequently. Moreover, two main directions can be pointed out.

1. *Patient-level discrimination* of epilepsy patients consists in discriminating TLE patients from healthy controls. Several studies have focused on this task for TLE such as [Focke et al., 2012, Cantor-Rivera et al., 2015]. Hippocampal sclerosis (HS) seems to play an important role in the discrimination of TLE patients from healthy controls. So, when HS is present, the classification accuracy reaches 89-96%, as opposed to 86% when it is not.

2. *Lateralization of the epileptogenic lesions* is the problem of discriminating the side of the brain (left versus right, typically) where the epileptogenic zone is located. Several studies have tackled this problem for TLE. Among those [Duchesne et al., 2006a, Keihaninejad et al., 2012] developed a system on MRI T1 data while [Focke et al., 2012, Pustina et al., 2015] considered multimodality data such as MRI, DTI and PET. Similarly to the previous case, the presence of HS is important. [Duchesne et al., 2006a, Keihaninejad et al., 2012] both achieved 100% accuracy in patients with HS.

Tables 3.3-3.5 summarize the current methods for FCD detection. Unlike for TLE, these systems aim to actually detect the epileptogenic lesions in patients, which is a more challenging problem. As it can be seen from the tables, most such studies consider only T1w MRI, although DTI and FLAIR were included in some of them [Thivard et al., 2006, Focke et al., 2008, Chen et al., 2008]. Many of these methods are based on GLM-analysis, comparing potential patients to a cohort of healthy controls [Srivastava et al., 2005, Bruggemann et al., 2007, Focke et al., 2008, Riney et al., 2012]. A significant number of approaches is developed in supervised contexts [Antel et al., 2003, Besson et al., 2008, Yang et al., 2011, Hong et al., 2014, Ahmed et al., 2015, Adler et al., 2017, Gill et al., 2017]. The used features are all hand-crafted and coincide with those presented in table 3.1. The lack of studies using data-driven features becomes immediately obvious. Indeed, only in

[Gill et al., 2018] the relevant features are learnt with neural networks. Another important consideration is the different metrics used for evaluation, including detection rate, AUC, accuracy, precision, recall, etc. Those metrics may be calculated differently depending on the convention adopted by the authors. So, in some studies the sensitivity may be computed at voxel level while calculated at subject level in others. Moreover, there is no benchmark data set to evaluate the methods which makes the comparison very difficult. Even more so when considering that some approaches were evaluated on MRI-negative patients as in [Ahmed et al., 2016] while others focused on 'easier' cases. For MRI-negative patients, the achieved sensitivity varies between 52 and 70% [Ahmed et al., 2015, El Azami et al., 2016, Ahmed et al., 2016]. Depending on the adopted method, some studies report specificity as to quantify the false detections in healthy controls while others skip this evaluation and report false positive detections per patient.
PET imaging is rarely exploited in the existing CAD systems. Among the few mentioned studies, [Chassoux et al., 2010] leveraged the PET imaging in an automated lesion detection system, based on GLM analysis on PET image intensity values. The statistical analysis performed in this study, however, was less efficient than a simple visual analysis of patients' PET images which greatly improved the epileptogenic lesion localization in patients whose T1-w MR images were considered normal in 13 out of 23 cases. A recent method proposed in [Tan et al., 2018] considers handcrafted features extracted from MRI and PET imaging and combines them in a 2-step approach based on SVM in order to identify FCD lesions. This is the only study leveraging MRI and PET data in a single system, using, however, handcrafted features. The results report an increase in maximum sensitivity from 82% to 93% and in the sensitivity, corresponding to the maximum specificity, from 61% to 64%, when PET imaging is considered alongside MRI data. The accompanying FP rate in FCD patients, however, increases as well.

We have presented a detailed description on epilepsy and the current features and approaches for automated epilepsy detection on neuroimaging. The main drawbacks of the current systems for epilepsy detection can be summed up in the following aspects. First, the proposed systems chiefly target a particular cause of epilepsy (such as FCD) and do not generalize for other epilepsy categories. Therefore, the *handcrafted* features considered in the current systems are relevant for the particular pathology. A wider range of features, however, may be more beneficial. The second aspect is the fact that the combination of several imaging modalities (such as T1w and FLAIR MRI and PET imaging) is rarely explored. As mentioned in a survey on computational analysis in epilepsy by [Kini et al., 2016], considering multimodal data may provide important complementary information, otherwise ignored in monomodal settings. The third aspect is that comparing different methods is not trivial in the absence of a unified protocol for performance evaluation.

Eventually, most studies are evaluated on *MRI-positive* cases where the lesions are visible (though subtle) on MR scans. The real challenge, however, are the *MRI-negative* patients.[1]

[1]Abbreviations in the tables below. MTL: medial temporal lobe; L: left; R: right; NC: normal control; HS: hippocampal sclerosis; nHS: without hippocampal sclerosis; CV: cross-validation; LOO: leave-one-out; LOPO: leave-one-patient-out.

Study	Data	Imaging	Ground Truth	Object definition	Features	Classifier	Evaluation	Results
[Duchesne et al., 2006b]	152 NC, 80 TLE HS, 47 TLE nHS	1.5T MRI: T1	video-EEG, clinical findings	predefined ROI centred on L and R MTL	T1 intensity, volume change measure	PCA and LDA	LOO CV	TLE HS vs nHS: Acc=100%, TLE HS L vs R: Acc=100%, TLE nHS L vs R: Acc=100%, TLE L vs R: Acc=96%
[Thivard et al., 2011]	40 NC, 13 non-lesional TLE	1.5T MRI: T1, DTI, PET	SEEG and sublobar region co-localization	voxels	T1 GM volume, DTI MD, PET normalized intensity	3 GLMs (ANCOVA) 1 per feature	single patient against controls	Sen: 4/13 GLM PET, 2/13 GLM DTI and 3/13 GLM GM
[Concha et al., 2012]	21 NC, 30 TLE	MRI DTI	video-EEG and neu-roimaging	clusters of group-wise difference	FA, MD, parallel and perpendicular diffusivity	LDA	LOO CV	TLE HS L vs R: Acc=91%, TLE nHS L vs R: Acc=71%
[Focke et al., 2012]	22 NC, 38 TLE HS	MRI: T1, T2, DTI	video-EEG and post-operative outcome	patient	T1 GM and WM segmentation, T2 relaxation, FA and MD	binary (NC vs R, NC vs L) and multi-class (one vs one) SVM	LOO CV	Acc=90-100% binary SVM, Acc=88-93% multi-class SVM
[Keihaninejad et al., 2012]	28 NC, 60 TLE HS, 20 TLE nHS	3T MRI: T1	consensus diagnosis by 2 experts	83 anatomical structures	structural volume, spectral features (volumetric difference)	RBF SVM, linear SVM	10 fold CV	TLE HS vs NC: Acc=96% TLE nHS vs NC: Acc 86% RBF, 91% linR SVM, TLE HS L vs R: Acc=100% TLS nHS L vs R: Acc=86% RBF SVM and 94% linR SVM
[Cantor-Rivera et al., 2015]	19 NC, 17 TLE	3T MRI: T1, T2, DTI	EEG and post-surgical pathology (8/17)	156 ROI subject specific atlas	mean and asymmetry values of T1, T2, FA, MD	PCA and/or ANOVA followed by SVM	LOO CV	TLE vs NC: ANOCA-PCA-SVM all features: Acc= 88.9%, T1: Acc=81%, MD: Acc=75, T2: Acc=74% and FA: Acc=67%
[Pustina et al., 2015]	3T PET, MRI, DTI	28 LTLE, 30 RTLE	EEG, video recording, intracranial EEG	patient	assymteris from PET (glucose metabolism), MRI (cortical thickness), DTI (white matter anisotropy)	stepwise logistic regression	bootstrapped split-sample validation	Detection rate : 80% - 100%

Table 3.2: State-of-the-art methods for TLE detection.

Study	Data	Imaging	Ground Truth	Object definition	Features	Classifier	Evaluation	Results
[Antel et al., 2003]	14 NC, 18 FCD (11 MRI+)	1.5T MRI: T1	manual lesion delineation using EEG and resection area	voxels (3000 sampled per subject)	cortical thickness, GM/WM contrast, relative intensity and 9 texture features derived from grey-level co-occurrence matrices	2 Bayes classifier (first stage computational model, second stage texture based model)	LOPO CV	Patient-level: Sen= 15/18, Spe= 100% no detection in controls, cluster-level: Sen=17/20, Spe=5/18
[Srivastava et al., 2005]	64 NC, 17 FCD (11 MRI+)	1.5T MRI: T1	manual delineation on T1	voxels	cortical thickness	GLM	single patient against controls	Sen= 9/17, Spe= FP detections in less than 4.5% of controls
[Colliot et al., 2006]	39 NC, 27 FCD	1.5T MRI: T1	manual segmentation of the lesion	voxels	GM parametric map (z-score map)	threshold of the parametric map	single subject against controls	Sen= 21/27, Spe= 1-4 FPs per subject
[Thivard et al., 2006]	40 NC, 16 (TLE, FCD 2 MRI+)	1.5T MRI: DTI	SEEG	voxels	FA, MD	GLM	single patient against controls	Co-localization in 7/16 cases *e. g.* with the irritative zone
[Bruggemann et al., 2007]	24 NC, 16 FCD children	1.5T MRI: T1	16 manual ROIs, TP if overlap >5%	voxels	GM, WM	GLM WM, GLM GM, GLM conjunction	single subject against controls	Sen: 10/16 GM-only, 14/16 (WM GM), 11/16 GM or WM, 3/16 GM and WM.
[Focke et al., 2008]	25 NC, 25 FCD	3T MRI: FLAIR	2 experts consensus and histology	voxels	FLAIR intensity	GLM	single patient against controls and LOPO for controls	Sen= 22/25, Spe= 1 FP in one control
[Besson et al., 2008]	41FCD, 48 +11 NC	1.5T MRI T1	manual segmentation of the lesion	cortical surface vertices	cortical thickness, WM/GM blurring, hyperintensity, depth, curvature	neural network + fuzzy k-NN	LOPO CV	Vertex-wise classification: 18/19, 23.1/7.1 FP in patients/controls, Cluster-wise classification: 13/19, 2.8/1.4 FP in patients/controls
[Chen et al., 2008]	40 NC, 15 FCD (MRI-)	3T MRI: DTI	EEG findings	voxels	FA and MD	2 GLM	single patient against controls	Sen: 7/15 with GLM MD, 2/15 with GLM FA

Table 3.3: State-of-the-art methods for FCD detection – part I

Study	Data	Imaging	Ground Truth	Object definition	Features	Classifier	Evaluation	Results
[Chassoux et al., 2010]	30 NC, 18 FCD-II	PET	histology	voxels	intensity	GLM	single patient against controls	GLM PET: 16/18 concordance with visual analysis in 13/18 cases
[Thesen et al., 2011]	48 NC, 11 FCD	3T MRI: T1	manual lesion segmentation on T1 and FLAIR if available	cortical surface vertices	thickness, GM/WM contrast, local gyrification, sulcal depth, curvature and Jacobian	Threshold of the parametric maps	single patient versus controls; LOPO CV on controls	Best operating point (Spe, Sen): (100%,84%) using thickness, (84%,61%) GM/WM contrast
[Yang et al., 2011]	21FCD	MRI T1	manual lesion segmentation	8x8x8 voxel cubes	statistics on cortical thickness, absolute gradient, gradient vectors	Naive Bayes	LOPO CV	62.49% detection rate and 19.31% false positive rate
[Riney et al., 2012]	29 NC, 8 FCD, 14 cryptogenic (children)	1.5T MRI: T1, FLAIR	experts' clinical findings	voxels	intensity scaled FLAIR, GM T1	GLM-FLAIR, GLM-T1	single patient against controls (pcorr<0.05)	FCD: GLM-T1: 3/8, GLM-FLAIR: 7/8, Cryptogenic: GLM-T1: 2/14, GLM-FLAIR: 4/14
[Strumia et al., 2012]	20 NC, 11 FCD Ia, Ib, IIa, IIb	3T MRI: T1, FLAIR	histopathology	voxels	intensity, texture, form-based features	fitting a Gaussian per voxel & Naive Bayes	train/test split	precision = 0.51 ± 0.04, re = 0.15 ± 0.09, Dice = 0.13 ± 0.04
[Hong et al., 2014]	24 NC, 19 FCD II-	3T MRI: T1	clinical findings and histology	cortical surface vertices	sulcal depth, curvature, gradient and statistical descriptors of the associated z-score maps	2 step LDA (imaging features and then statistical features)	LOPO CV	Step 1: 18/19 Sen, and 32 FP per patient. Step 2: 14/19 Sen, 1-3 FP per patient
[Ahmed et al., 2015]	62 NC, 31 FCD (24 MRI-)	3T MRI: T1	manual segmentation for MRI+ and resection zone for MRI-	cortical surface vertices	thickness, GM/WM contrast, sulcal depth, mean curvature, Jacobian distortion	bagging and logistic regression (stratified classification)	LOPO CV; comparison against VBM-thickness	MRI+: 6/7 both logistic regression and VBM-thickness, MRI-: 14/24 logistic regression and 9/24 VBM-thickness

Table 3.4: State-of-the-art methods for FCD detection – part II

Study	Data	Imaging	Ground Truth	Object definition	Features	Classifier	Evaluation	Results
[El Azami et al., 2016]	37 + 40 NC, 13 FCD	1.5T MRI T1	manual segmentation, histology, SEEG	voxels	gray/white matter junction, extension	voxelwise oc-SVM	patient group validation	Sensitivity: 10/13, Avg. # of FPs: 3.2
[Ahmed et al., 2016]	115 NC, 20 FCD	3T T1	post-surgical MRI of resected zone	cortical surface vertices	cortical thickness, gray/white-matter contrast, sulcal depth, curvature	semi-supervised HCRF + LoOP	patient group validation	Detection rate: 52-70% (depending on the feature)
[Adler et al., 2017]	28 NC, 22 FCD	1.5T T1, FLAIR	manual segmentation	cortical surface vertices	cortical thickness, GM/WM contrast, sulcal depth, mean curvature, "doughnut" thickness etc	neural network	LOPO validation	AUC: 0.51 - 0.83 (depending on the feature)
[Gill et al., 2017]	41 FCD, 38 HC	3T T1, FLAIR	manual segmentation	cortical surface vertices	cortical thickness, sulcal depth, curvature, (intra/sub)cortical intensity maps, gradient maps	vertex-level and cluster-level AdaBoost / RUSBoost	5-fold CV	Sensitivity: 83% (4 ± 5 FPs), Specificity: 92% (0.08 ± 0.27 FPs)
[Jin et al., 2018]	120 NC, 61 FCD	3T T1	manual segmentation	cortical surface vertices	cortical thickness, gray-white matter contrast, curvature, sulcal depth, "doughnut" maps, local cortical deformation	neural network	5-fold CV	Sen: 73.7%, Spe: 90.0% at optimal threshold
[Gill et al., 2018]	40 FCD, 38 NC, 67 FCD, 63 TLE/HS	1.5T/3T T1, FLAIR	manual segmentation	voxels	raw images	2 CNNs	train/test split	Sensitivity: 91% (3 FPs ± 2), Specificity: 92% (1 FPs ± 0)
[Tan et al., 2018]	23 TLE, 28 FCD	3T MRI, PET	histopathology	cortical surface vertices	cortical thickness, blurring, sulcal depth, PET hypointensity, PET asymmetry, etc	2 SVMs	LOPO CV	Step 1: 28/28 max sen, 63.5 ± 26.5 FPs. Step 2: 26/28 max sen, 35.8 ± 12.2 FPs

Table 3.5: State-of-the-art methods for FCD detection – part III. HCRF: Hierarchical Conditional Random Fields.

Chapter 4

Problem formulation

In the previous chapters we covered various aspects of the modern CAD systems. We started by a general introduction on the main components of such systems, detailing the existing approaches. Chapter 2 delved into an overview of the most popular category of methods used in the state-of-the-art CAD systems for various medical applications, namely deep learning based methods. Eventually, we presented the current CAD systems designed for epilepsy lesion detection on neuroimaging data. We summarized the main methodological choices implemented in those systems and outlined their particularities. In this chapter we present our considerations for the problem at hand and state the choices we have made when proposing our solution.

4.1 Motivation and strategy

The objective of this work is to propose a conceptual framework aimed at subtle anomaly detection on brain imaging. The clinical application of such a CAD system consists in automated epilepsy lesion detection in MRI-negative patients. In chapter 3 we discussed the specific characteristics of epilepsy lesions and the challenges associated with their detection on neuroimaging. We further presented the state-of-the-art methods for epilepsy detection.

Several limitations can be observed in the existing CAD systems for epilepsy lesion detection. First, most of them are specifically designed for FCD detection and do not present a unified system for other epilepsy causes. The second aspect concerns the methods chosen as the core decision-making mechanisms. Supervised learning methods, despite having impressive results in other medical applications, are less adapted to this context. The main reason, as explained in the previous chapter, is the high heterogeneity of epilepsy lesions in terms of size, shape and localization, and the lack of labeled training data sets, adequately representing this variability. Moreover, while it is possible to gather a huge

data set of epilepsy patients with voxel-level lesion annotations when the latter are visible on MRI, it is unclear what protocol should be followed to obtain such meticulous annotations for MRI-negative patients, which are the main challenge of the CADs for epilepsy detection. Assuming that a careful analysis of post-surgical exams, histopathology and/or invasive EEG findings could provide some sort of lesion delineation, the annotated zone is very likely to contain healthy tissue which means introducing label noise to the chosen supervised learning algorithm. This case have been argued in [Ahmed et al., 2015] where replacing the initial 'generous' annotations (based on resection masks) for MRI-negative patients with tighter boundaries resulted in improving the detection rate from 12% to 58%. In our case, the considered data set contains too few patients to cover the complexity of epilepsy lesions. Moreover, the concerned patients are chiefly MRI-negative patients. Supervised learning in these circumstances does not seem realistic.

The next category of the existing methods is based on GLM-analysis. The main strategy of these methods is to perform mass statistical analysis so as to compare a given subject to a group of normal healthy controls. Originally, the method has been developed in the univariate setting. When extending it to the multivariate setting, so as to account for multiple features (effects), the performance obtained with individual features is not always improved or even preserved. [Bruggemann et al., 2007] performed a GLM-based analysis in 4 settings - 1. GM-only, 2. GM-WM aggregate, 3. GM *or* WM and 4. GM *and* WM (GM: gray matter, WM: white matter). The corresponding detection rates were 10/16, 14/16, 11/16 and 3/16, respectively. It is not clear what method should be adapted to account for multiple effects; the conjunction null hypothesis (both effects should be significant) may clearly underperform.

We therefore adapt an unsupervised strategy for the difficult task of epilepsy lesion detection. In such a setting, we seek to identify abnormalities (including epilepsy lesions) on brain imaging through learning the normality of the brain on healthy examples. Therefore, we employ the outlier detection approach. Identifying epilepsy patients by detecting outliers at image-level is not realistic; the lesions are too subtle and are not likely to discriminate themselves more than other healthy anatomical variations in the brain. We therefore adopted the approach proposed in [El Azami et al., 2016] which consists in learning the normality of each voxel and detecting abnormalities as local neighborhoods of voxels with high abnormality scores. The general structure of the CAD system is shown on fig. 4.1. This approach allows to bypass the need for a large voxel-level labeled training data set, representing the heterogeneity of epilepsy lesions.

4.2 Challenges and objectives

The system proposed in [El Azami et al., 2016] was trained on two handcrafted features relevant for FCD lesions - gray/white matter junction and extension, and evaluated on a

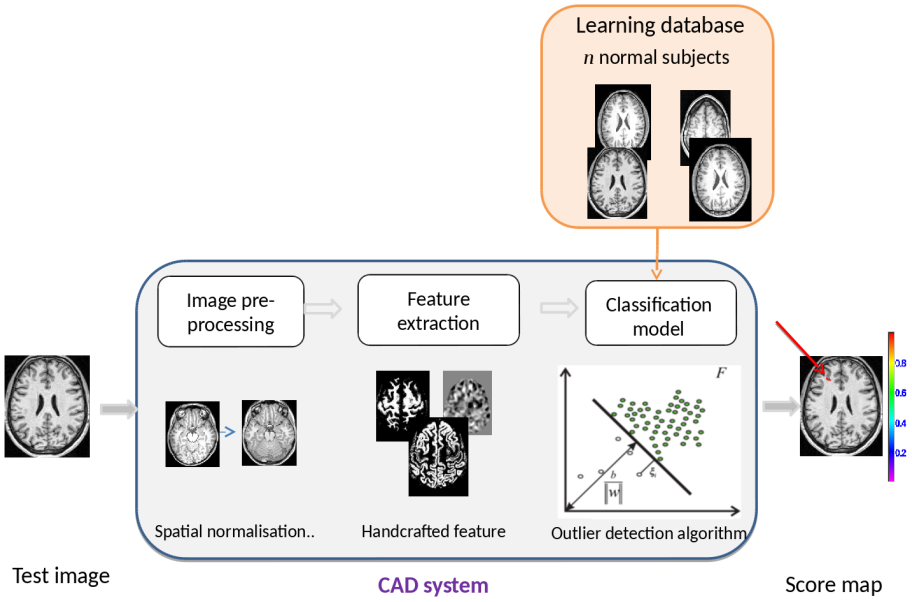


Figure 4.1: General representation of the CAD system by [El Azami et al., 2016]. The framework consists of three major steps - 1. Image normalization to a common template, 2. Extraction of handcrafted feature maps i.e. gray-white matter junction and extension maps and 3. oc-SVM model learning per voxel in the brain. For a new test image, each oc-SVM yields a score corresponding to the anomalousness of the voxel.

small set of 3 MRI positive and 10 MRI negative patients. These clinically guided FCD characteristics are limited to the current knowledge on the appearance of such lesions on MRI. However, they might not be the optimal ones to identify the lesions. While the reported performance is adequate (10/13 detection rate), the evaluation on a larger MRI negative patient set is yet to be conducted. Moreover, the CAD was designed on features computed on T1w MRI. A multimodal analysis was not conducted.

In this work we attempt to address both issues. We first hypothesize that representations learnt in a data-driven fashion may capture information left out when computing handcrafted features and, therefore, may lead to a better performance. As representation learning mechanisms, we consider deep learning architectures. Our objective, therefore, consists in proposing and developing deep learning architectures suitable for extracting representations to perform outlier detection per voxel. We thus define the objectives of this work as follows.

1. We will propose and develop unsupervised deep architectures to extract voxel-level representations to be used to train a oc-SVM model for each voxel in the brain.

2. We will evaluate the performance of the overall system on a group of 21 epilepsy patients with 18 *MRI negative/cryptogenic* cases and compare the results with those obtained with the handcrafted features in [El Azami et al., 2016].

3. We will next explore strategies to integrate multiple imaging modalities into a single framework, in order to leverage the complementary information present in different modalities of neuroimaging data.

Eventually, the strategies implemented in this work are summed up in the following contributions.

4.3 Contributions

Chapter II starts by presenting the general framework of the proposed CAD system and introducing the data set of healthy controls and epilepsy patients that will be explored throughout this work. We then move to the core of this study and introduce our first contribution which is to propose and exploit various unsupervised deep architectures as potential feature extraction mechanisms for outlier detection. We then identify the limitations of the existing architectures and propose a new configuration of siamese networks to better fit the context of subtle outlier detection on neuroimaging. We compare the learnt features with their handcrafted alternatives within the same framework and, further, with the currently popular SPM analysis. To our knowledge, data-driven representations have only been used in one recent study by [Gill et al., 2018] and, therefore, deserve to be looked at.

Chapter III addresses the issue of integration of multiple modalities for outlier detection. The choice of a strategy of integrating the information from different modalities is not a trivial one. A decision should be made on what is the optimal level of integration within the framework and what methods can be used to achieve it. As such, we propose and compare two strategies. The first strategy consists in training multichannel deep architectures, i.e. networks that combine the images of different modalities as input channels and, therefore, learn representations on their combination. The second approach represents an intermediate level fusion strategy. In this case the representations are learnt with deep networks for each modality individually and later combined through a multiple kernel learning paradigm. We compare the two approaches and eventually report the best performance achieved in multimodal outlier detection on T1-w and FLAIR MRI. Multimodal epilepsy lesion detection has been addressed only in a few works as shown in chapter 3.

The final chapters present our exploratory efforts to leverage the PET imaging as a potential source for epilepsy lesion detection. This imaging modality has not received the same attention as, for example, T1-w MRI, as can be seen from the summary on the state-of-the-art methods (tables 3.3 - 3.5). In our context, the available PET images are fewer in number than the corresponding T1-w and FLAIR MRIs. Hypothesizing that the insufficient number of training examples would not allow us to leverage the PET images at best, we explore strategies of their indirect integration. As such, we make an attempt of

PET image synthesis from MRI and evaluate the performance of the system with synthesized images substituting the missing ones. We show the improved sensitivity of the CAD system when both real and synthetic PET images are considered.
Eventually, the manuscript concludes with a general conclusion and our considerations for perspective work.

II Unsupervised representation learning for anomaly detection

Chapter 5

CAD pipeline and data description

In the previous chapters we described the principles of CAD systems for medical image analysis and further reviewed deep learning based methods for various medical applications, neuropathologies in particular. Further, we presented a detailed overview of the state of the art approaches for epilepsy lesion detection on neuroimaging. Eventually, the constraints and the specifics of the difficult task of epilepsy lesion detection were discussed in chapter 4 where we explained and formalized our approach to the problem at hand.
Our aim in the scope of this project is to adapt and develop unsupervised deep architectures as representation learning mechanisms, tailored to the task of per-voxel anomaly detection on brain MR images. Among others, such a system should capture such subtle abnormalities as epilepsy lesions. This chapter presents the general pipeline of the proposed approach and introduces the methods used in its implementation. The chapter concludes with a detailed description of the data set considered in this study, composed of a set of healthy controls and patients with confirmed epilepsy lesions.

5.1 General framework

The general framework of the CAD system is shown on fig. 5.1. The main components of the system represent the two main stages of the approach i.e.

1. representation learning

2. per voxel outlier detection model learning

Both components are trained on a data set composed of healthy examples only. First, a neural network is trained on patches of MR images extracted from healthy patients. Once the training is completed, a oc-SVM model is built for every voxel taking at input the representations corresponding to the patches of healthy images centered at the voxel.

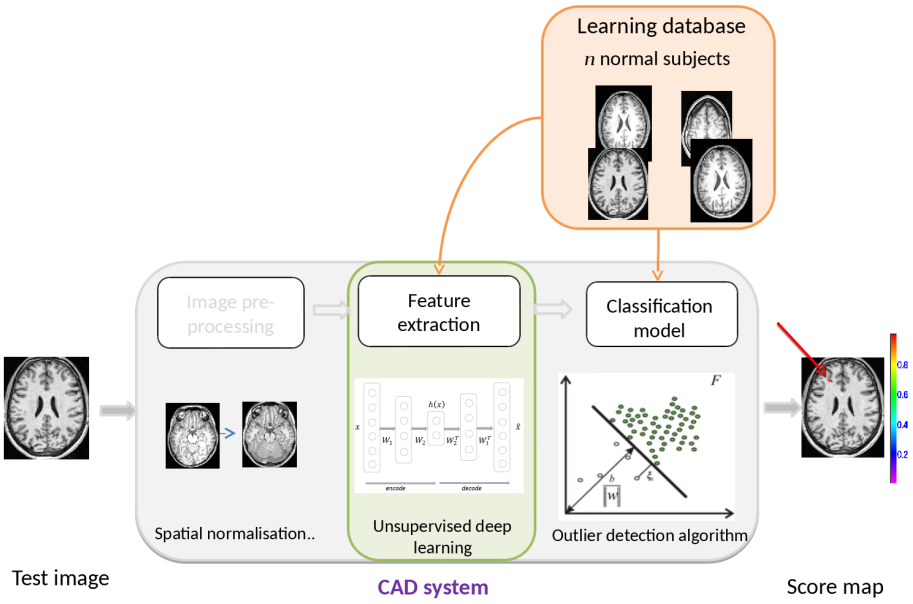


Figure 5.1: General representation of the CAD system.

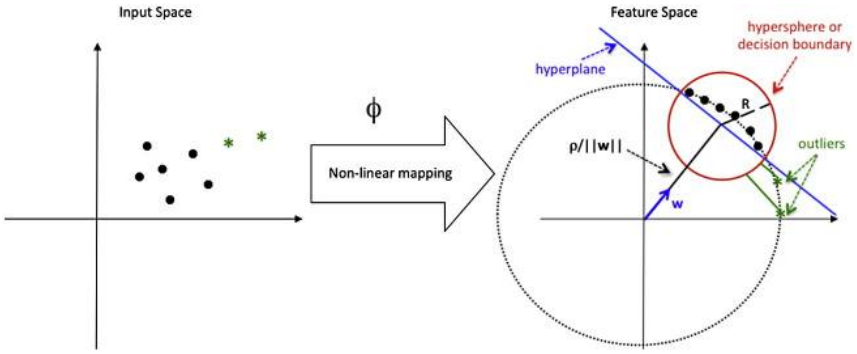
When the oc-SVM models are built for all the voxels in brain, a given test image can be run through the system. As an immediate output, the system produces a score map of the same size as the input image, where each voxel value is the signed score, output by the corresponding oc-SVM model for the representation corresponding to the test image patch, centered at the voxel. Below we present the components of the CAD system.

5.1.1 Data pre-processing

The first step of the CAD pipeline is the data pre-processing module. At this step, all the available acquisitions are first aligned to a common template which establishes a voxel-to-voxel correspondence between all the subjects. The detailed pre-processing routine is described in section 5.2.4. Next, the images are processed according to the format required in the second step of feature extraction. Typically, the images are turned into a stack of fixed size patches to be fed as input to the representation learning architectures. Moreover, depending on the type of architecture, the extracted patches may be fed to the network in a standalone fashion (monomodal architecture) or stacked in channels representing different modalities (multichannel architecture). Once the format of the input is decided upon, the corresponding data are introduced to the second component of the CAD system.

5.1.2 Feature extraction

The next component consists in learning representations for the provided input. This step is the main interest and focus of this work and therefore will be described in details in chapter 6, comprising the existing approaches and our contribution.



.

Figure 5.2: The principle behind the oc-SVM method. The points in the original space are projected into a higher dimensional space where their separation from the point of origin is sought through maximizing the margin. *Illustration from [Mourão-Miranda et al., 2011]*

5.1.3 Per-voxel outlier detection: oc-SVM

Principle

The one-class SVM (oc-SVM) introduced in [Schölkopf et al., 2001] is a particular case of the binary SVM which seeks to find a hyperplane separating the examples of the data set with different labels. In oc-SVM all the examples of the given data set $X = \{\mathbf{x}_i\}_{i=1,..,n}$ where $\mathbf{x}_i \in \mathcal{R}^d$ are normal/positive and therefore a hyperplane is sought to separate all the data points from the origin. Since most real-life data sets are not linearly separable in the original data space, the points are first mapped into a higher dimensional space through a mapping $\phi(\mathbf{x})$ with a corresponding kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the inner product. Kernels satisfying the Mercer's conditions [Vapnik, 1999] are guaranteed to have a corresponding mapping $\phi(\mathbf{x})$. Though different kernels may be chosen for the problem at hand, the most common choice is the RBF kernel due to its locality preserving properties. On the other hand, polynomial and sigmoid kernels seem to fail systematically in the scope of outlier detection [Bounsiar and Madden, 2014]. An illustration of the oc-SVM principle is shown on fig. 5.2.

Primal formulation

Let X be a set of n observations $X = \{\mathbf{x}_i\}_{i=1,..,n}$ where $\mathbf{x}_i \in \mathcal{X}$ and ϕ be a feature map $\mathcal{X} \to \mathcal{F}$ from the original space to a dot product space $\mathcal{F}$, corresponding to some kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

The oc-SVM algorithm seeks to project the data points into a new feature space corresponding to the kernel K and to separate them from the origin with maximum margin. To that end, the following problem is solved:

$$\begin{aligned} \min_{\mathbf{w},\rho,\xi_i} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \mathbf{w}\cdot\phi(\mathbf{x_i}) \geq \rho - \xi_i \qquad i \in [1,n] \\ & \xi_i \geq 0 \qquad i \in [1,n] \end{aligned} \tag{5.1}$$

where n is the number of training examples, $\mathbf{x_i}$ is the i-th example in the training data set X, ξ_i-s are slack variables relaxing the inequality constraints, $\mathbf{w}$ and ρ define the separating hyperplane, $\nu \in (0,1)$ is a parameter that sets a boundary to the fraction of outliers allowed. When the optimal solution $\mathbf{w}^*, \rho^*$ is found, the decision for an example $\mathbf{x}$ depends on the side of the hyperplane it falls in and is expressed with

$$f(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \phi(\mathbf{x}) - \rho^*)$$

The penalization of the slack parameters in the objective function assures that the decision function will yield 1 for most of the points in the training set with a reasonably small $||\mathbf{w}||$. The parameter ν controls the number of misclassified points.

Dual formulation

Often the optimal solution of the problem in 5.1 is found through its dual formulation. To arrive at the dual formulation, Lagrangian multipliers $\alpha, \beta \geq 0$ are introduced for each constraint. The Lagrangian form therefore is:

$$L(\mathbf{w},\xi,\rho,\alpha,\beta) = \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(\mathbf{w}\cdot\phi(\mathbf{x_i}) - \rho + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i \tag{5.2}$$

Setting the derivatives of the Lagrangian with respect to the primal variables $\mathbf{w}$, ξ and ρ to 0 yields the following:

$$\begin{aligned} \nabla_{\mathbf{w}}\mathcal{L} = 0 \quad &\Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x_i}) \\ \nabla_{\xi}\mathcal{L} = 0 \quad &\Rightarrow \quad \alpha_i = \frac{1}{\nu n} - \beta_i \\ \frac{\partial\mathcal{L}}{\partial\rho} = 0 \quad &\Rightarrow \quad \sum_{i=1}^{n}\alpha_i = 1 \end{aligned} \tag{5.3}$$

Introducing 5.3 into the Lagrangian 5.2, the dual formulation becomes

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i = 1 \\ & 0 \leq \alpha_i \leq \frac{1}{\nu n} \qquad i \in [1,n] \end{aligned} \tag{5.4}$$

The last inequality constraint stems from $\beta_i = \frac{1}{\nu n} - \alpha_i \geq 0$.

The solution of the dual problem yields an optimal α^* which can be used to recover the

optimal values of the primal variables. The simplest is the computation of $\mathbf{w}^*$ from the expression in 5.3.
We classify the training observations into 3 categories depending on the corresponding values of the Lagrangian multipliers. So,

1. $\alpha_i = 0$, hence $\beta_i = \frac{1}{\nu n}$. For these points $\mathbf{w} \cdot \phi(\mathbf{x}_i) > \rho$ and $\xi_i = 0$ in the primal formulation. These are the points that were correctly considered inliers (the slack variables are 0) and are referred to as *normal*.

2. $0 < \alpha_i, \beta_i < \frac{1}{\nu n}$. These are the points lying on the decision boundary. Their primal variable values are $\mathbf{w} \cdot \phi(\mathbf{x}_i) = \rho$ and $\xi_i = 0$. These points are called *(essential) support vectors (SV)*.

3. $\alpha_i = \frac{1}{\nu n}$. These are the points that were misclassified i.e. left outside of the decision boundary (the corresponding slack variables are positive $\xi_i > 0$). They are called *errors (non-essential support vectors)*.

It should be noted that the data points where $\alpha_i^* = 0$ do not contribute to the decision boundary and later to the decision function $f(\mathbf{x})$ for a new point $\mathbf{x}$, as can be seen from the first equation of 5.3. Moreover, for a support vector $\mathbf{x}_i$, $\rho^* = \mathbf{w}^* \cdot \phi(\mathbf{x}_i)$ and can be computed as

$$\rho^* = \mathbf{w}^* \cdot \phi(\mathbf{x}_i) = \sum_{j=1}^{n} \alpha_j^* K(\mathbf{x}_i, \mathbf{x}_j)$$

and the decision function for an example $\mathbf{x}$ therefore becomes

$$f(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \phi(\mathbf{x}) - \rho^*) = sgn(\sum_{i=1}^{n} \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) - \rho^*)$$

Solving the dual problem amounts to solving a quadratic programming (QP) problem. The main difficulty is to store the kernel matrix when its dimensionality is large. The most common tool for solving SVM problems is the LIBSVM library introduced in [Chang and Lin, 2011]. LIBSVM solves the dual QP problem in 5.4 using a decomposition method that relies on updating a subset of α coefficients in each iteration [Fan et al., 2005].

On the ν coefficient As it was shown in [Schölkopf et al., 2001], the ν coefficient in the formulation of oc-SVM is also an upper bound on the fraction of permitted errors and a lower bound on the fraction of support vectors. In other words, by setting $\nu = 0.01$ we allow around 1% of the training examples to be misclassified as outliers.

$$\frac{|errors|}{n} \leq \nu \leq \frac{|errors| + |SVs|}{n}$$

oc-SVM design

In the proposed pipeline each voxel is associated with a oc-SVM model. Applying the oc-SVM algorithm implies several design choices. First, the kernel function should be chosen. As such, we used the RBF kernel throughout this study, which is the most common kernel function used in oc-SVM problems. The kernel choice assumes appropriate values for its parameters. For the RBF kernel, the kernel width must be chosen. The details of the adopted heuristics for this parameter value are given in section 7.3.2. Eventually, the parameter ν, controlling the ratio of allowed outliers in the oc-SVM formulation 5.1, should be chosen. Our experimental choice for this parameter is given in section 7.3.2 as well. For a new test image, each voxel is assigned the signed score output by the corresponding oc-SVM, computed as $\mathbf{w}^* \cdot \phi(\mathbf{x}) - \rho^*$.

5.2 Data description

The proposed CAD system was evaluated on a set of patients with confirmed epilepsy lesions. 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. In this study we had access to patient data coming from the Neurological Hospital in Lyon, through our collaboration with Dr. J. Jung, in the scope of an ongoing research program PHRC (programme hospitalier de recherche clinique) initiated by Pr. F. Maugière and Dr. J. Jung. This research program is aimed at evaluating the diagnostic value of multimodal neuroimaging data in the pre-surgical evaluation of intractable epilepsy. The healthy control data was accessed through our collaboration in the scope of the same project.

5.2.1 Study group

The data set considered in this study consists in a training set of healthy subjects and a test set of epilepsy patients. The details of the data set are summarized in table 5.1.

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 and FLAIR MRI sequences. All but two had a PET exam as well. Additionally, the patients underwent intracranial EEG exam 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 T1-weighted and FLAIR MRI sequences. 35 of them had PET exams as well.

5.2.2 Imaging protocol

All the healthy controls and patients had 3D anatomical T1-weighted 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).
PET scans were conducted on a Biograph mCT PET-CT tomograph (Siemens). Subjects were positioned in the scanner such that the acquired planes would be parallel to the orbital-meatal line. Head movement was minimized with an airbag. A camera allowed a visual control over the head position during the acquisition. Measures for tissue and head support attenuation were performed with a 1min low-dose CT scan acquired before emission data acquisition. A dynamic emission scan was acquired in list mode during 60 min after the injection. Static ^{18}F-FDG uptake images were reconstructed for 50 to 60 min post injection using 3D-ordinary Poisson-ordered subset expectation maximization iterative algorithm (12 iteration, 21 subsets) incorporating point spread function and time of flight (with a Gaussian filter of 4mm) after correction for scatter and attenuation. Reconstructed volumes consisted of 109 contiguous slices (2.03mm thickness) of 200 x 200 voxels (2.036 x 2.036mm^2). Actual resolutions for reconstructed images were approximately 2.6mm in full width at half maximum in the axial direction and 3.1mm in full width at half maximum in the transaxial direction measured for a source located 1cm from the field of view [Jakoby et al., 2011].

	# of controls	**# of patients**	**T1** (1.5T Siemens Sonata)	**FLAIR** (1.5T Siemens Sonata)	**PET** (mCT PET-CT Siemens tomograph)
DB1	35	19	160 x 192 x 192 1.2mm cubic voxels	176 x 196 x 256 1.2mm cubic voxels	109 x 200 x 200 2.036mm cubic voxels
DB2	40	2	160 x 192 x 192 1.2mm cubic voxels	176 x 196 x 256 1.2mm cubic voxels	–
Total	**75**	**21**			

Table 5.1: Summary of the data set obtained through our collaboration with Dr. *J. Jung*.

5.2.3 Patient lesion location reference

The information on the patients' epileptogenic lesions is summarized in table 5.2. The table details the clinical justification on the true lesion localization. As it can be seen, all patients had an intracranial EEG exam while most of them had a resective surgery and became seizure free at most 6 months after the surgery. A few patients had thermocoagulation instead, which happened to be successful and therefore confirmed the EEG results. For most patients, the encountered lesions did not fall under any common epilepsy

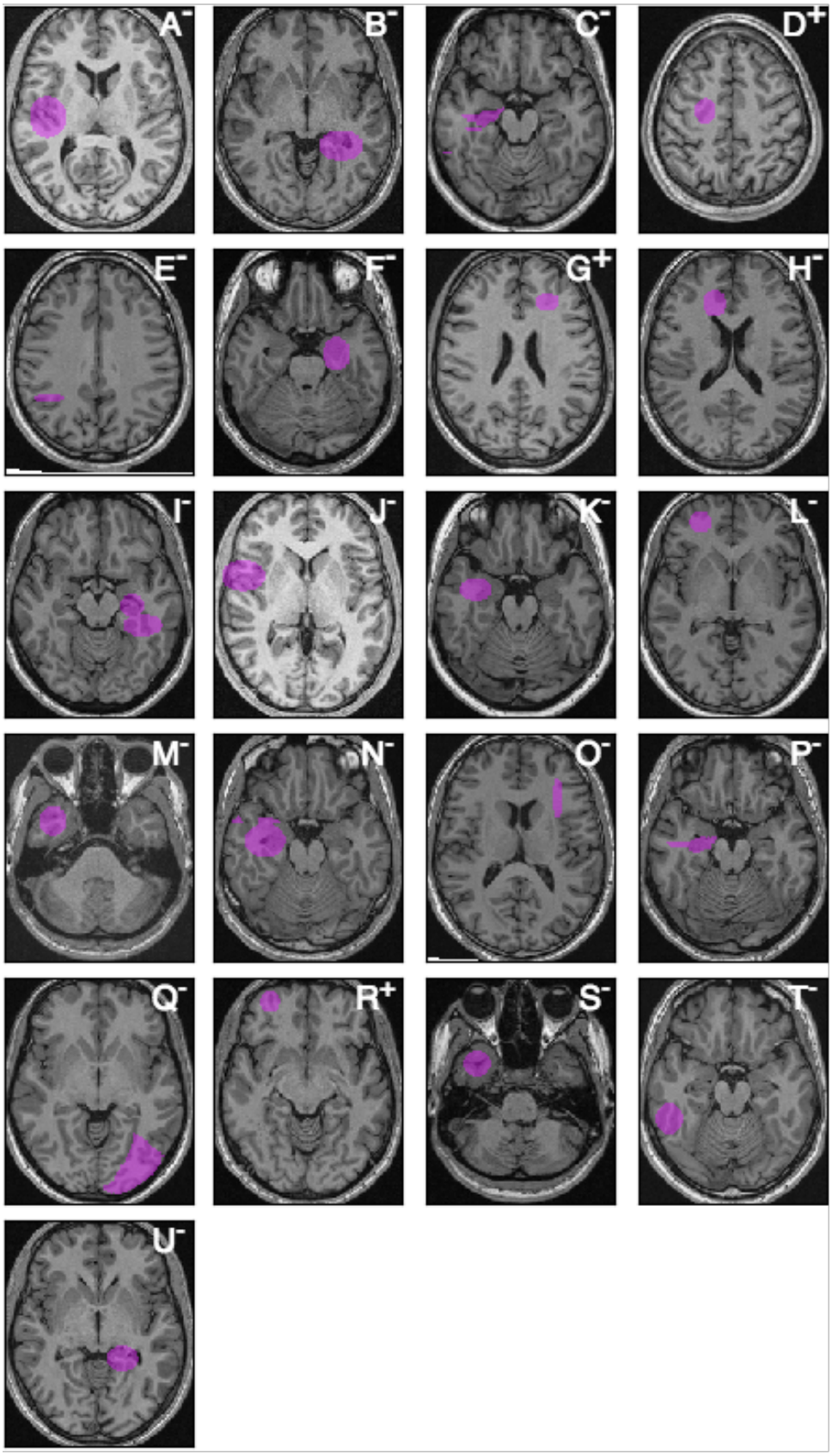


Figure 5.3: The ground truth annotations shown in purple circles overlaid onto the transverse slices of patients' T1-weighted MRI.

Figure 5.4: A slice of the maximum probability atlas (left) and the resulting volume of interest (right).

cause/category, as introduced in section 3.1. Typically, histopathological analysis of cryptogenic epilepsy patients is informative with respect to the lesion category in only 30-50% cases [Bernasconi et al., 2011]. Therefore, their lesion type is marked as *unknown*.
Of the 21 patients, 3 had epilepsy lesions visually detectable on the pre-surgical MR scans (MRI^+). However, only one (patient D^+) was detected on T1-w MRI. The MR images of the remaining 18 patients were considered normal on *multiple examinations* i.e. no lesion was detected visually (pure MRI^-). For the visually remarkable lesions, the manual annotations were obtained by contouring the visible lesions on the raw images. For the MRI^- patients, the ground truth references were traced by an expert neurologist after carefully verifying the corresponding intracranial EEG results as well as the post-surgical report and MR scans following surgery or thermocoagulation. The obtained ground truth annotations, projected onto the corresponding T1-w MR transverse slices, are illustrated on fig. 5.3. Localizing the true lesion in such manner naturally comes with a lack of precision as the precise delineation of the lesion is still impossible. Instead, the surgically resected area is considered a slightly "generous" expansion of the lesion localization. In both cases the evaluation of the CAD system proceeds as follows: if a detected lesion is largely located in the defined "reference" zones, the detected lesion is considered a true positive (TP). Otherwise it is considered a false positive (FP).

5.2.4 Data pre-processing

For the proposed CAD system approach to be advantageous, certain pre-processing steps are necessary. All the pre-processing steps have been done using the SPM8 software [Ashburner, 2009], a common tool in the neuroimaging community. The volumes were first spatially normalized with the unified segmentation algorithm (UniSeg) [Ashburner and Friston, 2005] implemented in SPM 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

Patient	**Lesion location**	**Lesion location confirmed with**	**Lesion type**	Age
Patient A^-	Insula R	Intracranial EEG & successful thermocoagluation	Unknown	17
Patient B^-	Temporal Lobe L	Intracranial EEG & surgical success	Unknown	32
Patient C^-	Hippocampus R	Intracranial EEG & surgical success	Histopathology: FCD type III with HS	41
Patient D^+	Superior frontal gyrus R	Intracranial EEG & surgical success	FCD type II	21
Patient E^-	Inferiolateral remainder of parietal lobe R	Intracranial EEG & surgical success	Unknown	25
Patient F^-	Hippocampus L, parahippocampus L	Intracranial EEG & surgical success	Unknown	28
Patient G^+	Middle frontal gyrus L	Intracranial EEG & successful thermocoagluation	FCD type II	43
Patient H^-	Superior frontal gyrus R	Intracranial EEG & surgical success	Unknown	29
Patient I^-	Hippocampus L, parahippocampus L	Intracranial EEG & surgical success	Unknown	41
Patient J^-	Precentral gyrus R	Intracranial EEG & surgical success	Unknown	19
Patient K^-	Superior temporal gyrus R	Intracranial EEG & surgical success	Unknown	44
Patient L^-	Middle frontal gyrus R	Intracranial EEG & surgical success	Unknown	25
Patient M^-	Anterior temporal lobe R	Intracranial EEG & surgical success	Unknown	25
Patient N^-	Anterior temporal lobe R Hippocampus R	Intracranial EEG & surgical success	Unknown	26
Patient O^-	Middle frontal gyrus L	Intracranial EEG & surgical success	Unknown	33
Patient P^-	Hippocampus R	Intracranial EEG & surgical success	Histopathology: FCD type III with HS	41
Patient Q^-	Lateral remainder of occipital lobe L	Intracranial EEG & surgical success	FCD type II	29
Patient R^+	Orbital gyrus R	Intracranial EEG & surgical success	Ganglioglioma	47
Patient S^-	Anterior temporal lobe R Hippocampus R	Intracranial EEG & surgical success	Histopathology: FCD type IIIa	31
Patient T^-	Posterior temporal lobe R	Intracranial EEG & surgical success	Unknown	36
Patient U^-	Posterior temporal lobe L	Intracranial EEG & surgical success	Unknown	18

Table 5.2: Summary of the epileptogenic lesions found in the patient group.

1 x 1 mm and the default parameter values. This step assures the voxel-level correspondence between all the subjects. Further, the other imaging modalities, namely FLAIR and PET sequences, were rigidly co-registered to the individual T1-w 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 and PET 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]. A slice of the maximum probability atlas is shown on fig. 5.4. After the elimination of the corresponding voxels, the number of remaining voxels adds up to around 1.5 million per volume.

Chapter 6

Unsupervised representation learning for anomaly detection

This chapter introduces the considered unsupervised architectures that could be applied in the CAD pipeline developed in chapter 5. We start by reviewing the existing and commonly used deep models. Moreover, we present the most relevant state-of-the-art studies where those architectures and their variations were exploited. In particular, we review the recent works where various unsupervised deep architectures were used for the problem of outlier detection, the context of this work. Eventually, we present a novel configuration of a siamese network, better suited for the context of anomaly detection when solely healthy / positive examples are available for training.

6.1 Unsupervised deep learning architectures

6.1.1 Autoencoders

A basic autoencoder is a one-hidden-layer neural network composed of two essential parts - an *encoder* and a *decoder* [Hinton and Zemel, 1994]. The encoder is a mapping that is applied on an input $\mathbf{x} \in R^d$ in order to transform it to a hidden representation $\mathbf{h} \in R^{d'}$ (typically, $d' < d$). Usually it is modeled with a non linear function applied to the affine transformation of the input i.e.

$$\mathbf{h} = f(W\mathbf{x} + \mathbf{b})$$

where W is a $d' \times d$ weight matrix and $\mathbf{b} \in R^{d'}$ is a bias vector associated with the mapping. f is the (non)-linear function of choice.
The decoder is the inverse mapping from the hidden representation space to the original

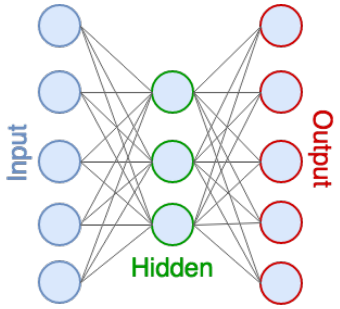


Figure 6.1: Autoencoder with a single hidden layer.

space with a similarly configured transformation i.e.

$$\hat{\mathbf{x}} = f'(W'\mathbf{h} + \mathbf{b}')$$

Frequently the weight matrices W and W' are chosen to be *tied* i.e. $W' = W^T$ in order to reduce the number of parameters. The parameter set $\Theta = \{W, \mathbf{b}, W', \mathbf{b}'\}$ is optimized so as to minimize the reconstruction error - a measure of deviation of $\hat{\mathbf{x}}$ from the original input $\mathbf{x}$ (e.g. the mean squared error or cross-entropy) - across all the instances of a given data set D composed of N examples

$$L(D;\Theta) = \frac{1}{N}\sum_{i=1}^{N} c(\mathbf{x}_i, \hat{\mathbf{x}}_i)$$

where $\mathbf{x}_i$ is the i-th example, $\hat{\mathbf{x}}_i$ is its reconstruction and c is the chosen measure of the reconstruction error. An autoencoder is illustrated on fig. 6.1.
Under some configurations it is possible for an autoencoder to learn the identity mapping. The latter is not an interesting objective as it means the network does not learn any meaningful/useful representation of the input. Therefore, employing such constraints as setting $d' < d$, tied weights and non-linear functions in the encoder-decoder mappings (e.g. sigmoid, ReLU, etc) has become the typical scenario.

Autoencoders can be stacked in a layer-wise manner where a sequence of layers performs the encoding and another sequence decodes the middle-layer representation. Precisely, the latent representation of one layer serves as input of the next one

$$\mathbf{h}^k = f(W^k\mathbf{h}^{k-1} + \mathbf{b}^k)$$

Stacked autoencoders allow to learn more complex mappings and, hence, more abstract representations in the middle layer.

6.1.1.1 Denoising autoencoders

A denoising autoencoder [Vincent et al., 2008] is a variation of autoencoders whose task is to recover the clean input from its corrupted version. Precisely, at input it is fed a

corrupted version $\tilde{\mathbf{x}}$ of $\mathbf{x}$ and its loss function measures the deviation of the reconstruction from the corresponding uncorrupted example

$$L(D;\Theta) = \frac{1}{N}\sum_{i=1}^{N} c(\mathbf{x}_i, \hat{\tilde{\mathbf{x}}}_i)$$

where $\mathbf{x}_i$ is the i-th example and $\hat{\tilde{\mathbf{x}}}_i$ is the reconstruction obtained for the corrupted input $\tilde{\mathbf{x}}$.

The common choices of input 'corruption' are salt and pepper noise (randomly setting a certain number of elements in the input vector to their possible minimum or maximum value), masking noise (randomly setting a certain number of elements in the input vector to 0) and adding Gaussian noise to the input. The reconstruction error may be designed in a way as to weigh separately the contribution made by the corrupted and uncorrupted elements of the examples. Such an autoencoder is called *emphasized denoising autoencoder*. A well-trained denoising autoencoder is capable of recovering the corrupted part of the input as it is trained to recognize structures inherent to the input and their relationships [Vincent et al., 2010]. This fact makes denoising autoencoders a good candidate for representation learning for various tasks.

6.1.1.2 Convolutional autoencoders

The main disadvantage of simple autoencoders is that they ignore the spatial relationships present in 2D or 3D images and are prone to learning redundant features. On the other hand, convolutional neural networks exploit the convolution operator and the shared weights so as to learn features that are common in various locations of the image. Convolutional autoencoders, hence, are built upon the same idea of shared weights that are applied to all the locations in the image. More precisely, the input image $\mathbf{x}$ is mapped to a number of *feature maps* where each feature map $\mathbf{h}^k$ is computed as

$$\mathbf{h}^k = f(\mathbf{x} * W^k + \mathbf{b}^k)$$

where W^k is the shared weight matrix, $\mathbf{b}^k$ is the shared bias, f is a chosen activation function and $*$ denotes the convolution operation. The reverse mapping to the input space is done similarly. The parameters are tuned to optimize, as in the case of basic autoencoders, a chosen loss function. Naturally, convolutional autoencoders too can be stacked to form more complex architectures.

6.1.1.3 Variational autoencoders

A variational autoencoder (VAE) introduced in [Kingma and Welling, 2013] is a directed probabilistic graphical model of two variables - an observed $\mathbf{x}$ and an unobserved $\mathbf{z}$. The variable $\mathbf{z}$ is generated from some prior distribution $p_\theta(\mathbf{z})$ while $\mathbf{x}$ is generated from

some conditional $p_\theta(\mathbf{x}|\mathbf{z})$. The posterior $p_\theta(\mathbf{z}|\mathbf{x})$ is intractable; therefore $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced as its approximation. Ideally, the objective of the graphical model is to maximize the log likelihood of a given data set $\mathbf{X} = (\{\mathbf{x}_1, ..., \mathbf{x}_n\})$ composed of n observations i.e. $\sum_{i=1}^{n} \log p_\theta(\mathbf{x}_i)$ where $\log p_\theta(\mathbf{x}_i)$ can be written as

$$\log p_\theta(\mathbf{x}_i) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)||p_\theta(\mathbf{z}|\mathbf{x_i})) + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(\mathbf{x_i}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)||p_\theta(\mathbf{z}))}_{\mathcal{L}(\theta,\phi;\mathbf{x}_i)}$$

where D_{KL} is the Kullback-Leibler divergence [Kullback and Leibler, 1951]. Due to the intractability of the first term, the framework instead optimizes $\mathcal{L}(\theta, \phi; \mathbf{x}_i)$ which is the *lower bound* of the overall likelihood (as KL-divergence is non-negative).
The second term of $\mathcal{L}(\theta, \phi; \mathbf{x}_i)$ imposes the similarity of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior on $\mathbf{z}$ while the first term can be interpreted as the reconstruction of $\mathbf{x}$ through the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{z})$.
In variational autoencoders, the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ is modeled through a neural network. In order to update the parameters through backpropagation, the reparameterization trick is applied which allows to replace the expectation over $\mathbf{z}|\mathbf{x}$ with an expectation over a noise variable (which is used in a transformation producing the reparameterized random variable $\mathbf{z}$). As opposed to regular autoencoders, the encoder component in a VAE produces an estimate of the parameters of the distribution over $\mathbf{z}$ and not the distribution itself. The produced parameter estimates, together with the introduced noise variable, are used to sample from $\mathbf{z}$ (which is why the VAE encoder and decoder are called *probabilistic*). For example, a common setting is to consider that $\mathbf{z}$ is drawn from a multivariate Gaussian and have the VAE encoder output its mean μ and standard deviation σ which are then used to draw a sample $\mathbf{z} = \mu + \sigma \cdot \epsilon$, with ϵ is the noise variable drawn from a normal distribution. Moreover, in this setting the KL-divergence has the following closed form

$$-D_{KL}(q_\phi(\mathbf{z})||p_\theta(\mathbf{z})) = 0.5 \sum_{m=1}^{M} [1 + \log(\sigma_m^2) - \mu_m^2 - \sigma_m^2]$$

where M is the dimensionality of the vector $\mathbf{z}$. A VAE corresponding to this setting is shown on fig. 6.2.
The advantage of VAEs is that they can be used to generate examples, acting as a typical generative model. Moreover, when trained properly, the latent representation can reveal novel information on the data and be used in other auxiliary tasks.

6.1.1.4 Recent applications

Autoencoders and their variations have been used in various contexts and applications. Various studies, such as [Vincent et al., 2010, Masci et al., 2011] proposed to use autoencoders to pre-train deep architectures, otherwise said, to initialize the weights of the layers of the architecture at hand. This technique, however, has recently lost its widespread use.

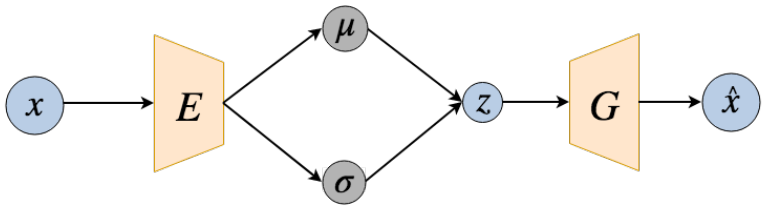


Figure 6.2: VAE framework when $\mathbf{z}$ is drawn from a multivariate Gaussian distribution. E denotes the encoder, G - decoder, μ and σ are the parameters of $q(\mathbf{z})$ output by the encoder.

In [Gehring et al., 2013], the authors used stacked autoencoders, that were first pretrained and later combined with additional layers, to train a classification system on speech data. Stacked denoising autoencoders were considered for noise reduction and speech enhancement in [Lu et al., 2013]. In another study [Deng et al., 2010], binary codings of speech spectrograms are learnt with deep autoencoders. Super-resolution problems, recovering a high resolution image from its low resolution version, are another application where autoencoders have been used frequently. For super-resolution problems, approaches, mainly involving autoencoder-based complex architectures, were proposed in [Cui et al., 2014, Zeng et al., 2017]. Denoising autoencoders were used in [Xie et al., 2012] for image denoising and inpainting while convolutional autoencoders were applied to the image restoration problem in [Mao et al., 2016].
In many tasks, autoencoders have been used as feature extracting modules, further to be combined with various classification models. So, in [Xing et al., 2016], the authors evaluate the potential of features learnt with stacked denoising autoencoders on hyperspectral imaging data, by feeding the produced representations into a SVM classifier. In several studies autoencoders are explored for anomaly detection. Such studies will be presented in section 6.2.

6.1.2 Generative adversarial networks

Generative Adversarial Networks (GAN) were first introduced in [Goodfellow et al., 2014] and have since seen a variety of extensions. A GAN is composed of two components (neural networks) - a *generator* G and a *discriminator* D. The generator's objective is to produce examples as realistic as the authentic ones in the training data set. Those generated examples, together with the real ones, are the input of the discriminator which aims at distinguishing perfectly the generated and the real images. Training a GAN implies improving the generator, so that the discriminator does not succeed at distinguishing the generated input from the observations in the data set, and at the same time improving the discriminator's performance. This corresponds to a typical two-player minimax game with a unique solution G^* reproducing the underlying data distribution and D^* equal to

1/2 everywhere. The objective of the game therefore is:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - D(G(\mathbf{z})))]$$

where $D(\mathbf{x})$ can be interpreted as the probability of $\mathbf{x}$ being real. Eventually, minimizing the value function amounts to minimizing the Jensen-Shannon divergence between the real distribution p_{data} and the distribution p_g modeled by the generator.

A GAN can be trained with a stochastic gradient and backpropagation, by updating the parameters of G and D. The GAN training, however, is not trivial because of its instability; frequently GANs do not converge or generator collapses into a single mode or generates unrealistic outputs. For one, the loss change over iterations does not necessarily correlate with the model convergence; the loss oscillates frequently, unlike in other deep architectures. An important contribution was the introduction of the DCGAN by [Radford et al., 2015], a reasonably stable architecture which gave a first glimpse on the potential of GANs. It should be noted that while it is straightforward to sample an example $\mathbf{x}$ given some $\mathbf{z}$, the reverse mapping is not achieved within GANs.

Further works proposed different versions/extensions of the basic GAN to overcome the instability issues, have a more informative loss and a reverse mapping to obtain a $\mathbf{z}$ given $\mathbf{x}$. So, Wasserstein GAN (WGAN) in [Arjovsky et al., 2017] revolves around the Wasserstein distance between the real distribution p_{data} and p_g and its properties. Exploiting the Kantorovich-Rubinstein duality [Villani, 2008], which can be expressed as an adversarial objective, it follows that if the discriminators were designed to model K-Lipschitz functions, the optimal Wasserstein distance between the real and generated distributions could be attained (up to a multiplicative factor). In [Arjovsky et al., 2017] weight-clipping was performed to assure the K-Lipschitz constraint on the discriminator. It did not solve the instability issues so other approaches were proposed. In [Gulrajani et al., 2017], the authors proposed to penalize the norm of the gradient of the discriminator with respect to its input and in [Salimans et al., 2016], the discriminator is enhanced with a *minibatch layer* measuring the closeness of the discriminator's representations of the examples in a minibatch. [Dumoulin et al., 2016] presented another framework, *ALI* (Adversarially Learnt Inference), where the discriminator aims at distinguishing jointly an input $\mathbf{x}$ and its encoded representation $\mathbf{z}$. One advantage is that the inverse mapping from $\mathbf{z}$ to $\mathbf{x}$ in this framework is explicit.

6.1.2.1 Wasserstein autoencoder

Despite the promising potential of GANs, the complicated training puts a halt on their widespread use, especially in the medical domain. Much effort was therefore invested in the development of related frameworks allowing to leverage the generative potential of GANs while ensuring a stable training. One such approach is Wasserstein autoencoder [Tolstikhin et al., 2017]. Wasserstein autoencoder (WAE) is a regularized autoencoder

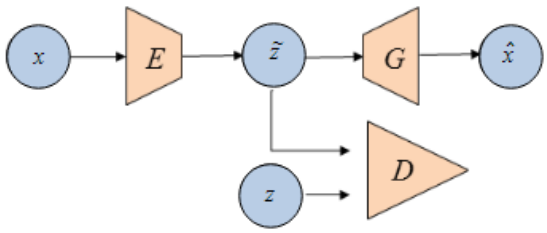


Figure 6.3: Wasserstein autoencoder (WAE) composed of an encoder E, a decoder G and a (adversary) module D to estimate the discrepancy between P_Z and Q_Z.

with a cost function similar to that of VAEs and a generative power resembling that of GANs. As illustrated on fig. 6.3, a WAE is composed of three components: an encoder E mapping an input from the data space $\mathcal{X}$ to the latent space $\mathcal{Z}$, a decoder G mapping a latent code from the latent space $\mathcal{Z}$ to the data space $\mathcal{X}$, and a module D that tries to minimize the discrepancy between the prior distribution of the latent code P_Z and the latent distribution Q_Z produced by the encoder. The resulting loss function can be expressed as

$$L(X;\Theta_{WAE}) = \frac{1}{N}\sum_{i=1}^{N} c(\mathbf{x_i},\hat{\mathbf{x}}_\mathbf{i}) + \beta \cdot D_Z(P_z,Q_z) \tag{6.1}$$

where D_Z measures the discrepancy between a given distribution P_z and Q_z for the data set $X = \{\mathbf{x_i}\}_{1,..,N}$ 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. When c is the squared error $c(\mathbf{x_i},\hat{\mathbf{x}}_\mathbf{i}) = ||\mathbf{x} - \hat{\mathbf{x}}_\mathbf{i}||_2^2$ and D_Z is the GAN objective, the WAE matches the Adversarial Autoencoders introduced in [Makhzani et al., 2015]. WAE is advantageous in practice as it comes with a built-in encoder-decoder architecture allowing the bidirectional mapping $\mathcal{X} \leftrightarrow \mathcal{Z}$. Moreover, examples can be generated by sampling from P_z and feeding it into the decoder.

6.1.2.2 Recent applications

GANs are rarely used as feature extraction modules in practical applications. First, the reverse mapping from the input space to the representation space is not explicitly present. To perform this mapping, two options are possible. The first one is to find a representation for a given input $\mathbf{x}$ through an iterative search of $\mathbf{z}$ that minimizes the deviation of a generated $\hat{\mathbf{x}}$ given $\mathbf{z}$ and the input $\mathbf{x}$. This requires iterating through the latent representation space until the optimal point is reached, for each given input, and is not practical to use. The second option is to train another network performing the inverse mapping of the generator. This requires the training of an additional network, but once done, extracting the corresponding representation of a given input is straightforward and

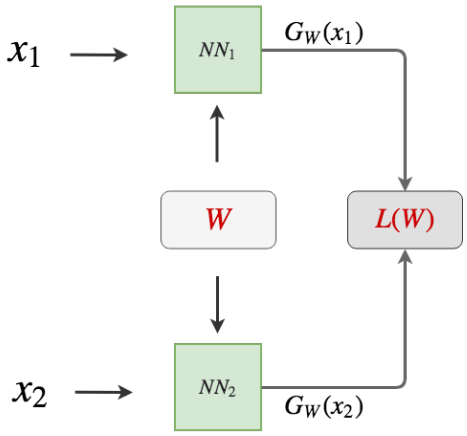


Figure 6.4: Siamese network.

efficient. On the other hand, the representations learnt in the discriminator layers were occasionally used to demonstrate the advantages of GANs as in [Wu et al., 2016].
GANs, however, have been applied in other applications for their generative capability. So, GANs were used for super-resolution image synthesis in [Ledig et al., 2017] where the discriminator aims to distinguish between real high resolution images and those generated from the low resolution images through the generator. Another work aims at text-to-image generation where textual information is introduced in the generator and the discriminator so as to generate/discriminate images conditioned to their textual content [Reed et al., 2016]. In medical imaging, recent works explored cross-modality synthesis which will be addressed in more details in section 10.2. Another important application is the outlier detection context using adversarial training. Such studies will be presented in section 6.2.

6.1.3 Siamese neural networks

Siamese neural networks were first introduced in [Bromley et al., 1993] for the problem of signature verification. A siamese network is composed of two sub-networks, with identical architecture and a shared parameter set, and a cost module, as illustrated on figure 6.4. It receives at input a pair of examples and propagates each of them through the corresponding sub-network that represents a function G_W parameterized with W. The sub-networks yield representations that are passed to the cost module. The main objective of such a system is to find a function G_W that maps examples in the original space to a space where their distance is small if they are 'close' in the input space and large otherwise. The 'close' input examples are referred to as *similar* and 'far' examples are referred to as *dissimilar*. With a well chosen loss function, the system learns to map similar inputs close to each other with respect to some simple measure and pulls apart the dissimilar examples. Moreover, no explicit distance measure in the input space is required.

Formally, a data set D is composed of pairs of examples and a label standing for their similarity/dissimilarity $D = \{(x_{i1}, x_{i2}, l_i)\}_{i=1,..,N}$ where l_i is equal to 0 when the pair is composed of similar inputs and 1 otherwise. A typical loss function in such a network is

$$L(D; W) = \frac{1}{N} \sum_{i=1}^{N} [(1 - l_i) \cdot L_S(D_W^i) + l_i \cdot L_D(D_W^i)] \tag{6.2}$$

where L_S is the loss function for similar examples and L_D is the loss function for dissimilar examples. D_W is a distance function between the representations learnt by the sub-networks for inputs and D_W^i designates its value for the i-th pair $D_W^i = D_W(G_W(\mathbf{x_{i1}}), G_W(\mathbf{x_{i2}}))$.
[Chopra et al., 2005] introduces a contrastive energy function composed of two terms decreasing the energy of similar pairs and increasing the energy of the dissimilar pairs for face verification. [Hadsell et al., 2006] proposes the following loss function:

$$L(D; W) = \frac{1}{N} \sum_{i=1}^{N} [(1 - l_i) \cdot \frac{1}{2}(D_W^i)^2 + l_i \cdot \frac{1}{2} max\{0, (m - D_W^i)\}^2]$$

where $D_W^i = ||G_W(\mathbf{x}_{i1}) - G_W(\mathbf{x}_{i2})||^2$.
[Simo-Serra et al., 2015] used a similar loss in the task of patch correspondence, with the sub-networks being convolutional networks. [Zagoruyko and Komodakis, 2015] explored siamese and pseudo-siamese (the sub-network weights are not shared) networks for the same problem, with a hinge-loss based objective function while [Han et al., 2015] minimized the cross-entropy error between the similarity label ($\{0, 1\}$) and the softmax activation computed on the values output by the sub-networks i.e.

$$L(D, W) = -\frac{1}{N} \sum_{i=1}^{N} [l_i \cdot log(\hat{l}_i) + (1 - l_i) \cdot (1 - log(\hat{l}_i))]$$

where $\hat{l}_i = \frac{e^{G_W(\mathbf{x}_{i2})}}{e^{G_W(\mathbf{x}_{i1})} + e^{G_W(\mathbf{x}_{i2})}}$.

[Bertinetto et al., 2016] exploited a fully convolutional siamese neural network for object tracking. In this case, each sub-network is a fully convolutional network producing a real-valued score. This score is later matched against the similarity label ($\{+1, -1\}$) in a logistic loss.
[Taigman et al., 2014] used siamese networks for face verification, by defining D_W distance as a weighted sum of unit-wise absolute differences of yielded representations for the pair (the coefficients being trainable parameters). [Zheng et al., 2016] defined a triangular similarity metric (closely related to the cosine similarity) in a siamese network with multi-layer perceptron sub-networks. Siamese networks have also been exploited in the context of one-shot learning by [Koch et al., 2015]. Other applications of this network include gesture classification [Zheng et al., 2016], text classification [Yih et al., 2011], speaker-specific information learning from speech [Chen and Salman, 2011], question retrieval [Das et al.,

2016] and sketch-based shape retrieval [Wang et al., 2015].
[Zeghidour et al., 2016] extended the classical setup of siamese networks to accommodate two similarity labels - speaker and phonetic similarities - and further used *triamese* networks that receive at input triplets composed of an example, its similar pair and its dissimilar pair at once. In the scope of face verification problems, some authors [Parkhi et al., 2015, Schroff et al., 2015] proposed the so called *triplet* networks analogous to *triamese* networks. The input to this network is a triplet (a, p, n) containing an *anchor* image, a *positive* image p of the same person, different from a and a *negative* image n belonging to a different person. The *triplet* loss hence is designed to bring closer the anchors and their positives and pull apart the former and their negatives. Though the underlying principle is shared between siamese and triplet/triamese networks, their difference lays in that the notion of similarity/dissimilarity coded with labels is replaced with direct triplet input examples.
It should be noted that while siamese networks do not require labels per example, similarity labels for input pairs are expected. Those similarity labels can be obtained by using specific labels of individual examples when available or by a sort of self-supervision by defining similar/dissimilar pairs as those whose distance in the input space is below/above some threshold.

6.2 Unsupervised deep learning and anomaly detection

Deep learning architectures have recently been exploited in many studies for anomaly detection. Two main strategies could be distinguished among the current methods.
The first group consists of reconstruction-based anomaly detection methods (introduced in 1.1.3). In such methods, the anomaly detection is based on some metrics quantifying the discrepancy between the original input and the reconstruction obtained with the chosen network. The examples with large deviations are considered outliers. In particular, variational autoencoders (VAE) have been used in a number of studies for anomaly detection. [An and Cho, 2015] used the reconstruction probability produced by a VAE trained on normal examples. [Xu et al., 2018] used a similar VAE-based approach to detect anomalies in web applications. [Munawar et al., 2017a] used the autoencoder reconstruction error to detect anomalies by training the network not only to minimize the reconstruction error of normal examples but also to maximize the same error for outliers. [Munawar et al., 2017b] used long short-term memory networks (LSTM) to predict a frame in a video given the previous frames and revealed anomalies based on the difference between the predicted and the actual frames. A more recent tendency exploits adversarially trained generative networks. An important contribution was the approach proposed in [Schlegl et al., 2017] where the authors defined a score function that measures how anomalous a given sample is based on the reconstruction and discrimination losses estimated with a GAN architecture

trained on normal samples only. Further works attempted to improve this approach. For example, [Hirose et al., 2017] used the same approach for robot navigation enhanced with an inverse generator that maps images to the representation space the generator produces images from (GAN architecture lacks such an explicit mapping, therefore the original work used an additional optimization step to approximate it). [Zenati et al., 2018] went further and replaced the proposed GAN architecture with a Bidirectional Generative Adversarial Network (BiGAN, [Donahue et al., 2016]) with the same score function for some general outlier detection tasks.

The advantage of such approaches is that the networks are trained in end-to-end fashion. The methods, however, suit well to the contexts where a network, trained on normal examples only, would fail to reconstruct outliers due to the significant differences between them. When it comes to subtle brain lesions, such an approach would not be realistic. First, the images containing lesions, especially when they were considered normal over a visual analysis, do not differ significantly from the healthy ones. It is not plausible, therefore, to assume that the reconstruction error of a lesionous area will be larger than that of any other anatomical difference between subjects. Moreover, if the network is trained on patches of images, the assumption would fail to account for the abnormalities resulting in normal patterns occurring in wrong parts of the brain (e.g. heterotopia, where normal looking gray matter cells end up within the white matter). Since the patterns would have been 'seen' and therefore learnt by the network, the reconstruction may be close to perfect.

In this work we consider another strategy employed in recent studies which consists in learning representations with a deep network and couple them with some outlier detection algorithm. So, [Erfani et al., 2016] builds on this approach, by first learning latent representations of normal samples with deep belief networks and then feeding the learnt representations to a one-class SVM model in order to estimate the boundaries of the normal examples. [Xu et al., 2015] proposed a framework for anomaly detection on image and video-based surveillance data by combining the decisions of three oc-SVMs trained on representations learnt with stacked denoising autoencoders on images and videos individually and the two data sources combined. Similarly, [Huang et al., 2018] trained a single oc-SVM on the concatenated representations of three convolutional restricted Boltzmann machines for energy, visual and motion data modalities. In [Seeböck et al., 2016] a deep convolutional autoencoeder is coupled with oc-SVM for anomaly detection in retinal imaging. The usefulness of the learnt features is additionally verified through classifying intraretinal cystoid fluid, subretinal fluid and the remaining part of the retina.

Among the studies mentioned above only a small number was developed for medical applications [Seeböck et al., 2016, Schlegl et al., 2017].

6.3 Contribution: Regularized siamese network with deep convolutional autoencoders

From the review of the architectures and their applications above, it seems that autoencoders are well-adapted for feature extraction, later to be coupled with additional models, specific to the task. On the other hand, siamese networks offer to learn a mapping to a representation space where the similarity of the data points is imposed. We will therefore make an attempt to leverage both autoencoder and siamese network structures in a unified framework by adapting them to the outlier detection problem.

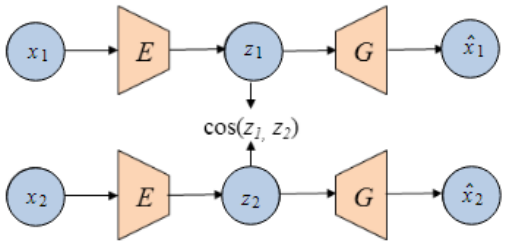


Figure 6.5: Regularized siamese network.

Given a data set $X = \{\mathbf{x}_i\}_{i=1,..,n}, \mathbf{x}_i \in \mathcal{R}^d$ composed of n normal points, we seek to find a mapping $G_\theta : \mathcal{X} \to \mathcal{Z}$ to project the original points to a representation space where the given examples form a close neighborhood. Such a mapping will be modeled with a neural network. The main task is therefore to assure the similarity of the representations.
To this end we propose to employ the underlying idea of siamese networks. Having a *similar* pair at input, the siamese subnetworks map them to a representation space where their proximity is imposed by the cost module. This coincides with our objective. However, the data set X is composed of normal examples only. While it is only natural to pair random couples of normal examples and label them similar, the notion of *dissimilar* pairs would not be defined. We will therefore introduce a regularizing term which will impose a certain structure on the representations produced by the mapping G_θ so that they do not collapse to a single point. We will refer to the proposed framework as a *Regularized Siamese Network*. It is important to note that the major difference with the classical siamese networks lays in the absence of dissimilar pairs and therefore the contrastive term typical to siamese cost modules (as in the expression 6.2). Moreover, while classical siamese networks require some level of supervision in order to provide the cost module with pair similarity labels, in our context of outlier detection all points are similar and no additional supervision is needed. Below we give a formal specification of the proposed model.
Fig. 6.5 shows the proposed framework. It receives at input a pair of normal points $(\mathbf{x}_1, \mathbf{x}_2)$ which are then propagated through two subnetworks - convolutional autoencoders (any other autoencoder could be considered; convolutional autoencoders suit better for

imaging data). The convolutional autoencoder-subnetworks have identically parameterized components - an encoder E and a decoder G. E performs the encoding to the space $\mathcal{Z}$ with a series of convolutional and downsampling operations while G performs the inverse mapping to the original space $\mathcal{X}$ through a series of deconvolutional and upsampling operations. The subnetworks output the reconstruction $\hat{\mathbf{x}}_i$ of the corresponding input $\mathbf{x}_i$. For a single input pair, the cost function associated with the framework is:

$$L(\mathbf{x_1}, \mathbf{x_2}; \Theta) = \sum_{t=1}^{2} ||\mathbf{x_t} - \hat{\mathbf{x}}_\mathbf{t}||_2^2 - \alpha \cdot cos(\mathbf{z_1}, \mathbf{z_2}) \tag{6.3}$$

The objective consists of two terms: the first one imposes the subnetworks to produce high quality reconstructions by minimizing the squared error between the subnetwork input and output; the second term imposes the similarity in the representation space by maximizing the cosine similarity of the *middle layer* feature vectors. α is the trade-off coefficient controlling the extent of similarity. Eventually, the representation $\mathbf{z}$ can be used for various tasks. In the scope of this work we will be considering the problem of outlier detection.
A regularized version of classical siamese networks was proposed in [Chen and Salman, 2011] for speech data. The network, composed of multilayer perceptrons, aims at learning similar representations for speech fragments of the same speaker while imposing the reconstruction error of the overall speech as a regularizing term in the loss function.

We reviewed several unsupervised deep architectures as potential representation learning mechanisms that can be coupled with an outlier detection algorithm. We proposed a novel configuration of siamese networks, regularized with the reconstruction error of the subnetworks - convolutional autoencoders. Such a regularized siamese network may be beneficial in the context of outlier detection. Precisely, the middle-layer representations may be coupled with an outlier detection algorithm which may model better the normality of the data points since they have been driven to be close when training the proposed framework. This aspect will be showcased in the next chapter.

Chapter 7

Epilepsy lesion detection on T1-weighted MR images

In this section we implement the CAD architecture proposed in chapter 5, with the regularized siamese architecture, developed in chapter 6, as the feature representation learning module. The architecture is applied to the automated detection of subtle epilepsy lesion detection on T1-weighted magnetic resonance images. To this end, we will consider the T1-w MR images of the data set presented in section 5.2. This amounts to 75 MR images of healthy controls and 21 images of patients with confirmed epilepsy lesions. The proposed framework is trained in an entirely unsupervised manner, using the images of the healthy controls. The evaluation is performed on the epileptogenic lesions found in patient scans. The following sections start by presenting the detailed pipeline of the CAD system. Next, we introduce our approaches of representation learning in the context of outlier detection. Eventually, we present the results obtained with the proposed framework on the task of epilepsy lesion detection.

7.1 Detailed CAD pipeline

The general pipeline of the CAD system is illustrated on fig. 7.1. It consists of two major steps - *patch-level representation learning* and *voxel-level outlier detection model learning*. In the first step, we propose to extract image patches of all the available volumes of the healthy controls and learn representations with deep learning architectures described in chapter 6. Once this step is performed, each voxel of a brain volume will be associated to a representation yielded by the deep network for the patch centered at the voxel. The second stage consists in building a oc-SVM model per voxel. Each voxel is associated with a classifier, hence the number of classifiers is equal to the number of voxels in the volume of

Algorithm 1: Algorithm to train oc-SVMs with the learnt representations per voxel.

Input : train set of registered images X,
deep model M,
number of voxels nb_v
Output: set of oc-SVMs $C = \{C_i\}_{i=1,nb_v}$

```
init C
for i ← 1 to nb_v do
    patches ← getPatchesCentereadAt(X, i)
    tr_matrix ← getRepresentations(patches, M)
    C[i] ← train_oc_SVM(tr_matrix)
end
return C
```

Algorithm 2: Algorithm to output the oc-SVM scores per voxel.

Input : test image I_p,
deep model M,
number of voxels nb_v,
set of oc-SVMs $C = \{C_i\}_{i=1,nb_v}$
Output: score map D_p

```
init D_p
for i ← 1 to nb_v do
    patch ← getPatchesCentereadAt(I_p, i)
    test_example ← getRepresentations(patch, M)
    D_p[i] ← output_score(C[i], test_example)
end
return D_p
```

interest (around 1.5 million voxels). 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 normal subjects centered at v_i.

For a new patient p, each voxel v_i is matched against the corresponding classifier C_i and is assigned the signed score output by the classifier. This yields a *distance map/score map* D_p for the given patient. The entire system is summarized in the pseudocode 1-2.

Implementation details The entire development of the CAD system was done in Python, using Theano and Keras libraries for the architecture training and feature extraction. The oc-SVM training was achieved by dividing all the voxels into distinct subsets, each subset being assigned to a separate thread. The oc-SVMs in each subset/thread were then trained sequentially. Testing was done in the same way. We used the oc-SVM implementation available in the Scikit-learn library [Pedregosa et al., 2011] which provides a Python wrapper to the LIBSVM library for Support Vector Machines [Chang and Lin, 2011]. Depending on the number of threads possible, the oc-SVM training takes between half an hour and 3 hours. Obtaining the cluster map for a given patient takes between 1 and 3 minutes.

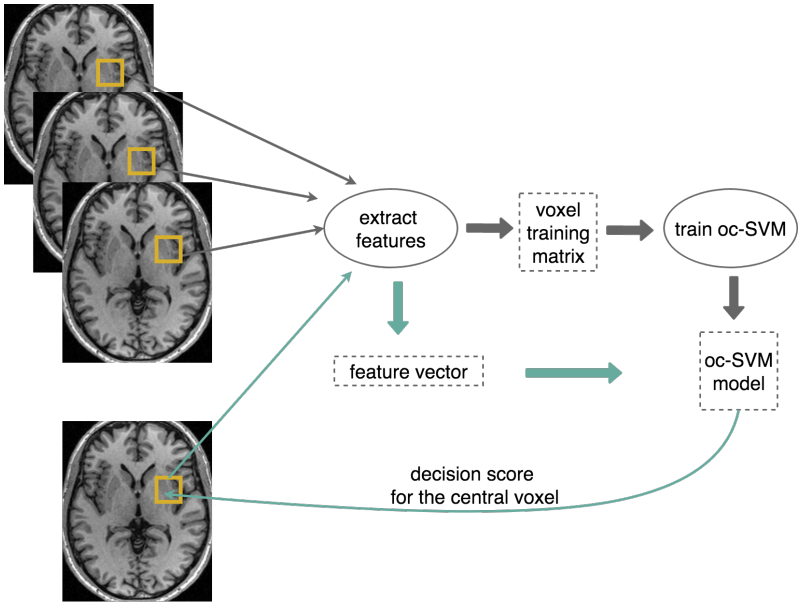


Figure 7.1: CAD general pipeline. The training is shown in a gray path, testing - in a green path.

7.2 Data description

The data set is composed of the healthy controls and patients introduced in chapter 5.2. In the scope of the experiments below, we will use the T1-weighted MRI sequences of the 75 healthy controls and 21 patients with confirmed epilepsy lesions. The detailed pre-processing steps are given in section 5.2.4. We recall the main aspects of the pre-processing routine. The T1-w MR images were normalized to the MNI space. Eventually, a voxel-level correspondence was established between the T1-w MRI acquisitions of different subjects. 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. Before feeding the volumes to the representation learning architectures, we removed top 1% intensities and scaled the images between 0 and 1 individually.

7.3 Experiments

In this section we describe and summarize the experiments done in the scope of the pipeline illustrated on fig. 7.1. More precisely, we have coupled the representations learnt with different networks, presented in chapter 6, with a oc-SVM model, on a per voxel basis. We have considered 4 such architectures

- stacked convolutional autoencoder (CAE)
- stacked denoising autoencoder (DAE)

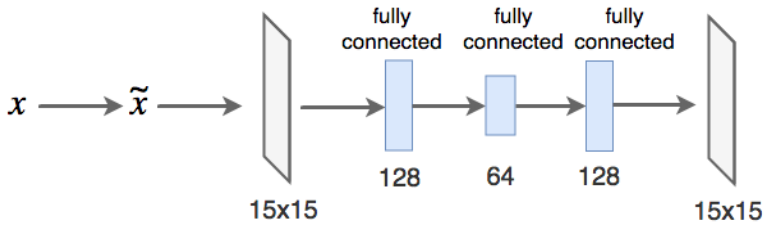


Figure 7.2: Stacked denoising autoencoder architecture (DAE).

- Wasserstein autoencoder (WAE)

- the proposed regularized siamese network (rSN)

In order to evaluate these architectures in comparable configurations, we chose the CAE, WAE and rSN architectures to have the same encoder E and decoder G structure. The dimension of the representations extracted from all the architectures is the same. Below we give the details of all 4 architectures.

7.3.1 Deep unsupervised architectures for representation learning

For all the architectures below, the training data set $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ consists of 15 x 15 patches extracted from all the volumes of healthy individuals with a fixed overlap of 8. This resulted in around $N = 3.5$ million patches. The choice of the input patch size is not arbitrary. It corresponds well to the sizes of subtle abnormalities linked to epilepsy. When considering larger patch sizes, the CAD system would detect abnormalities at a larger scale whereas epilepsy related abnormalities require features over local contexts. An illustration of this observation is included in appendix A.

Stacked denoising autoencoder architecture (DAE)

Fig. 7.2 shows the considered architecture. The square patches were flattened at input. The encoding path consists of 2 fully connected layers, later decoded by another fully connected layer. ReLU activation function was used everywhere except for the last one where sigmoid is used. The loss function to optimize is the mean squared error of the input patches and the corresponding 'reconstructions' output by the network

$$L_{DAE}(X;\Theta) = \frac{1}{N}\sum_{1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2$$

The masking noise was used on all the images with the masking probability P_{mask} (which was varied among the values 0.1, 0.3 and 0.5). The network is optimized with Adam optimization algorithm with learning rate=0.001 and momentum=0.5. Fig. 7.3 shows 10 randomly selected patches along with their reconstructions.

Stacked convolutional autoencoder architecture (CAE)

Fig. 7.4 shows the considered architecture. The encoding path consists of 3 hidden layers

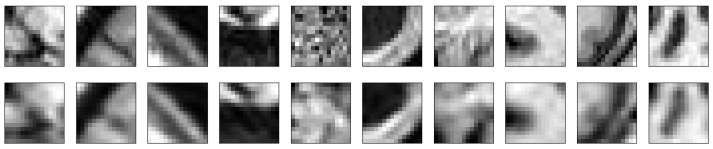


Figure 7.3: First row: 10 randomly selected patches. Second row: the corresponding patches reconstructed with the stacked denoising autoencoder (DAE).

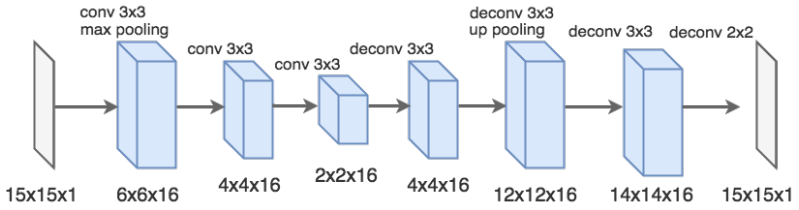


Figure 7.4: Stacked convolutional autoencoder architecture (CAE).

with kernel size 3x3 where only the first layer is followed by a max pooling layer. The decoding path is designed in a similar fashion. We used ReLU activation function in all the layers except for the last one where sigmoid is used. The loss function to optimize is the mean squared error of the input patches and the corresponding 'reconstructions' output by the network

$$L_{CAE}(X;\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2$$

The network is optimized with Adam optimization algorithm with learning rate=0.001 and momentum=0.5. Fig. 7.5 shows 10 randomly selected patches along with their reconstructions below. The model manages to preserve the main structures, with a slight blurring effect however.

Wasserstein autoencoder architecture

Fig. 7.6 shows the architecture for the encoder E, decoder (generator) G and discriminator D composing the proposed Wasserstein autoencoder shown on the right of the fig. 7.7. Similarly to the architectures above, the input patches are mapped to vectors $\mathbf{z} \in \mathcal{R}^{64}$ via the encoder and then mapped back to the original space through the generator. The loss function in this configuration is:

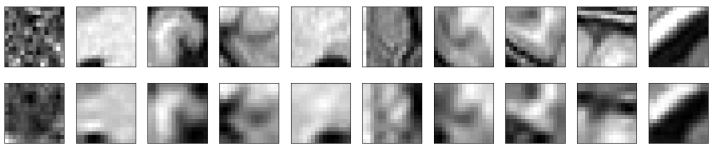


Figure 7.5: First row: 10 randomly selected patches. Second row: the corresponding patches reconstructed with the stacked convolutional autoencoder (CAE).

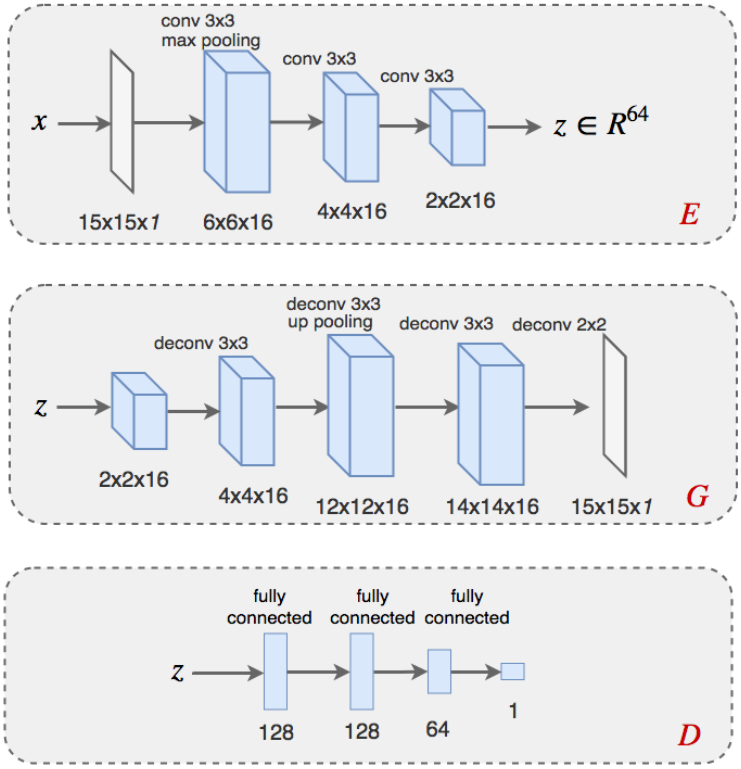


Figure 7.6: The architectures of the encoder E, generator G and discriminator D used in the Wasserstein autoencoder (WAE). Same E and G were used in the regularized siamese network (rSN).

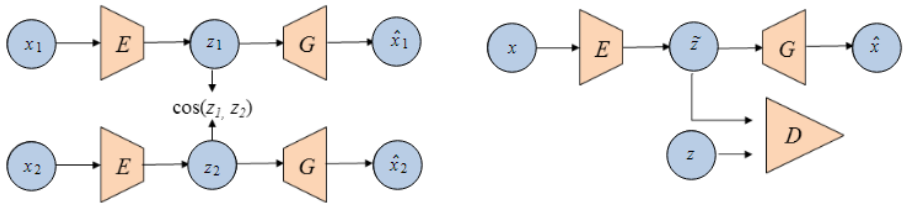


Figure 7.7: Global representation of the regularized siamese network (left) and Wasserstein autoencoder (right). The components E, G and D are shown on fig. 7.6.

$$L_{WAE}(X;\Theta_{WAE}) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i}||^2 + \beta \cdot D_{JS}(P_z, Q_z) \tag{7.1}$$

where D_{JS} is the Jensen-Shannon divergence, P_z is a multivariate Gaussian distribution and Q_z is explained in section 6.1.2.1. In this setting, $D_{JS}(P_z, Q_z)$ is estimated with the discriminator D that aims to distinguish the $\mathbf{z}$ produced with the encoder and the samples from the apriori distribution P_z. LeakyReLU was used as activation in the WAE discriminator with a scale of 0.02 for negative values. ReLU was used in the generator and the encoder, except for the last layer of G where sigmoid was applied. We varied the parameter β in the L_{WAE} expression 7.1 among the following values - 1,5,10 and 20.

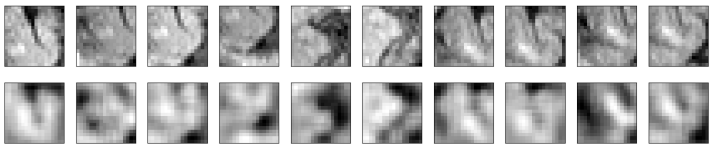


Figure 7.8: First row: 10 randomly selected patches. Second row: the corresponding patches reconstructed with the Wasserstein autoencoder (WAE).

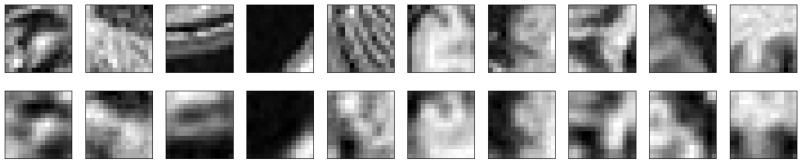


Figure 7.9: First row: 10 randomly selected patches. Second row: the corresponding patches reconstructed with the regularized siamese network (rSN).

Regularized siamese network architecture (rSN)

The proposed regularized siamese network (rSN) has the same encoder-decoder components as the CAE and WAE architectures described above. Fig. 7.7 illustrates a general scheme of the network while the details are shown on fig. 7.6. The input of the network consists of similar pairs of patches, defined as patches of different subjects centered around the same spatial voxel. The pairs were composed in the following way. First, patches were extracted from all the healthy subjects with a stride 8. Next, for each patch of a subject, a pair was composed by randomly selecting its similar patch among those belonging to the remaining subjects. The number of pairs is again around 3.5 million. The tradeoff coefficient α was varied during the training in the following way. It was set to 0 for 10 epochs, then grew linearly up to some pre-chosen value α_{max} and plateaued for another 5 epochs. In our experiments, α_{max} was varied among the values 0.25, 0.5, 0.75 and 1. The network is optimized with Adam optimization algorithm with learning rate=0.001 and momentum=0.5.

Alternative 3D architecture

In the scope of the proposed CAD we have also evaluated a limited number of architectures on 3D patches. Our intuition is that 3D patches may provide a richer context for the representation learning component in the CAD system and, hence, improve its performance. The encoder and decoder components considered for an alternative 3D architecture are illustrated on fig. 7.10. We have combined them into a regularized siamese network, as it is depicted on fig. 7.7. The structures of the encoder and the decoder follow those presented earlier for 2D patches. We considered 15 x 15 x 5 patches since most epilepsy lesions take up around 5 consecutive transverse slices. We therefore aimed at having a comparable view on the possible abnormalities.

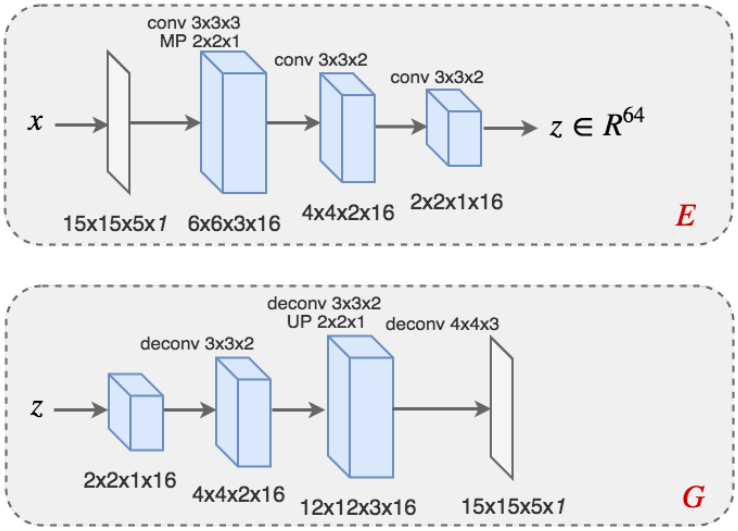


Figure 7.10: Alternative 3D encoder and decoder to be used in an experimental 3D rSN.

7.3.2 oc-SVM classifier design

Each voxel v_i is associated with a oc-SVM classifier C_i which is trained on the matrix $M_i = [\mathbf{z_{i1}}, ..., \mathbf{z_{in}}]$ where $\mathbf{z_{ij}}$ is the representation vector corresponding to the patch centered at v_i of subject j and n is the number of subjects.
Each classifier C_i was used with a RBF kernel defined as

$$K_{RBF}(\mathbf{z}_{ik}, \mathbf{z}_{ij}) = e^{-\gamma||\mathbf{z}_{ik}-\mathbf{z}_{ij}||^2}$$

where $\gamma = \frac{1}{\sigma^2}$ is the inverse of the kernel width σ.
For large values of γ, $K_{RBF}(\mathbf{z}_i, \mathbf{z}_j)$ gets close to 0 for any distinct i and j and therefore all the data points turn into support vectors. The opposite extreme case results in a small number of support vectors which leads to tighter boundaries. We adopted the heuristic described in [Caputo et al., 2002], proposing to choose γ within the range between the 10th and the 90th percentiles of the pairwise distances in the data set. For each oc-SVM individually, we chose to set γ to the median of the standardized euclidean pairwise distances of the corresponding matrix M_i. The parameter ν, the upper bound of the fraction of allowed outliers in the oc-SVM formulation 5.1, was set to 0.03. Varying this parameter had no effect on the performance of the CAD. Indeed, the fraction of outliers is actually controlled in the post-processing stage, described in section 7.3.3.
For each voxel v_i, the corresponding oc-SVM model C_i outputs the score for the voxel, i.e. the distance to the found optimal hyperplane, corresponding to

$$score(v_i) \leftarrow \mathbf{w}^* \cdot \phi(\mathbf{z}_i) - \rho^*$$

where $\mathbf{w}^*$ and ρ^* define the optimal hyperplane, as explained in section 5.1.3. Eventually, all voxel distance scores combined together yield the *distance map* D_p for the given patient p.

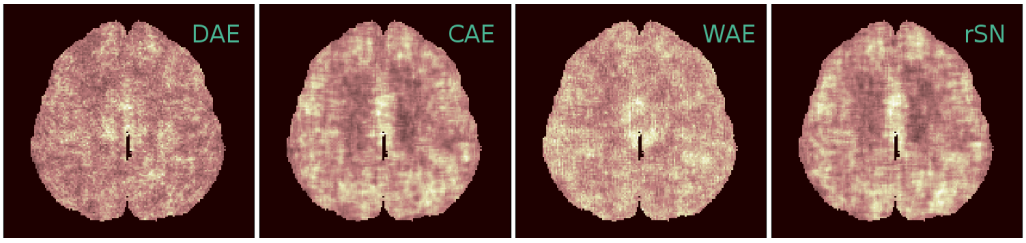


Figure 7.11: Normalizing maps N_S derived from the distance score maps of the healthy population, obtained with features learnt with denoising autoencoder (DAE), stacked convolutional autoencoder (CAE), Wasserstein autoencoder (WAE) and regularized siamese network (rSN), from left to right. The distance score maps of healthy subjects were computed with a 10-fold validation. Darker shades correspond to zones with low standard deviation.

7.3.3 Post-processing

For a given patient, the output of the previous step, the *distance map* D_p, is then post-processed to obtain the final detections. A 3-step post-processing is proposed as follows.

The first step consists in normalizing the distance maps to account for the intra-subject spatial variability. For that purpose, the distance maps of the control subjects are computed by performing a 10-fold evaluation of the controls in the training set. For each fold of normal subjects, the distance maps are obtained based on the oc-SVM models trained on the remaining subjects. These distance maps, estimated on the healthy subjects constituting the training data set X, are used to estimate the standard deviation of the *normal subjects' distance* distribution at voxel-level. In other words, a normalizing map N_S is computed where

$$N_S(v_i) \leftarrow std(\{D_s(v_i)\}_{s \in X})$$

where D_s is the distance score map for the healthy subject s, X is the training data set. Examples of such maps are shown on fig. 7.11. 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 deviation map N_S.
The final distance map F_p is then derived by averaging D_p and $\acute{D}_p$ i.e.

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

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 take into account 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 was then approximated by a non-parametric distribution using a kernel density estimator [Bowman and Azzalini, 1997]. The approximated patient distance score distribution is then

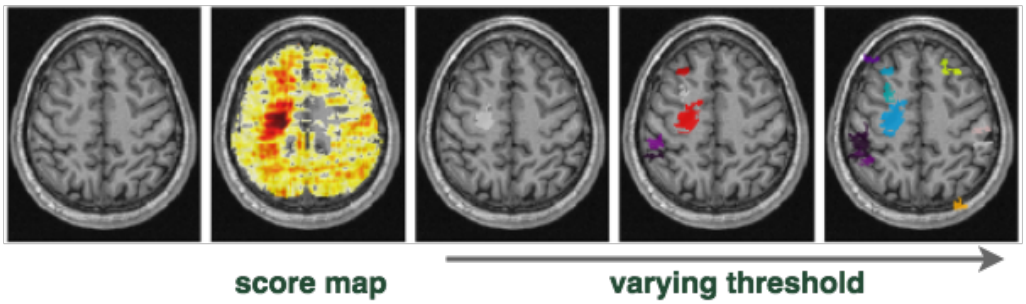


Figure 7.12: An example of post-processing on a patient. The first column shows the original slice centered at the lesion, the second column corresponds to the normalized distance map F_p and the last three columns are obtained by thresholding F_p at three different p-values and identifying the connected components as detections.

thresholded at some pre-chosen *p-value* and a 26-connectivity rule is applied to identify the connected components. These components are referred to as *clusters*. By varying the *p-value* the number of clusters can be controlled according to a clinician's needs. Fig. 7.12 shows an example of post-processing by varying the threshold with a p-value. We empirically set the *p-value* to the value that results in at most 15 clusters. We noticed that larger *p-values* resulting in more clusters typically produce a number of very large ones, as observed among the normal population whose output maps were obtained through the 10-fold validation done in the previous step. The voxel clusters smaller than a fixed size are discarded. In this study, the minimal cluster size was set to 82 voxels corresponding to the expected cluster size calculated with the SPM analysis of the T1 MRI data. This allows quick elimination of small and very negative clusters which usually represent isolated intensity peaks. The size of the majority of the detected clusters varies between 500 and 3000, this threshold therefore does not affect the performance in any significant way. The *clusters* are what we refer to as *detections* by the proposed method.

The third step consists in ranking the detected clusters to help the analysis of the detections. We use the following *ranking criterion* to assign a rank to a cluster c_i, inspired from [Ahmed et al., 2016]

$$rank(\mathbf{c_i}) \sim \omega * \frac{score(\mathbf{c_i})}{min_j score(\mathbf{c_j})} + (1-\omega) * \frac{size(\mathbf{c_i})}{max_j size(\mathbf{c_j})} \tag{7.2}$$

where $score(c_i)$ is the average of the voxel scores in the cluster c_i and $size(c_i)$ is the number of voxels in the cluster. Since the scores are thresholded at some *p-value*, $score(\mathbf{c_j})$ are, in turn, bounded by that value. In our experiments we rank the clusters on both cluster size and average score, therefore we set ω to 0.5. Using this ranking, we keep at most the top 10 detections and discard the rest. Eventually, keeping at most 10 clusters has a practical consideration from the medical perspective. Allowing more false positives may complicate the lesion screening by clinicians.

7.3.4 Evaluation protocol

In order to evaluate the described CAD system with different representation learning architectures, the produced final cluster maps are matched against the defined ground truth annotations. Each of the patients in our data set has only one lesion. A given cluster map is compared to the ground truth image and the overlap between the found clusters and the *ground truth cluster* is computed. When such an overlap exists for one or several detected clusters, we refer to it (them) as a *true positive detection*. The remaining clusters, falling outside the true lesion zone, are counted as *false positive detections*. Eventually, we report

1. the sensitivity as the proportion of the patients where there is at least one true positive detection by the system

2. the average number of false positive detections per patient.

We did not measure the number of detections in healthy patients since the framework seeks to find a wide range of anomalies and a certain number of anomalies (resulting from healthy anatomical variability of brains) would be found among healthy controls as well. It should not be indicative of CAD performance.
Since the post-processing described in the previous section produces cluster maps with ranked clusters, our main interest is also to evaluate how many true detections are found among top n clusters in all the patients.
From the clinical perspective, the number n of detected clusters would be rather limited to 10 since a larger number may limit the capacity of a clinician to focus on the true predictions. For this reason, we will evaluate the performance of the CAD system by varying the number of top n clusters from 3 to 10, hence tolerating at most 9 false positives per patient.

7.4 Results

7.4.1 Comparison of deep feature-based CADs

The first comparison involves evaluating the performance of the proposed CAD system using the representations learnt with various unsupervised deep architectures as described in section 7.3.1. Fig. 7.13 illustrates the performances of the architectures mentioned above in various configurations. More precisely, the figure depicts the sensitivity of the system as a function of the number of false positive detections. In other words, the performance is quantified as the proportion of patients whose lesions were detected among the top n clusters, where n varies across the horizontal axis. Ideally, the performance should behave as a straight line i.e. the maximum sensitivity should be achieved with the smallest number of false positives.

We varied the masking probability P_{mask} of the DAE among 3 values - 0.1, 0.3 and 0.5. The tradeoff coefficient β of WAE was set to 4 values - 1, 5, 10 and 20. We considered 4 values for the coefficient α in rSN - 0.25, 0.5, 0.75 and 1[1]. The best performance with DAE was achieved for $P_{mask} = 0.1$. For WAE, the different values of β resulted in rather similar sensitivities. Among the considered values of the α coefficient in the rSN model, 0.25 resulted in the best sensitivity. Eventually, the best configurations for DAE, WAE and rSN architectures were compared with each other and CAE on the bottom right plot on fig. 7.13. The latter clearly illustrates that both WAE and rSN features outperform CAE features which supports our initial intuition that the reconstruction loss alone does not fully seize the potential of unsupervised feature learning for outlier detection. Eventually, the rSN architecture with $\alpha = 0.25$ gives the best performance achieving 43% detection rate for 8-9 false positive clusters, followed by rSN with $\alpha = 0.5$ and WAE with $\beta = 1$ resulting in 38% sensitivity for 8-9 FPs. Larger values of the α coefficient probably impose too strong a similarity constraint among the normal points which smears out the contribution of the subnetwork reconstructions (therefore, ignoring the necessary variability) and degrade the overall performance. These results show that the CAD system with automatically learnt features, in particular, those obtained with WAE and rSN, allows a sensitivity by far superior to the human performance ($1MRI^+$ out 21 patients on T1-weighted MRI).

7.4.2 2D versus 3D representations

We compare the 2D and 3D patch-based approaches by evaluating the described 2D and 3D regularized siamese networks. The proposed 3D architecture is an early experimental one. We only seek to demonstrate the possible advantages of 3D representations over their 2D alternatives.

Fig. 7.14 illustrates the difference in the performances of 2D and 3D CAD systems for two values of the coefficient α - 0.25 and 0.5. As it can be seen, the performance is significantly improved with respect to the sensitivity. It can also be noted that the different α values seem to offer less performance variability in 3D than in 2D. Starting from 5 FPs, the two α values result in identical sensitivities. Eventually, the 3D architectures allow to detect around 48% of epilepsy lesions for 6-9 false positives.

7.4.3 Comparison with handcrafted features and GLM

We compared our approach of automated feature learning coupled with per voxel oc-SVM learning with two current approaches. The first is the approach described in [El Azami et al., 2016] which consists in associating each voxel with two clinically guided features and learning a oc-SVM per voxel. Two feature maps were thus computed for all subjects from the probabilistic tissue maps modeling

[1]Hereafter, rSN with $\alpha = t$ actually refers to $\alpha_{max} = t$. We use this shorthand for practical reasons.

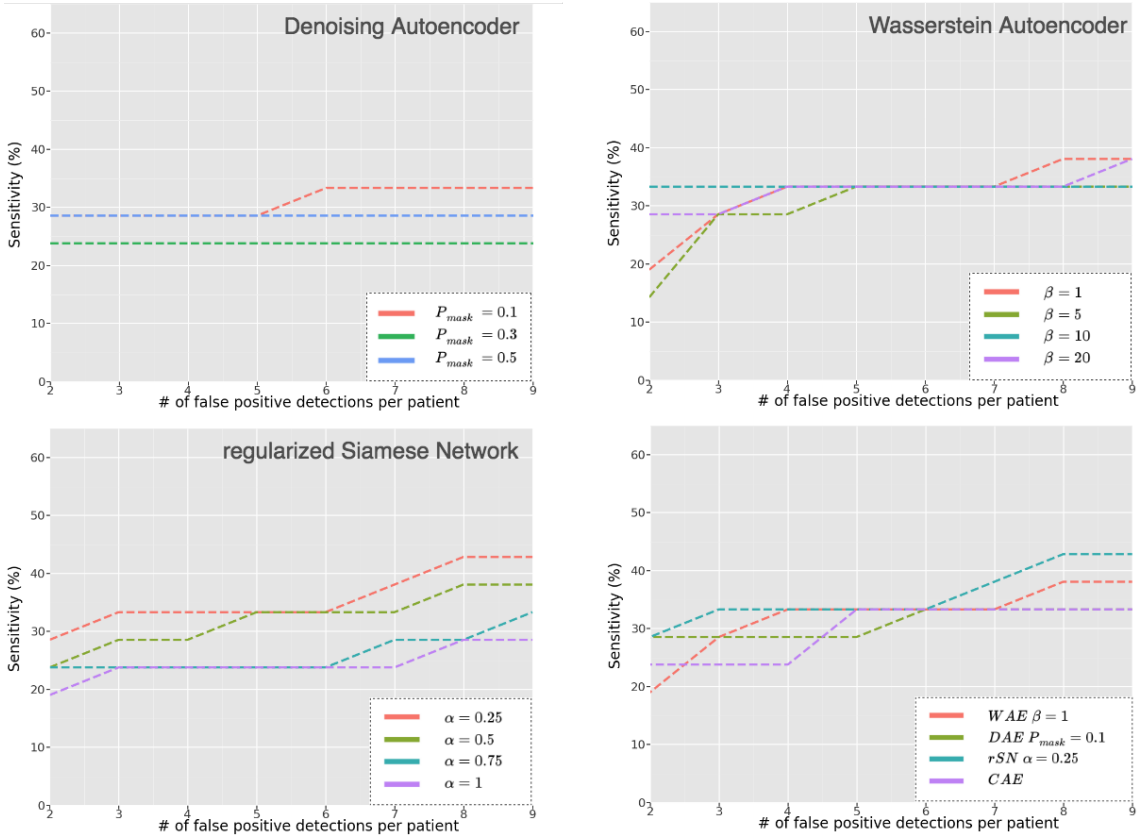


Figure 7.13: Comparative performance of the CAD system based on features learnt with 4 unsupervised architectures. Top left: denoising autoencoder (DAE) with 3 choices of masking probability; Top right: Wasserstein autoencoder (WAE) with 4 values of the tradeoff coefficient $\beta = 1, 5, 10, 20$; Bottom Left: regularized siamese network (rSN) with 4 values of the tradeoff coefficient $\alpha = 0.25, 0.5, 0.75, 1$; Bottom Right: the best configurations of the DAE, WAE, rSN and CAE (convolutional autoencoder).

1. the extension of the gray matter into the white matter (*extension map*)

2. the junction between the gray and white matters (*junction map*)

The extension map was obtained from the segmented gray matter image, smoothed with a 6 mm Gaussian kernel. To compute the junction map, the T1-weighted intensity corrected MR image was transformed into a binary image by selecting the voxels with a gray value between $mean_{GM} + \frac{1}{2}std_{GM}$ and $mean_{WM} - \frac{1}{2}std_{WM}$ where GM/WM refers to gray matter/white matter, *mean* and *SD* values correspond to the mean and standard deviation of the gray values in the respective tissue class, with $mean_{WM} > mean_{GM}$. A smoothing with a 6mm Gaussian kernel was then applied to the binary image. These features were shown to be discriminant for FCD and heterotopia [Huppertz et al., 2005, Wagner et al., 2011].

The oc-SVM classifiers were designed in the same way as descried in section 7.3.2. The same post processing routine was applied, with the exception of the distance map normalization step described in 7.3.3. This comparison should reveal the advantages and the

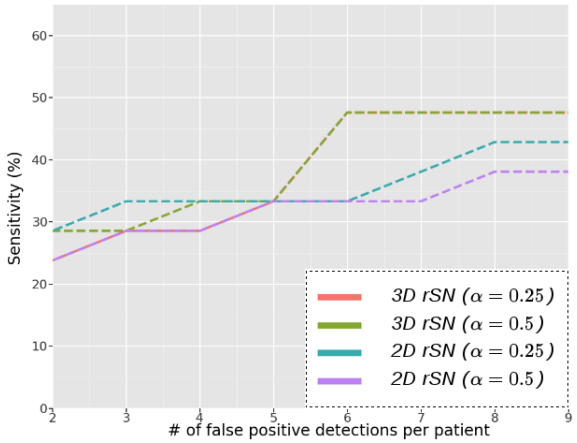
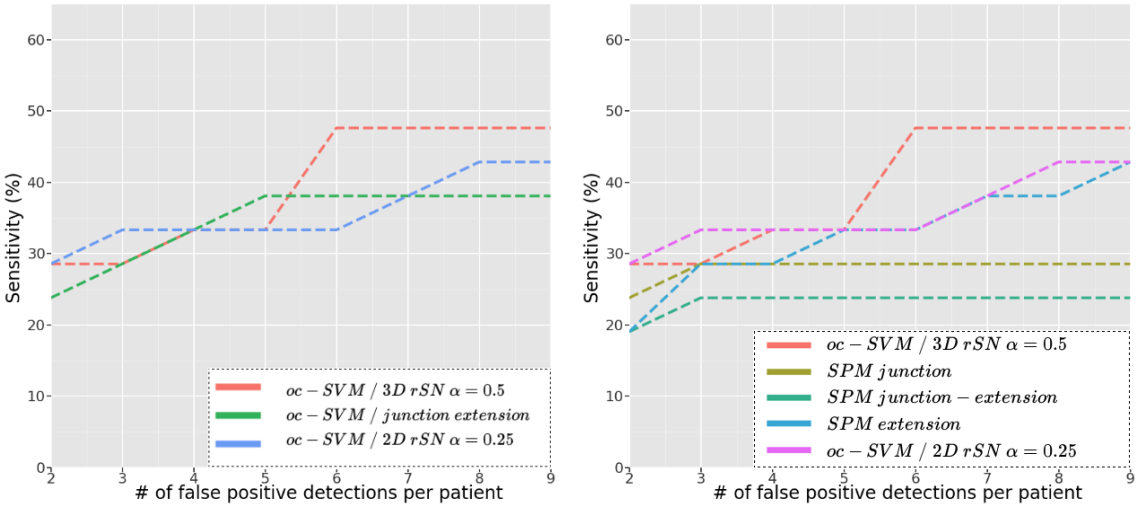


Figure 7.14: Comparative performance of the CAD system based on features learnt with 2D and 3D regularized siamese networks with different tradeoff coefficients $\alpha = 0.25$ and $\alpha = 0.5$. Starting from 5FPs, the curves for both 3D architectures overlap.

drawbacks of automatically learnt versus handcrafted features, coupled with per-voxel oc-SVMs.

As shown on fig. 7.15a, the CAD system based on handcrafted features reaches 38% of detection rate for 5-9 false positive detections. The sensitivity of the CAD as the number of FPs varies is rather steady; these handcrafted features, being chosen to capture common FCD characteristics on T1-w MRI, detect the relevant lesions early with less false positives as opposed to automatically learnt features that do not target a specific lesion type. The drawback of the handcrafted features can be deduced right away - when a lesion does not meet the typical characteristics, it is easily missed by the system. The comparative results show that the features learnt with 2D rSN are comparable to the handcrafted features, moreover, outperforming them when allowing 8-9 FPs. As for the 3D rSN features, they clearly outperform the handcrafted ones with around 48% sensitivity for 6 FPs versus only 38%.

The second comparison we performed was against the mass univariate statistical analysis implemented in the statistical parametric mapping software SPM. Within this method, a general linear model (GLM) is first fitted to each voxel based on the chosen factors of interest. As such factors, similarly to [El Azami et al., 2016], we considered the same junction and extension maps that were used to train the oc-SVM models. Later the estimated parameters of the fitted models are used to produce statistical t-value maps. Local neighborhoods of voxels, where the statistic is above a certain threshold, form clusters. The statistical significance of those clusters is assessed with the Gaussian random fields (GRF) theory which accounts for the correlation among the neighboring voxels. In our experiment, one-way ANOVA was performed based on four factors of interest: patient junction map, control junction maps, patient extension map and control extension maps. Two contrasts - [1,-1,0,0] and [0,0,1,-1] - were used to test the significant differences in the patient junction and extension maps compared to those of the controls. We also performed



(a) Comparative performance of the oc-SVM based CAD systems with rSN features versus handcrafted features (gray-white matter junction/extension).

(b) Comparative performance of the CAD system with rSN features versus SPM analysis on junction, extension and conjunction of junction and extension.

Figure 7.15: fROC curves of the CAD system. x-axis: number of false positives per patient, y-axis: sensitivity.

a conjunction analysis to test the global null hypothesis that the chosen factors - junction and extension maps - are consistently and jointly significant. In all three cases, the obtained t-score maps were thresholded at the value corresponding to p-value $= 0.001$ where higher t-scores indicate a higher level of anomalousness. These maps were post-processed in the same way as described in section 7.3.3 (with the exception of the normalization step which otherwise degraded the corresponding results).

Fig. 7.15b illustrates the comparison of our CAD system with rSN features with the SPM analysis performed in three settings

1. SPM analysis on gray-white matter junction
2. SPM analysis on gray matter extension into the white matter
3. SPM conjunction analysis on junction/extension

As it can be seen, the SPM analysis on the extension maps yields the best results among the three settings. Indeed, for 9 FPs SPM analysis on extension achieves around 43% sensitivity while the same analysis on junction and the conjunction analysis on junction/extension result in only 28% and 24% sensitivity, respectively. The rSN features coupled with oc-SVMs, however, outperform the statistical analysis approach, for both 2D and 3D contexts. Moreover, the difference is even more significant with 3D rSN features. These results show that the quantitative performance of the proposed CAD system is superior to the statistical approach with SPM.

Table 7.1 summarizes the performance of the best model with automatically learnt features with oc-SVM (in both 2D and 3D settings), handcrafted features with oc-SVM and

handcrafted features with GLM analysis. The table shows which patients' lesions were detected among the top 10 clusters and, hence, 9 FPs. The rank of each correct detection among the top 10 clusters is given inside parentheses in each cell.
As can be deduced from these results, GLM on the extension maps reaches 9 out of 21 detections while GLM on junction detects 6 out of 21 lesions. A drawback can be seen immediately in the joint analysis of these features which degrades the results to 5 detections out of 21. Indeed, it seems to us it is not trivial to extend the SPM analysis to consider multiple factors. On the other hand, some patients (I, J and T), while occasionally detected with some of the SPM analysis settings, are never detected with oc-SVM CAD.
The comparison between oc-SVM based CAD system with handcrafted features and deep representations reveals that for a few patients', the lesions are detected with the former and not with the latter and vice versa. So, patients A^-, G^- and N^- are detected with 2D rSN features but missed with the handcrafted features. Conversely, patients E^- and P^- are detected with the handcrafted features. When comparing the handcrafted features with 3D rSN, all the detections found with the former, except patient P^-, are found as well with the latter.

7.4.4 Qualitative results

Below we present the visual results obtained with the CAD system trained on the features learnt with the architectures presented in section 7.3.1. Fig. 7.16 shows the normalized score maps (centered at the slice of interest) output by the CAD systems based on the representations of different architectures. The columns correspond to DAE with $P_{mask} = 0.1$, CAE, WAE with $\beta = 1$ and rSN with $\alpha = 0.25$. Darker shades indicate more negative values which, in turn, means higher anomalousness. As it can be seen, the different architectures allow to identify the lesions to different extent. The most striking difference is noticed for patient G^- where the lesion contrast is the most visible with the rSN representations.
Fig. 7.17 illustrates the maximum intensity projections of the clusters found by the CAD system onto a transverse slice centered at the presumed lesion. Similarly, each column corresponds to a particular architecture. In each case, the illustrated cluster map shows the minimal number of FPs allowing to achieve the detection of the epilepsy lesion. In other words, the number of shown clusters for each patient corresponds to the rank of the true detection by the CAD. When a patient's lesion is not detected by the system, top 10 clusters are shown. It should be noted that the maximum intensity projections may result in overlapping clusters, therefore their number may be visually underestimated.
The figure shows the differences in the detection quality and the FPs accompanying the true detection. Patient G^- is detected with only 3 FPs with rSN while with many more for other architectures. The same is true for patient D^+, though with less variation between the number of FPs across different architectures. It should be noted that we employ the

Patient	**Lesion location**	**oc-SVM junction-extension**	**GLM on junction**	**GLM on extension**	**GLM on conjunction**	**oc-SVM 2D rSN** $\alpha = 0.25$	**oc-SVM 3D rSN** $\alpha = 0.5$
Patient A^-	Insula R	✗	✗	✗	✗	✓(8)	✗
Patient B^-	Temporal Lobe L	✓(1)	✗	✓(3)	✗	✓(1)	✓(1)
Patient C^-	Hippocampus R	✗	✓(3)	✓(4)	✗	✗	✗
Patient D^+	Superior frontal gyrus R	✓(1)	✓(4)	✓(1)	✓(1)	✓(2)	✓(3)
Patient E^-	Inferiolateral remainder of parietal lobe R	✓(5)	✗	✓(8)	✗	✗	✓(7)
Patient F^-	Hippocampus L, parahippocampus L	✗	✗	✗	✗	✗	✗
Patient G^-	Middle frontal gyrus L	✗	✗	✗	✓(4)	✓(4)	✓(5)
Patient H^-	Superior frontal gyrus R	✓(3)	✗	✗	✗	✓(1)	✓(1)
Patient I^-	Hippocampus L, parahippocampus L	✗	✗	✓(10)	✗	✗	✗
Patient J^-	Precentral gyrus R	✗	✗	✓(2)	✓(1)	✗	✗
Patient K^-	Superior temporal gyrus R	✗	✗	✗	✗	✗	✗
Patient L^-	Middle frontal gyrus R	✗	✗	✗	✗	✗	✗
Patient M^-	Anterior temporal lobe R	✗	✗	✗	✗	✗	✗
Patient N^-	Anterior temporal lobe R	✗	✓(1)	✓(4)	✓(1)	✓(9)	✓(3)
Patient O^-	Middle frontal gyrus L	✓(1)	✓(1)	✓(6)	✓(1)	✓(1)	✓(1)
Patient P^-	Hippocampus R	✓(6)	✗	✗	✗	✗	✗
Patient Q^-	Lateral remainder of occipital lobe L	✓(4)	✗	✓(1)	✗	✓(2)	✓(7)
Patient R^-	Orbital gyrus R	✗	✗	✗	✗	✗	✗
Patient S^-	Hippocampus R	✗	✓(3)	✗	✗	✗	✓(7)
Patient T^-	Posterior temporal lobe R	✗	✓(1)	✗	✗	✗	✗
Patient U^-	Posterior temporal lobe L	✓(1)	✗	✗	✗	✓(1)	✓(1)
Overall # of detections		**8**	**6**	**9**	**5**	**9**	**10**

Table 7.1: Comparative results of the CAD systems with oc-SVM on junction and extension, GLM on junction, extension and junction-extension conjunction and oc-SVM with 2D and 3D rSN features. ✓ denotes a detected lesion while ✗ denotes no true detection. The rank of each true detection is given inside parentheses.

term false negatives in a rather generous context. Some of the 'false' detections may be identified as anatomical abnormalities captured by the system and, hence, from the abnormality detection perspective are not actually $false$. Other false detections are likely to emerge from the registration step. For example, patient N^- seems to have several narrow clusters on the scalp borderline. Such patterns may be a result of the registration step that were not corrected through the normalization step during postprocessing.
Fig. 7.18 demonstrates the differences between the CAD system based on 2D rSN and 3D rSN representations. The two middle columns show the normalized output of each system while the two rightmost columns demonstrate the eventual detected clusters. Similarly as before, the difference between the two models can be seen in the number of FPs accompanying the true detection. So, for patient N^- this number is significantly reduced with 3D features. For patients D^+ and G^- the number of FPs is augmented by one in the 3D context as compared to the 2D context. Overall, it seems to us that the 3D context has certain advantages to offer, including the size and clarity of the true detections, particularly observed for patients D^+, N^- and G^-, and should be studied more extensively.

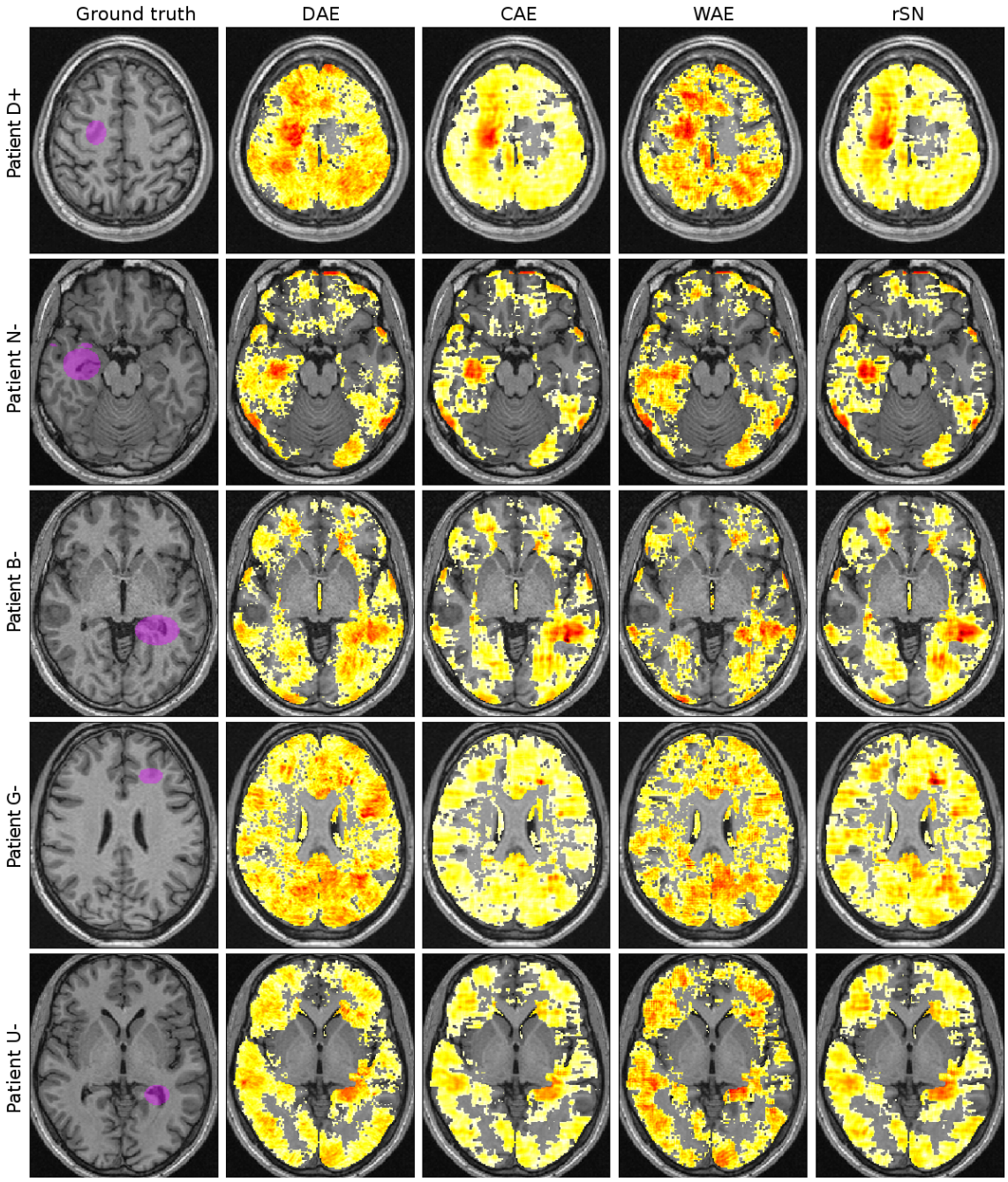


Figure 7.16: Normalized output maps obtained with features of various architectures. First column: The original image slice centered at the lesion, highlighted in a purple circle; Second column: DAE ($P_{mask} = 0.1$); Third column: CAE; Third column: WAE ($\beta = 1$); Fourth column; rSN ($\alpha = 0.25$). Darker shades on the output maps indicate higher suspicion of anomalousness. Notice the differences in the contrast between non-pathological areas and epilepsy lesions for various architectures. The most obvious difference is seen for patient G^- with high anomalousness found with rSN features and not with others.

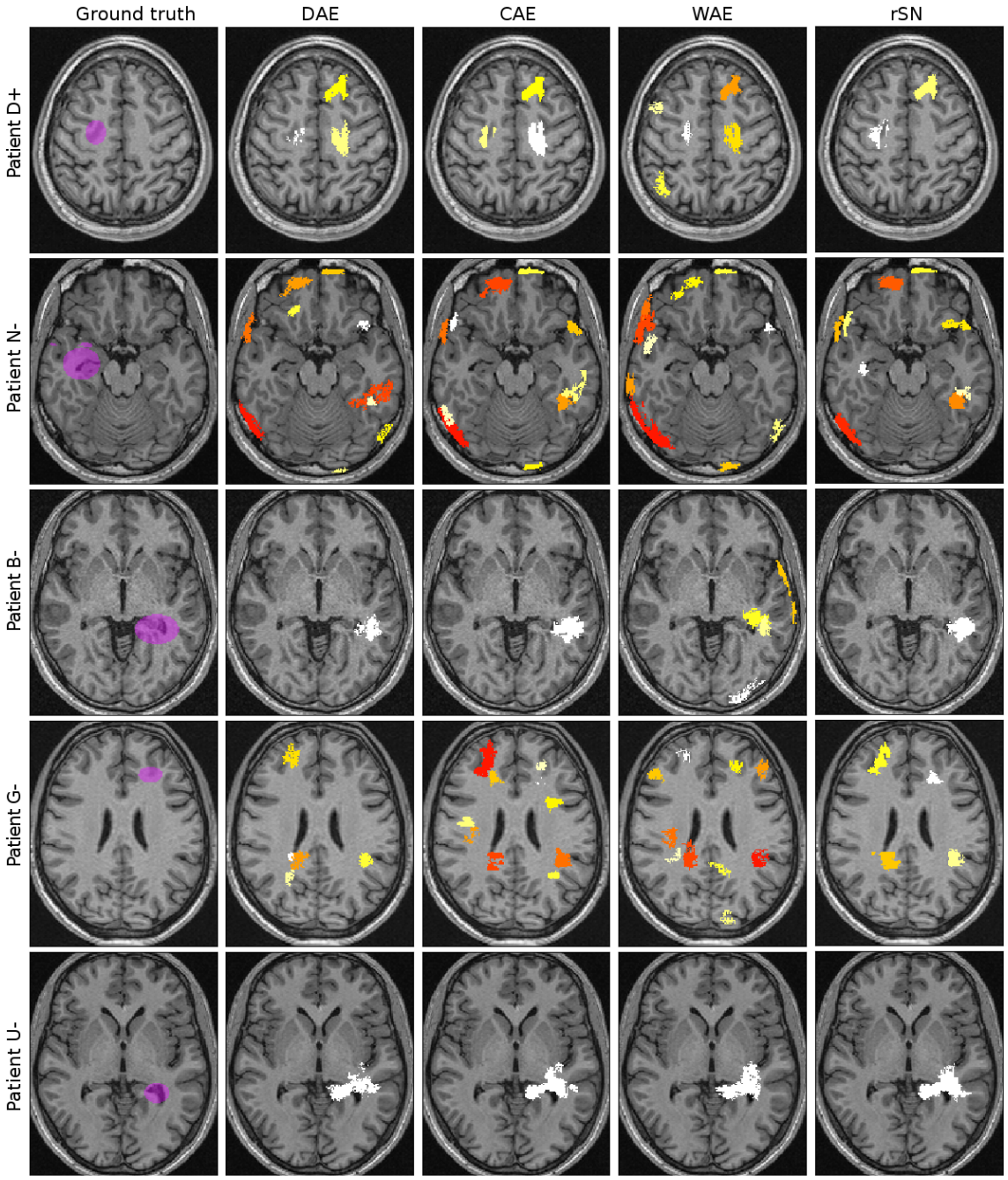


Figure 7.17: Maximum intensity projections of the cluster maps obtained with features of various architectures. The maps show the minimum number of false positive clusters allowing to detect the lesion, when it is detected, and top 10 clusters, when it is not. Some clusters' projections may overlap so visually their number might be underestimated. In reality, the clusters are distributed across the 3D brain volume, so the projections may sometimes seem to appear outside the scalp. First column: The original image slice centered at the lesion, highlighted in a purple circle; Second column: DAE ($P_{mask} = 0.1$); Third column: CAE; Fourth column: WAE ($\beta = 1$); Fifth column; rSN ($\alpha = 0.25$). Notice the variation in the number of clusters (hence, FPs) where the true detections are found.

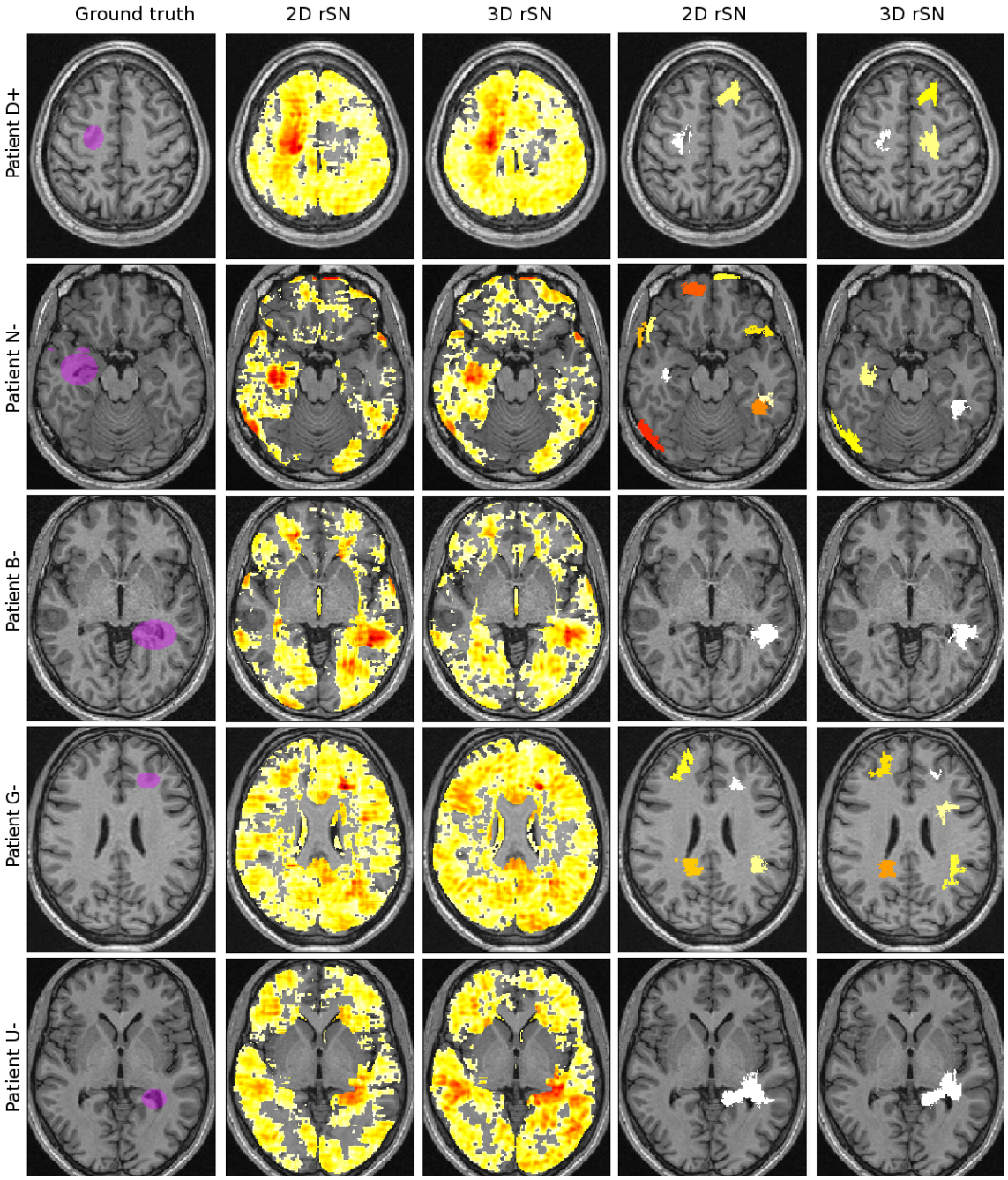


Figure 7.18: Visual comparison of normalized output maps and maximum intensity projections of the cluster maps obtained with features learnt with 2D/3D rSN architectures. The cluster maps show the minimum number of false positive clusters allowing to detect the lesion, when it is detected, and top 10 clusters, when it is not. Some clusters' projections may overlap so visually their number might be underestimated. In reality, the clusters are distributed across the 3D brain volume, so the projections may sometimes seem to appear outside the scalp. First column: The original image slice centered at the lesion, highlighted in a purple circle; Second column: 2D rSN ($\alpha = 0.25$) output maps; Third column: 3D rSN ($\alpha = 0.5$) output maps; Fourth column; 2D rSN ($\alpha = 0.25$) cluster maps; Fifth column; 3D rSN ($\alpha = 0.5$) cluster maps. Notice the variation in the number of clusters (hence, FPs) where the true detections are found.

7.5 Conclusion

In this chapter we presented a CAD framework for subtle anomaly detection on brain MR images. Such a system does not target any particular pathology but helps identify brain regions deviating from the population of healthy controls used for training. Such a CAD system can be easily integrated into the clinical routine and serve as a tool for preliminary screenings.
The proposed framework is based on learning feature representations with various unsupervised deep architectures and coupling them with a voxelwise outlier detection algorithm (oc-SVM). Our main contribution consists in

1. formulation of a regularized siamese network suitable for feature extraction for anomaly detection tasks

2. evaluation of the proposed CAD framework with various unsupervised deep architectures on a data set of epilepsy patients

3. comparison with clinically guided handcrafted features within the same framework as well as a statistical mass univariate analysis as implemented in the SPM software, for the task of epilepsy lesion detection.

The epilepsy lesion detection task on brain MRI is known to be extremely challenging as the lesions are subtle and do not have striking biomarkers. Moreover, in many cases patients' MRI are considered normal over routine visual examinations (*MRI-negative* patients). Our data set comprised mostly difficult cases (20 *MRI-negative* out 21 patients). We showed that with features learnt with deep architectures it is reasonable to achieve between 38 and 42 % of sensitivity for at most 9 FP detections. When evaluating different unsupervised architectures, we showed the advantageous performance of the features learnt with the proposed regularized siamese network. Moreover, when considering an experimental 3D rSN architecture, the CAD system achieved around 48% sensitivity for only 6 FPs.
When comparing our approach of data driven representations with the handcrafted features used in [El Azami et al., 2016], we established at least similar (and rather superior) performance, in particular, with 2D and 3D rSN features. This shows the potential of the considered approach which, despite not being tailored to any specific pathology, including epilepsy, can be successfully used to detect subtle pathological lesions. It should be noted, however, that in [El Azami et al., 2016], the reported sensitivity of 10/13 with 3-4 FPs corresponds to a slightly different configuration of the CAD system. First, the training and evaluation was done per data set, while in our system we pulled all the available subjects into a single one. We found this to be more coherent with real-life scenarios where the training subjects do not constitute a perfectly homogeneous set. Second, unlike our approach for tuning the RBF kernel parameter for each voxel individually, [El Azami et al.,

2016] chose a common parameter value for all the voxels, found through a cross validation procedure minimizing the number of FPs in healthy controls. These choices may have driven the reported performance in [El Azami et al., 2016] to be higher than the one revealed in our comparison.

When comparing the CAD system with the currently used SPM analysis we have shown that the proposed system is at least as good if not better when considering certain factors and their combination. An obvious bottleneck of this kind of approaches is the not trivial fashion the combination of multiple factors are treated in. So, extension alone in the SPM analysis results in 9 correct detections while the conjunction of extension and junction - in only 5. Our framework is more flexible towards the consideration of different characteristics.

One aspect of our CAD system that should be acknowledged is the interpretation of the so called false positives. When evaluating our system on the task of epilepsy lesion detection, we considered the detections not coinciding with the ground truth, false positives. Some of these false positives, however, correspond to anatomical abnormalities with respect to the healthy population used for training. From the abnormality detection perspective, those are not actual false positives. Some abnormalities detected by the proposed system could also be relevant to epilepsy, even though they do not overlap with the ground truth. The reason behind this is the fact that many epilepsy patients, especially those with normal MRIs, have other brain abnormalities that may be linked to the epilepsy seizures. It is not a trivial task to confirm these false detections as potentially epilepsy-relevant since their connection to epilepsy may not be validated clinically. Another set of false positives corresponds to imaging artifacts and registration discrepancies which end up detected as abnormalities. Such false positives are clearly a drawback and should be addressed in the future work.

The next chapter aims at extending the proposed CAD system to accommodate multiple imaging modalities. The motivation behind this stems from the fact that the visual confirmation of subtle abnormalities is rarely fully achieved with a single imaging modality. Typically, in the clinical practice neuroimaging data from different sources is extensively reviewed in order to capture the complementary information present in each of them. Thus, we aim at improving the diagnostic performance of the proposed CAD system through the integration of FLAIR MRI data.

III Multimodal outlier detection

Chapter 8

Modality fusion methods

For most phenomena and applications, it is not uncommon to have multiple sources of information providing different perspectives on the problem at hand. Those sources could include various measurements, experiments, multiple sensors, etc., depending on the application. In medical imaging, different types of images of the same subjects can be acquired, either from different modalities such as MRI and PET or from different protocols of the same modality such as various MRI sequences (FLAIR, diffusion, etc). Each such source results in a data *modality* or *view*. Since different modalities highlight different aspects of a subject, combining the information present in all of them has the advantage of giving a more comprehensive overview of the problem. This notion has built the foundation of the so called *multiview learning*, a category of algorithms that seek to learn a model per view and jointly optimize them in order to boost the generalization performance.
In a clinical routine, analyzing the medical images obtained with a number of distinct techniques is a standard practice as it gives a fuller picture on the state of the patient. It is therefore only natural for machine learning based automated diagnosis systems to seek to integrate multimodality data into a single framework.
Combining multimodality data involves adopting a multimodal data fusion strategy. The main aspects to consider for data fusion are

1. when in the pipeline the fusion should be introduced

2. the optimal method to perform the fusion for the task at hand

8.1 Fusion level

The current methods of multimodal data fusion fall into three major categories illustrated on fig. 8.1.

Early fusion consists in merging the modalities at the earliest level so that the input to the learning algorithm consists of their combined representations. The input per modality may be the raw imaging data or features relevant for each modality. The most straightforward option is to concatenate the input from different modalities. The advantage of the early fusion strategy is that the learning algorithm may leverage the correlation between the modalities. This is a natural approach especially when the multimodal inputs are homogeneous. Moreover, in an early fusion strategy, only one learning model is trained as opposed to training individual models per modality. However, when the modalities are significantly heterogeneous, a single learning model may not be able to capture the most relevant information. Moreover, the nature of the extracted features may be quite different (as is the case for textual, speech and visual features) and combining them in a unified representation may be challenging [Poria et al., 2016].
Early fusion has been used in various medical applications. [Niaf et al., 2012] combined various features from T2-weighted, diffusion-weighted and dynamic contrast-enhanced MRI modalities to discriminate prostate cancer from healthy tissue. [Kabir et al., 2007] combined gray level intensity values of different MRI sequences before feeding them into a Markov random field for ischemic stroke lesion segmentation. Recent state-of-the-art deep architectures for many pathologies are designed to perform a sort of early fusion by combining raw image modality data as channels. So, T1-weighted, contrast enhanced T1c, T2-weighted and FLAIR MR images were joint as channels in the architectures proposed in [Havaei et al., 2017, Kamnitsas et al., 2016] for brain tumor segmentation. The same strategy was implemented, among others, in [Brosch et al., 2016] for MS lesion segmentation.

Late fusion, also referred to as decision fusion, methods aim at combining the decisions output by the models trained individually for each modality. In this case the individual potential of each modality is exploited maximally, however, at the cost of training multiple learning models, one per modality. Combining the decisions in a meaningful way is another aspect to consider, depending on the type of the decision output by the chosen models (probabilistic score, numerical score, etc).
Late fusion was applied to combine the decisions of audio and video speech recognition models through adaptive weighting in [Lee and Park, 2008] or using a genetic algorithm optimisation technique in [Rajavel and Sathidevi, 2015]. In medical imaging, late fusion has been frequently performed in segmentation [Wang et al., 2013, Sabuncu et al., 2010, Asman and Landman, 2013]. [Isgum et al., 2009] applied late fusion on multi-atlas cardiac

segmentation while [Heckemann et al., 2006] performed late fusion on brain MR image segmentations obtained with label propagation.

Intermediate fusion methods, considered a separate fusion level in [Noble et al., 2004], are an intermediate strategy between early and late fusion. In this case, the modalities are combined after some transformation (e.g. kernel computation) has been applied to the individual modality data and the decision is made based on the computed combination. A representative example of intermediate fusion is multiple kernel learning (MKL) [Bach et al., 2004, Sonnenburg et al., 2006a, Sonnenburg et al., 2006b]. This category of methods is often acknowledged in the literature on multiview learning [Sun, 2013, Xu et al., 2013, Zhao et al., 2017]. The advantage of this fusion level is that only a single learning model is trained while the difficulty of combining heterogeneous inputs is alleviated with the kernel computation which can yield a common representation for all the modalities. Intermediate fusion has been applied in many medical contexts. [Zhang et al., 2011] proposed a framework where multiple kernels are associated with three modalities, namely MRI (measuring brain atrophy), PET (measuring hypometabolism) and CSF (quantifying specific proteins), while an SVM-like formulation combines them into a single model. The method was used to classify healthy subjects versus AD (Alzheimer's disease) patients and healthy subjects versus MCI (Mild Cognitive Impairment) patients. An MKL SVM on image-based markers was proposed again in [Hinrichs et al., 2011] for the same medical application. Another approach for the same task was proposed in [Suk and Shen, 2013] that applied the MKL paradigm coupled with representations learnt with stacked autoencoders. [Suk et al., 2014] proposed a deep architecture consisting of two subnetworks for MRI and PET data, joined only at the topmost layers to discriminate AD patients.

[Pavlidis et al., 2002] compared early fusion (concatenation of the vectors), intermediate fusion (kernels are computed separately and added later) and late fusion (individual SVMs are trained with their scores added later) for the problem of gene classification from a heterogeneous data set consisting of DNA microarray expression measurements and phylogenetic profiles from whole-genome sequence. Intermediate fusion showed the best results, probably since it provides a reasonable tradeoff between the assumptions made in early and late fusion strategies. Such comparisons, of course, depend on the application and the data.

8.2 Fusion methods

Depending on the fusion level, different combination strategies may apply. For early fusion the simplest approach is the vector concatenation or joining the raw images as channels (for neural networks mostly). Eventually, linear combination of multiple inputs is possible.

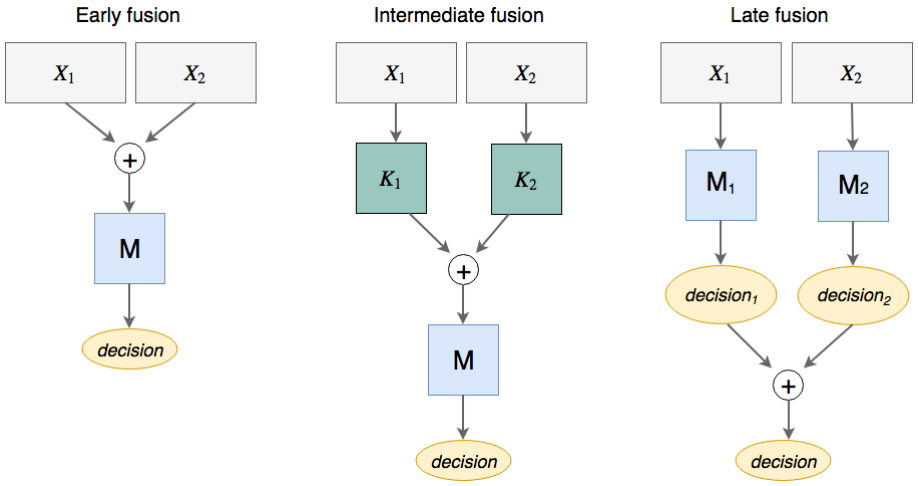


Figure 8.1: Comparative visualization of fusion levels. X_1 and X_2 are the inputs of the two modalities, K denotes a kernel, M denotes a learning model, *decision* is the output of M. $\oplus$ denotes the combination method.

Late fusion usually combines the decisions provided by the learning models built per modality. Common combination methods are majority voting, min/max operations, linear combination and/or logical operations. Another option is to train an additional model on the decisions of the trained models for each modality, having as output the final decision. This, however, involves training a supplementary model which comes with additional challenges. Multiple kernel learning is the most representative intermediate fusion strategy which will be explained in the next section. A more comprehensive review can be found in [Gönen and Alpaydın, 2011].

8.3 Multiple kernel learning for intermediate data fusion

In section 5.1.3 we presented the formulation of the oc-SVM algorithm. The primal formulation is the following problem

$$\begin{aligned} \min_{\mathbf{w},\rho,\xi_i} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \mathbf{w}\cdot\phi(\mathbf{x_i}) \geq \rho - \xi_i \qquad i \in [1,n] \\ & \xi_i \geq 0 \qquad i \in [1,n] \end{aligned} \tag{8.1}$$

where n is the number of training examples, $\mathbf{x_i}$ is the i-th example in the training data set X, ξ_i-s are slack variables relaxing the inequality constraints as to account for the non-separable classes, $\mathbf{w}$ and ρ define the separating hyperplane, $\nu \in (0,1)$ is a parameter that sets a boundary to the fraction of allowed outliers. ϕ is the feature map $\mathcal{X} \to \mathcal{F}$ from the original space to a dot product space $\mathcal{F}$, corresponding to some kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

In this formulation, as is the case for the binary SVM, a single kernel is used to project the data points into a new space. The choice of the kernel and its parameters requires a

tuning of its own and is one of the challenges of the algorithm.

In multiple kernel learning (MKL), the single kernel is replaced with a combination of several kernels, each parameterized accordingly [Bach et al., 2004, Sonnenburg et al., 2006a, Sonnenburg et al., 2006b]. A combined kernel can be written as:

$$K_{MKL}(\mathbf{x}_i, \mathbf{x}_j) = f_\theta(\{K_m(\mathbf{x}_i, \mathbf{x}_j)\}_{m=1}^{M}|\theta)$$

where $\mathbf{x}_i$ is the i-th observation, K_m is the m-th kernel, M is the number of modalities. f_θ is the kernel combination function parameterized with θ. θ may refer to the parameters injected into the individual kernel functions, optimized when solving the problem. Most common scenario, however, is to predefine the kernels per modality by setting their parameters while θ refers solely to their combination. The most common choice for f_θ is the linear combination

$$K_{MKL}(\mathbf{x}_i, \mathbf{x}_j) = f_\theta(\{K_m(\mathbf{x}_i, \mathbf{x}_j)\}_{m=1}^{M}|\theta) = \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i, \mathbf{x}_j)$$

This setting of multiple kernel learning can be injected into several kernel-based methods such as Kernel Fisher discriminant analysis, regularized kernel discriminant analysis and kernel ridge regression. The MKL formulation for one-class SVM can be written as

$$\begin{aligned}
\min_{\mathbf{w},\rho,\xi_i,d} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i & \\
\text{subject to} \quad & \mathbf{w}\cdot\sum_{m=1}^{M} d_m\phi_m(\mathbf{x_i}) \geq \rho - \xi_i & i \in [1,n] \\
& \xi_i \geq 0 & i \in [1,n] \\
& d_m \geq 0 & m \in [1,M] \\
& \sum_{m=1}^{M} d_m = 1 &
\end{aligned} \tag{8.2}$$

where d_m is the weighing coefficient of the m-th kernel, M is the total number of kernels, n is the number of observations. The Lagrangian of the problem is:

$$\begin{aligned}
L(\mathbf{w},\xi,\rho,d,\alpha,\beta) = & \quad \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\
& - \sum_{i=1}^{n}\alpha_i(\mathbf{w}\cdot\sum_{m=1}^{M} d_m\phi_m(\mathbf{x_i}) - \rho + \xi_i) \\
& - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{m=1}^{M}\gamma_m d_m - \mu(\sum_{m=1}^{M} d_m - 1)
\end{aligned} \tag{8.3}$$

Setting the derivatives of the Lagrangian with respect to the primal variables $\mathbf{w}$, ξ, ρ and d to 0 yields the following:

$$\begin{aligned}
\nabla_{\mathbf{w}}\mathcal{L} = 0 \quad &\Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i \sum_{m=1}^{M} d_m \phi_m(\mathbf{x_i}) \\
\nabla_{\xi}\mathcal{L} = 0 \quad &\Rightarrow \quad \alpha_i = \frac{1}{\nu n} - \beta_i \\
\frac{\partial \mathcal{L}}{\partial \rho} = 0 \quad &\Rightarrow \quad \sum_{i=1}^{n} \alpha_i = 1 \\
\nabla_{d_m}\mathcal{L} = 0 \quad &\Rightarrow \quad \sum_{i=1}^{n} \alpha_i (\mathbf{w} \cdot \sum_{m=1}^{M} \phi_m(\mathbf{x_i})) - \gamma_m - \mu = 0
\end{aligned} \tag{8.4}$$

Introducing 8.4 into the Lagrangian 8.3 the dual formulation becomes

$$\begin{aligned}
\max_{\mu,d,\alpha} \quad & \mu \\
\text{subject to} \quad & \sum_{i=1}^{n} \alpha_i = 1 \\
& 0 \leq \alpha_i \leq \frac{1}{\nu n} \qquad\qquad i \in [1, n] \\
& \sum_{m=1}^{M} d_m \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j K_m(\mathbf{x_i}, \mathbf{x_j}) \geq \mu \\
& d_m \geq 0 \qquad\qquad m \in [1, M] \\
& \sum_{m=1}^{M} d_m = 1
\end{aligned} \tag{8.5}$$

[Rakotomamonjy et al., 2008] proposed an algorithm (SimpleMKL) to solve the MKL problem for SVMs. More precisely, the algorithm solves problems that can be put in the following form

$$\begin{aligned}
\min_{d} \quad & J(d) \\
\text{subject to} \quad & d_m \geq 0 \qquad m \in [1, M] \\
& \sum_{m=1}^{M} d_m = 1
\end{aligned}$$

where the form $J(d)$ takes is rather flexible and therefore, can accommodate various SVM formulations. For oc-SVM

$$J(d) = \begin{aligned}
\min_{\mathbf{w},\rho,\xi_i} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i \\
\text{subject to} \quad & \mathbf{w} \cdot \sum_{m=1}^{M} d_m \phi_m(\mathbf{x_i}) \geq \rho - \xi_i \quad i \in [1, n] \\
& \xi_i \geq 0 \qquad\qquad i \in [1, n]
\end{aligned} \tag{8.6}$$

[Loosli and Aboubacar, 2017] proposed an extension to the SimpleMKL paradigm (slim-SimpleMKL) for oc-SVM which aims at building tight boundaries around the normal class by controlling the number of support vectors. This is achieved through the following formulation:

$$\begin{aligned} \min_{d} \quad & J(d) - \lambda \, card(\alpha) \\ \text{subject to} \quad & d_m \geq 0 \qquad\qquad m \in [1, M] \\ & \sum_{m=1}^{M} d_m = 1 \end{aligned} \tag{8.7}$$

where $card(\alpha)$ is the number of support vectors, λ is a tradeoff parameter and $J(d)$ is as in 8.6. By controlling the number of support vectors, tight normality bounds can be achieved which in turn can lead to an improved performance of the outlier detection. Note that $\lambda = 0$ amounts to the original formulation of SimpleMKL.
Multiple kernel learning was initially proposed to achieve better generalization performance by considering multiple kernels instead of restricting the kernel space to a single one. Ever since, however, MKL methods have been widely considered in multiview learning problems. The reason behind this is that the different kernels in MKL can be interpreted as different views on the data and combining them, therefore, is analogous to multiview learning.

Formally, MKL may be used to capture different similarity notions on the same observations or to combine the observations of different modalities, each encoded with a separate kernel. In the first scenario, all the kernels are computed on the same data set at hand, as described so far. In the second scenario, different kernels are associated with different modalities of the given set of observations. This means that the combination of kernels is computed with

$$K_{MKL}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i^m, \mathbf{x}_j^m)$$

where $\mathbf{x}_i^m$ is the i-th observation of m-th modality, K_m is the kernel corresponding to the m-th modality. Each modality may be associated with multiple kernels, without any major change in the formulations above.

The application we are concerned with is the optimal fusion of neuroimaging data modalities in the context of outlier detection. We are interested in comparing two fusion strategies. The first is the early combination of the imaging modalities as channels in a deep architectures, in order to learn common representations, later to be coupled with a oc-SVM model per voxel. We will then consider an intermediate fusion strategy, in particular, applying MKL oc-SVM on the features learnt with networks trained on each imaging modality separately. More precisely, the kernels will be computed individually on the features corresponding to the given modalities. In this case, an optimal combination of such kernels is sought in order to provide the best separation of the data points from the origin.

8.4 Multiview learning with incomplete data

So far, we have considered the scenario where the data of all modalities are available. In many applications, however, this is not the case. Some observations may not have the information from all of the views, but from some of them. A simple approach would be to train learning models on the examples with no missing entries. This, however, leaves out the available useful information. Eventually, this limits the number of training examples to consider. Classical strategies of missing data imputation, based on the estimation of missing entries from the observed ones, could be applied such as k-nearest neighbor, expectation-maximization, low-rank matrix completion [Hastie et al., 1999, Troyanskaya et al., 2001, Schneider, 2001, Candès and Recht, 2009]. However, such methods are better suited for problems where values are missing at random. In multiview learning on medical imaging, the information is missing in blocks (entire modalities are missing).

In the scope of multiview learning, several methods have been proposed to account for the missing data in the multimodal learning framework. One recent approach is based on multi-task learning where several learning models are trained simultaneously within a single framework. [Thung et al., 2017] leveraged the multi-task learning paradigm by dividing the given data set of incomplete observations into subsets of complete ones and associating a classification task to each subset. The approach is implemented through a deep learning architecture that is composed of one subnetwork per subset, each subnetwork having its own final classification layer, corresponding to multi-stage Alzheimer's disease diagnosis. There are, however, common intermediate layers, shared between the subnetworks, which means that the available data is fully exploited in all the classification tasks simultaneously. A similar multi-task learning framework was proposed previously in [Yuan et al., 2012] where, additionally, a constraint is imposed that results in all the tasks using a particular modality to select a common subset of the modality features.

A multiple kernel learning based approach, dealing with incomplete data, was proposed in [Zhu et al., 2017c]. The method relies on a multiple kernel learning problem formulation, enhanced with two additional terms. The first term is based on maximum mean discrepancy criterion, forcing the different data modalities to have a similar distribution in a common space, to which a mapping is performed via a kernel function. The second term involves a consistency criterion, assuring that the different modalities of the same subject are similar in the kernel space. The approach allows to leverage the full set of observations within each modality. The clinical application, again, is the Alzheimer's disease diagnosis.

Another approach consists in explicit generation (synthesis) of one modality given the examples of another one. The synthesized examples are then used in a hybrid data set combining real and generated examples and any further learning model, depending on the task, can be applied. Recent cross-modality synthesis methods, chiefly based on deep learning, will be further reviewed in section 10.2.

We presented an overview of the existing methods for multimodality data fusion, mentioning the problem of incomplete data as well. We are interested in applying the discussed options, more precisely, the early fusion and the intermediate fusion with MKL paradigm, in the context of anomaly detection. We will, therefore, evaluate the chosen strategies on the task of epilepsy lesion detection on T1-weighted and FLAIR MRI, in the scope of the proposed CAD system. As a next step, we will make an exploratory effort to integrate the PET imaging as well. In this case, the multimodal learning will take place on incomplete data. To overcome this problem, we will present our strategy of PET image synthesis from the corresponding MRI acquisitions.

Chapter 9

Epilepsy lesion detection on T1-w/FLAIR MR images

In chapter 4 we formulated the problem of subtle lesion detection on brain imaging as a per voxel outlier detection problem. In chapter 6 we presented our strategy of representation learning using various unsupervised neural architectures, namely denoising, convolutional and Wasserstein autoencoders and our own variation of siamese networks. Further, in chapter 7 we exploited these architectures as feature extraction mechanisms, coupled with per voxel oc-SVM learning, in the task of epilepsy lesion detection on T1-weighted MRI. The best performance of the proposed CAD system for epilepsy detection was achieved with regularized siamese networks as the representation learning component, reaching 42% sensitivity for 8-9 false positive detections.
In this chapter we explore strategies for multimodal outlier detection in order to integrate FLAIR MR images into the proposed framework. As explained in chapter 8, two fusion strategies will be considered i.e. early fusion and intermediate fusion. The early fusion strategy consists in proposing relevant unsupervised architectures that combine the given imaging modalities as input channels, with the rest of the pipeline remaining as described in chapter 5. The intermediate fusion consists in training individual networks for each modality and combining the learnt representations in a MKL paradigm. We will next evaluate the performance obtained with each strategy and eventually compare them.

9.1 Data description

The data set is composed of the same healthy controls and patients introduced in chapter 5.2. In the scope of the experiments below, we will use the T1-weighted and FLAIR MRI sequences of the 75 healthy controls and 21 patients with confirmed epilepsy lesions. The

T1-w MR images were normalized to the MNI space. The FLAIR sequences were first rigidly co-registered with the corresponding T1-w volumes and further normalized to the MNI space as well. Eventually, a voxel-level correspondence was established between the T1-w and FLAIR acquisitions of the same subject in addition to the correspondence of the acquisitions of different subjects. As in chapter 7, 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. Before feeding the volumes to the representation learning architectures, we removed top 1% intensities and scaled the images between 0 and 1 individually.

9.2 Experiments

In the experiments below, we extend the original framework proposed in chapter 5, combining features learnt with various unsupervised architectures with voxelwise one-class SVM (oc-SVM), to the multimodal setting. In the first set of experiments, we consider a number of architectures allowing the integration of multiple imaging modalities as input channels and compare their performance on the task of epilepsy detection on T1-w and FLAIR MRI. In the second part we pick the best architectures for each modality individually and couple the learnt features via the multiple kernel learning paradigm.

9.2.1 Early fusion with multichannel architectures

As an early fusion strategy, we first consider the architectures presented in chapter 7.3.1 for their ability to accommodate multiple modalities as input channels. In particular, we focused on stacked convolutional autoencoders, Wasserstein autoencoders and regularized siamese network, adapted to the multichannel setting. Similarly as in chapter 7.3.1, for all the architectures, the training data set $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ consists of 15x15 patches extracted from all the volumes of the healthy individuals with a fixed overlap of 8 which resulted in around $N = 3.5$ million patches.
The three considered models are shown on fig. 9.1b with the architectures of their components shown on fig. 9.1a. The encoder and decoder components have identical structures in all three architectures. In all cases the representations extracted with the architectures are 64 dimensional vectors when flattened.

Stacked convolutional autoencoder was trained with a batch size of 128 using the mean square error as loss function. ReLU activation was applied in every layer, except the last layer of the generator where the sigmoid was applied. Adam optimizer with a learning rate 0.001 was used for 25 epochs.

Wasserstein autoencoder was trained with the Jenssen-Shanon divergence as the discrepancy measure between $Q_{\mathbf{z}}$ and $P_{\mathbf{z}}$ distributions, estimated with a discriminator. P_z was modeled with a multivariate Gaussian distribution. LeakyReLU was used as activation in the WAE discriminator with scale 0.02 for negative input values. ReLU was used in the generator and the encoder, except for the last layer of G where sigmoid was applied. The parameter β in the L_{WAE} loss (7.1) was varied among the following values - 1,5,10 and 20.

Regularized siamese network was trained in a similar manner as in the monochannel case. In all layers, except the last one in G, ReLU activation was used. The last layer was followed by the sigmoid function. The input of the network consists of pairs of patches that were composed in the following way. First, patches were extracted from all the healthy subjects with a stride of 8. Next, for each patch of a subject, a pair was composed by randomly selecting its similar patch among those belonging to the remaining subjects. The number of pairs is again around 3.5 million. The α coefficient of the loss function was set to 0 for 10 iterations, then grew linearly to some α_{max} value for 15 more epochs and remained at α_{max} for 5 more epochs. Adam optimizer with a learning rate 0.001 was used. The batch size was set to 128. We considered two values for α_{max} - 0.25 and 0.5.

Experimental 3D regularized siamese network

In the scope of the early fusion strategy for the proposed CAD we have also evaluated an experimental rSN on 3D patches. Combining the two modalities in the 3D context may improve the performance obtained with 2D early fusion. The encoder and decoder components considered for an alternative 3D architecture are illustrated on fig. 9.1c. The structure of the encoder and the decoder follows those presented earlier for 2D patches. We considered 15 x 15 x 5 patches since most epilepsy cases take up around 5 consecutive transverse slices. The network was trained identically to its 2D analogue with the same strategy of pair constitution.

Outlier detection

The per voxel outlier detection step is identical to the setup described in section 7.3.2. In fact, the only difference in the pipeline introduced earlier is the representation learning stage which now comprises both modalities. Therefore, each voxel v_i is associated with a oc-SVM classifier C_i which is trained on the matrix $M_i = [\mathbf{z_{i1}}, ..., \mathbf{z_{in}}]$ where $\mathbf{z_{ij}}$ is the representation vector corresponding to the multimodal patches centered at v_i of subject j and n is the number of subjects.

The RBF kernel was chosen for each classifier C_i. For each oc-SVM individually, we chose to set γ to the median of the standardized euclidean pairwise distances of the corresponding matrix M_i. The parameter ν, the upper bound of the fraction of allowed outliers in the oc-SVM formulation 5.1, was set to 0.03.

For each voxel v_i, the corresponding oc-SVM model C_i outputs the score for the voxel, i.e.

the distance to the found optimal hyperplane, corresponding to

$$score(v_i) \leftarrow \mathbf{w}^* \cdot \phi(\mathbf{z_i}) - \rho^*$$

where $\mathbf{w}^*$ and ρ^* define the optimal hyperplane, as explained in section 5.1.3. Eventually, all voxel distance scores combined together yield the *distance map* D_p for the given patient p.

9.2.2 Intermediate fusion with multiple kernel learning

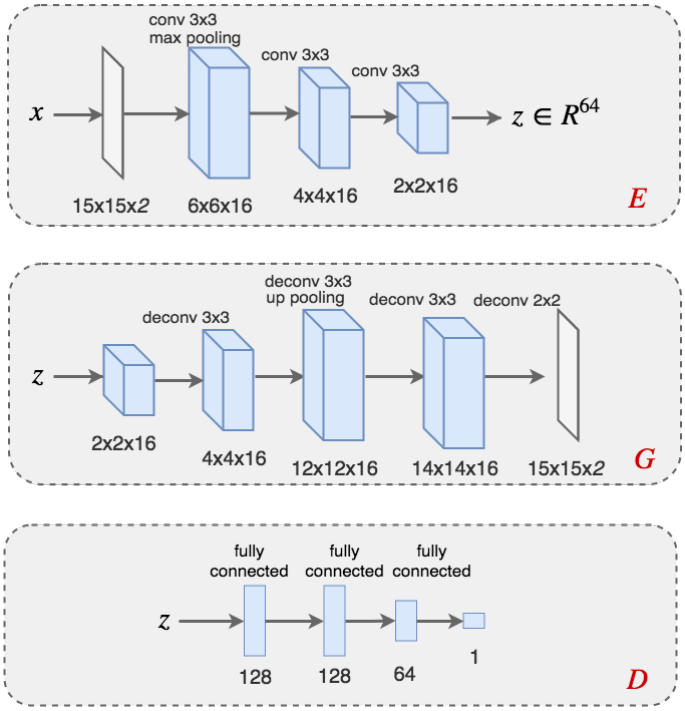
In order to apply the multiple kernel learning paradigm as an intermediate fusion method, we first need to extract representations with networks trained on each modality individually. In chapter 7 we considered several architectures for the T1-weighted MRI modality and showed the superior performance achieved with regularized siamese networks for epilepsy detection. We carried out the same series of experiments on the FLAIR modality and found that the regularized siamese network performed the best on this modality as well. We therefore chose regularized siamese networks as feature extraction components for both T1-weighted and FLAIR data. In both cases the maximal value α_{max} of the tradeoff coefficient α was set to 0.25, corresponding to the best configuration for both modalities.

Outlier detection

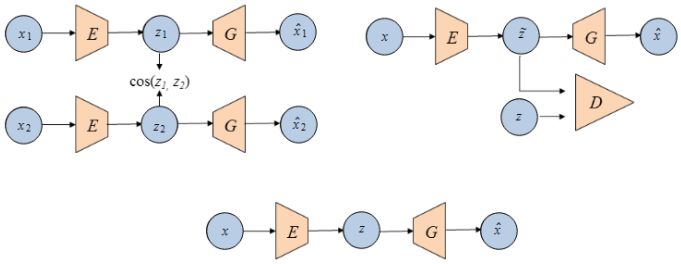
With the individual networks trained, we employ multiple kernel learning on the features extracted per modality. Each voxel v_i is associated with a slimSimpleMKL model C_i, as described in section 8.3, which is trained on the matrices $M_{i,m} = [\mathbf{z_{i1,m}}, ..., \mathbf{z_{in,m}}]$ where $M_{i,m}$ is the matrix of representations for voxel v_i of modality m, $\mathbf{z_{ij,m}}$ is the representation vector corresponding to the patch centered at v_i of subject j for modality m and n is the number of subjects. A kernel is computed on each view i.e. each matrix $M_{i,m}$, and their optimal combination is sought so as to separate the points from the origin. The RBF kernel was used with the same method of γ choice as described in 7.3.2. More precisely, for each voxel v_i and each modality m, the corresponding $\gamma_{i,m}$ was set to the median of the standardized euclidean pairwise distances of the corresponding matrix $M_{i,m}$. We employed slimSimpleMKL by varying the parameter λ among the values - 0, 0.05, 0.1 and 0.5. We used a Matlab implementation of the method provided by [Loosli and Aboubacar, 2017]. Similarly to the oc-SVM employed in chapter 7, the slimSimpleMKL models yield a signed score for each voxel which eventually amounts to a distance score map for a given patient.

9.2.3 Post-processing and performance evaluation

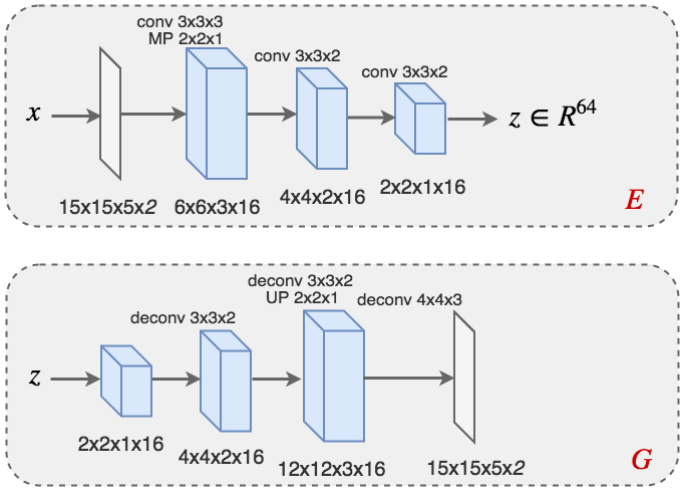
For both early and intermediate fusion based CAD systems the raw output of the system is a signed distance score map where negative scores denote voxels found anomalous by the outlier detection module. Following the strategy described in details in section 7.3.3, we first normalize the raw output maps with a voxel-level score standard deviation map



(a) Encoder E, generator G and discriminator D used in the convolutional autoencoder, Wasserstein autoencoder and regularized siamese network multi-channel architectures.



(b) Global representation of the regularized siamese network (left), Wasserstein autoencoeder (right) and convolutional autoencoder (center). The components E and G have the same structure and are shown in fig. 9.1a.



(c) Alternative 3D encoder and decoder architectures to be used in an experimental 3D multichannel rSN.

Figure 9.1: Multichannel architectures considered for early fusion.

to obtain normalized score maps. The eventual distance map is obtained by averaging the original raw and normalized maps.
Next, all the voxel score values of the distance map are pooled together into a histogram which was then approximated by a non-parametric distribution using a kernel density estimator. The approximated patient's distance score distribution is then thresholded at some pre-chosen *p-value* and a 26-connectivity rule is applied to identify the connected components. These components are referred to as *clusters*. For each patient individually, the distance map is thresholded at the *p-value* that produces at most 15 clusters.
Finally, we rank the detected clusters according to a ranking criterion that privileges large clusters with low average score values. Eventually, the topmost 10 detections are kept.
The evaluation protocol is identical to the one described in section 7.3.4. A given cluster is considered a *true positive* when an overlap exists between the cluster and the ground truth lesion. Otherwise it is considered a *false positive*. A patient is considered detected when at least one true positive cluster is found. Eventually, we calculate the sensitivity (the proportion of the detected patients) and the average number of false positive detections per patient, represented through a fROC curve in the following results.

9.3 Results

9.3.1 Comparison of multichannel architectures for early fusion

We have implemented the CAD system using the features learnt with the multichannel architectures described above. The results are illustrated on fig. 9.2a. The baseline architecture, that is the convolutional autoencoder, is, as was the case in the mono-modal setting, the least successful. The 2 choices for the coefficient α[1] in the rSN showed the best results, especially given that the number of parameters is the same for both CAE and rSN. This only emphasizes the advantage of the proposed regularized siamese architecture in the context of anomaly detection. Eventually, with $\alpha = 0.5$ the CAD system detects around 62% of epilepsy lesions for 8-9 false positive detections. As it can be seen, the performance achieved with WAE features varies significantly for different choices of β. The maximum sensitivity is around 53% for 8-9 false positive detections, obtained with $\beta = 1$.
The performance gain obtained with the additional FLAIR channel at input becomes evident. Indeed, the maximum sensitivity achieved with rSN on T1-w/FLAIR data is 62% for 8-9 FPs while being 42% for the same number of false positives in the T1-w CAD system. Same is true for the WAE architecture as well, increasing from 38% to 53%. The comparative performances of monochannel versus multichannel architectures are illustrated on fig. 9.2c.

[1]Hereafter, rSN $\alpha = *$ actually refers to $\alpha_{max} = *$. We use this shorthand for practical reasons.

We have also evaluated the performance obtained with an experimental 3D multichannel regularized siamese network as a feature extractor component for the proposed CAD. This should give us an insight on the importance of the 3D view for the problem at hand. Fig. 9.2e shows the fROC curves with for the 3D rSN for 3 different values of $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 0.75$. When comparing the results to the best performance achieved in the 2D setting, as shown on fig. 9.2f, it can be seen that the maximum sensitivity achieved remains at 62%. However, the 3D architecture seems to outperform the 2D alternative for the sensitivities at earlier fROC curve points, in other words, for less FPs. So, for 2FPs the 3D network achieves around 48% sensitivity while the 2D rSN achieves only 38%.

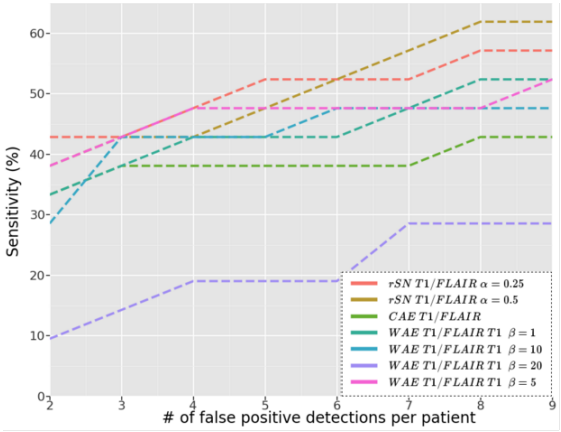
9.3.2 Intermediate fusion strategy with MKL

As an intermediate fusion strategy, we considered the multiple kernel learning paradigm and its implementation with (slim) SimpleMKL algorithm. This implies varying the parameter λ in the formulation 8.7. We have tried the following values - 0 (original SimpleMKL formulation), 00.5, 0.1, 0.25 and 0.5. Fig. 9.2b illustrates the performances in the scope of this experiment. The $\lambda = 0.1, 0.25, 0.5$ all resulted in an identical performance and, therefore, the corresponding curves overlap and are seen as a single one. Apparently, there is a limit on how many support vectors are kept eventually when solving the problem 8.7 which in our case happened to be bound to the value $\lambda = 0.1$. From this comparison, it is apparent that even a slight regularization of the number of support vectors with $\lambda = 0.05$ offers a significant improvement over the original SimpleMKL problem. The sensitivity jumps from 32% for 9 false positives to 52%.

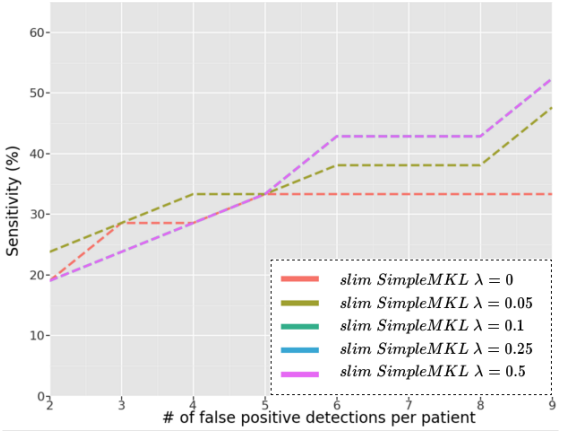
9.3.3 Comparison of fusion levels

Despite the promising performance obtained with intermediate fusion, the results are still inferior to those obtained with multichannel rSNs as shown on fig. 9.2d. This may be due to the fact that certain properties based on the combination of raw image modalities are learnt during the multichannel network training while the intermediate fusion operates on the already learnt features which may skip those properties. Additionally, multichannel fusion has an advantage in terms of efficiency of implementation. Indeed, only one network is trained and the eventual outlier detection algorithm (oc-SVM) is computationally lighter than the multiple kernel learning alternative. We do find multiple kernel learning an interesting approach which deserves to be explored more thoroughly in the future.

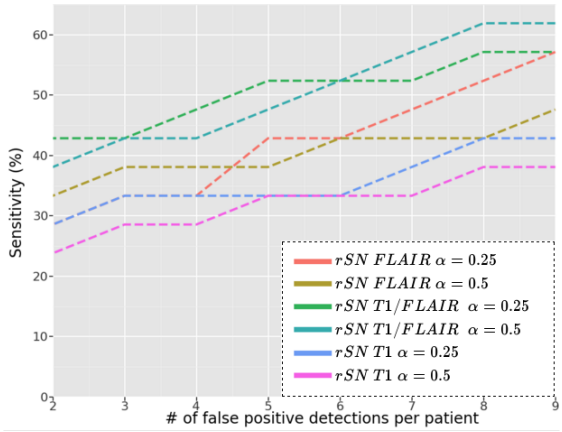
Table 9.1 summarizes the results obtained for each patient in the scope of our comparison of early and intermediate fusion strategies. Several observations could be made. First, multimodal CAD system, with both early and intermediate fusion strategies, offers an improvement over the T1-only CAD. The early fusion with multichannel networks allows to detect almost all the patients detected with at least one modality independently. The



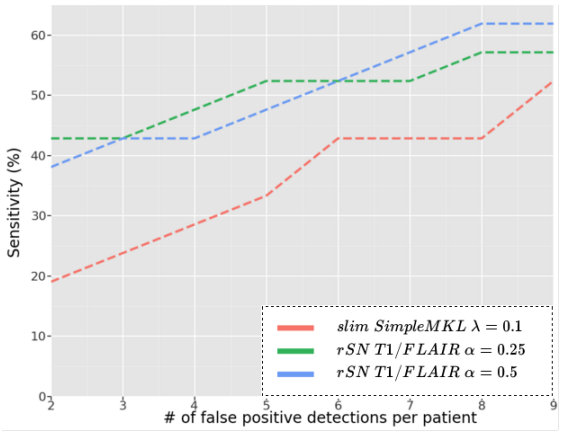
(a) fROC curves of the CAD system with multichannel architectures for early fusion. The architectures correspond to convolutional autoencoder (CAE), Wasserstein autoencoder (WAE) with $\beta = 1, 5, 10, 20$ and regularized siamese network (rSN) with $\alpha = 0.25$ and $\alpha = 0.5$.



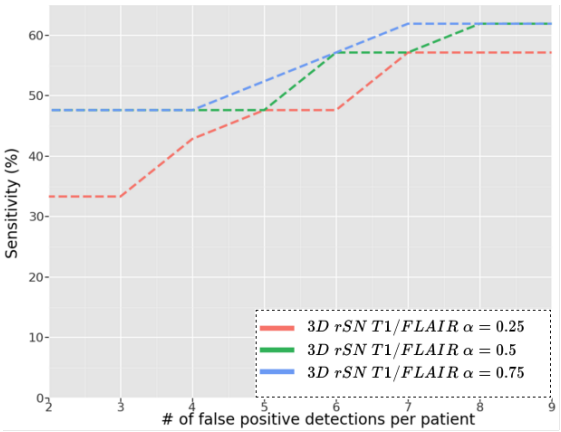
(b) fROC curves of the CAD system with multiple kernel learning for intermediate fusion. The curves correspond to the performance obtained with slim SimpleMKL for different values of $\lambda = 0, 0.05, 0.1, 0.25, 0.5$. The λ values above 0.1 give identical performances and therefore overlap.



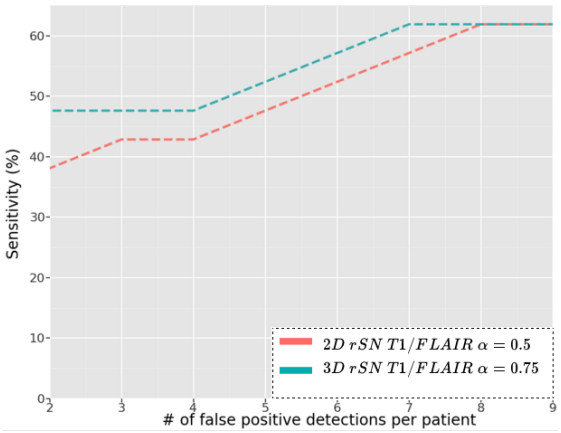
(c) fROC curves of the CAD system with monochannel and multichannel regularized siamese networks (rSN) on T1-w, FLAIR and T1-w/FLAIR MRI.



(d) fROC curves of the CAD system with early and intermediate fusion. The best performance obtained with intermediate fusion is compared to the 2 best multichannel architectures for early fusion.



(e) fROC curves of the CAD system with 3D multichannel regularized siamese networks (rSN) on T1-w/FLAIR MRI for 3 values of α.



(f) fROC curves of the CAD system with the best 2D and 3D multichannel regularized siamese networks (rSN) on T1-w/FLAIR MRI.

Figure 9.2: The proposed CAD system performance with early and intermediate fusion.

only two exceptions are patients A^- and R^+. The former was detected with the T1-based CAD, however, with a low rank of 8. It is likely for lesions, ranked at the bottom half of the allowed top 10 clusters, to be smeared out when an even more normal looking imaging of the second modality is added. Patient R^+ had an extremely subtle lesion which was visually identified on FLAIR imaging. It consists of a very small hypersignal which is ranked 6 by the FLAIR-only CAD. It seems to us that the normality of its T1-w MRI smeared out the contribution of the FLAIR data in the multichannel setting. Another interesting observation is that two patients (L^- and M^-) were detected with the multichannel network while not being detected with either of monomodal CADs. This supports our intuition that there is complementary information present in different imaging data that, when considered jointly, allows a more coherent detection on some patients. It should also be stated that despite the quantitative advantage shown by the early fusion strategy, the intermediate fusion yielded some plausible results as well. So, patients A^- and C^- were detected with MKL while being missed with the multichannel network. We do find the MKL approach an interesting method of coupling multiple views on the problem, that should be looked into in the future work.

Eventually, some remarks should be made on the 2D early fusion versus 3D early fusion. The 3D context seems to significantly improve the detection ranks of two patients - F^- and Q^-. This speaks of the potential advantage of exploiting the 3D information, even though the maximum sensitivity achieved with both 2D and 3D multichannel architectures is the same. We should also mention that, though some patients (A^- and C^-) were missed by the 2D rSN and detected by the 3D alternative, the contrary is also true. Patients F^- and M^- were detected by the system based on 2D rSN and not with the 3D one. These discrepancies should be considered in a further analysis. It is likely that improving the 3D network structure will resolve the detection of such cases.

9.3.4 Visual analysis

We have implemented two fusion strategies of imaging modalities within our CAD pipeline and have shown that early raw image combination in a multichannel regularized siamese network achieves promising performance. We will next visualize the detection maps of both strategies and compare them with each other, as well as with those obtained with the maps of the T1-w only CAD system.

Fig. 9.3 illustrates the normalized output maps obtained with intermediate fusion with MKL, early fusion with 2D rSN and early fusion on 3D rSN. The last column shows the outputs of the system with 2D rSN on T1-w MRI. As it can be seen, the MKL output maps demonstrate a very mild contrast between the scores around the lesion and elsewhere on the image. Conversely, the output maps corresponding to the early fusion strategies exhibit significantly negative scores in the areas around the lesions. The abnormalities are more striking on the multimodal CAD output (3rd column) than on the monomodal one

Patient	**Lesion location**	**T1 rSN** $\alpha = 0.25$	**FLAIR rSN** $\alpha = 0.25$	**T1/FLAIR rSN** $\alpha = 0.5$	**slimSimpleMKL** $\lambda = 0.1$	**T1/FLAIR 3D rSN** $\alpha = 0.75$
Patient A^-	Insula R	✓(8)	✗	✗	✓(7)	✓(2)
Patient B^-	Temporal Lobe L	✓(1)	✓(2)	✓(1)	✓(1)	✓(1)
Patient C^-	Hippocampus R	✗	✗	✗	✓(10)	✓(7)
Patient D^+	Superior frontal gyrus R	✓(2)	✓(3)	✓(1)	✓(4)	✓(1)
Patient E^-	Inferiolateral remainder of parietal lobe R	✗	✓(10)	✓(8)	✓(5)	✓(6)
Patient F^-	Hippocampus L, parahippocampus L	✗	✓(3)	✓(9)	✗	✗
Patient G^+	Middle frontal gyrus L	✓(4)	✓(1)	✓(1)	✓(10)	✓(1)
Patient H^-	Superior frontal gyrus R	✓(1)	✓(8)	✓(3)	✓(1)	✓(2)
Patient I^-	Hippocampus L, parahippocampus L	✗	✗	✗	✗	✗
Patient J^-	Precentral gyrus R	✗	✗	✗	✗	✗
Patient K^-	Superior temporal gyrus R	✗	✗	✗	✗	✗
Patient L^-	Middle frontal gyrus R	✗	✗	✓(1)	✗	✓(1)
Patient M^-	Anterior temporal lobe R	✗	✗	✓(4)	✗	✗
Patient N^-	Anterior temporal lobe R	✓(9)	✓(1)	✓(1)	✓(7)	✓(1)
Patient O^-	Middle frontal gyrus L	✓(1)	✓(6)	✓(2)	✓(1)	✓(1)
Patient P^-	Hippocampus R	✗	✗	✗	✗	✗
Patient Q^-	Lateral remainder of occipital lobe L	✓(2)	✓(3)	✓(7)	✓(6)	✓(2)
Patient R^+	Orbital gyrus R	✗	✓(6)	✗	✗	✗
Patient S^-	Hippocampus R	✗	✓(8)	✓(6)	✗	✓(8)
Patient T^-	Posterior temporal lobe R	✗	✗	✗	✗	✗
Patient U^-	Posterior temporal lobe L	✓(1)	✓(4)	✓(3)	✓(1)	✓(2)
Overall # of detections		**9**	**12**	**13**	**11**	**13**

Table 9.1: Comparative results of different configurations of the CAD system at patient level. For each patient, column 2 reports the lesion location while columns 3 to 7 indicate, for each CAD setting, if the lesion was detected (✓) or missed (✗), as well as the rank of the true detection inside parentheses.

(last column). This seems to justify the contribution of the FLAIR modality.
Fig. 9.4 depicts the output cluster maps obtained with the early and intermediate fusion strategies. The number of clusters in each image correspond to the smallest number of FPs allowing to detect the true lesion, when it is detected, and the top 10 clusters, when not. The differences in the number of FPs allows to compare which models rank the true epilepsy abnormalities higher than others. The most significant qualitative differences between the two fusion strategies are seen for patients D^+, N^- and G^-. For the former, the detection with MKL is very slight while for the latter two patients the number of FPs is much higher than those obtained with early fusion.
Comparing the multichannel and monochannel architectures in the middle and rightmost columns, the advantage of considering both T1-w and FLAIR MRI becomes clear. The patient E^- illustrates one example when the T1-weighted MRI alone is unable to detect the subtle lesion. On the other hand, the early fusion strategy with 3D and 2D rSNs seems to result in quite similar cluster maps.

9.4 Conclusion

In this chapter we presented possible scenarios for the integration of multimodal imaging data into the proposed CAD system. We focused on two fusion strategies - early fusion and intermediate fusion. Our main contributions consist in

1. formulating and proposing multichannel unsupervised architectures as an early fusion strategy
2. proposing a multiple kernel learning approach as an intermediate fusion strategy
3. evaluating both strategies on problem of epilepsy lesion detection on T1-weighted/FLAIR multimodality imaging data.

The early fusion strategy consists in learning joint representations of multimodality data fed at input to a multichannel architecture. The intermediate fusion approach learns the boundary of normal training points by assigning kernels and combining the representations learnt with each modality separately. We evaluated both methods on the combination of T1-weighted and FLAIR MRI data and offered a comparison of the achieved performances. Overall, several remarks could be deduced from the described experiments. First, the performance gain achieved with the integration of the additional FLAIR modality improved the performance obtained with T1-weighted MRI only. Second, the early fusion method of learning joint representations with multichannel networks results in the best performance when regularized siamese networks are used. The performance is further improved when such an architecture is based on 3D patches. The MKL approach, even though showing promising results, was rather inferior to the results obtained with early fusion.
As it can be seen from tables 3.2, 3.3-3.5, summarizing the state-of-the-art methods for

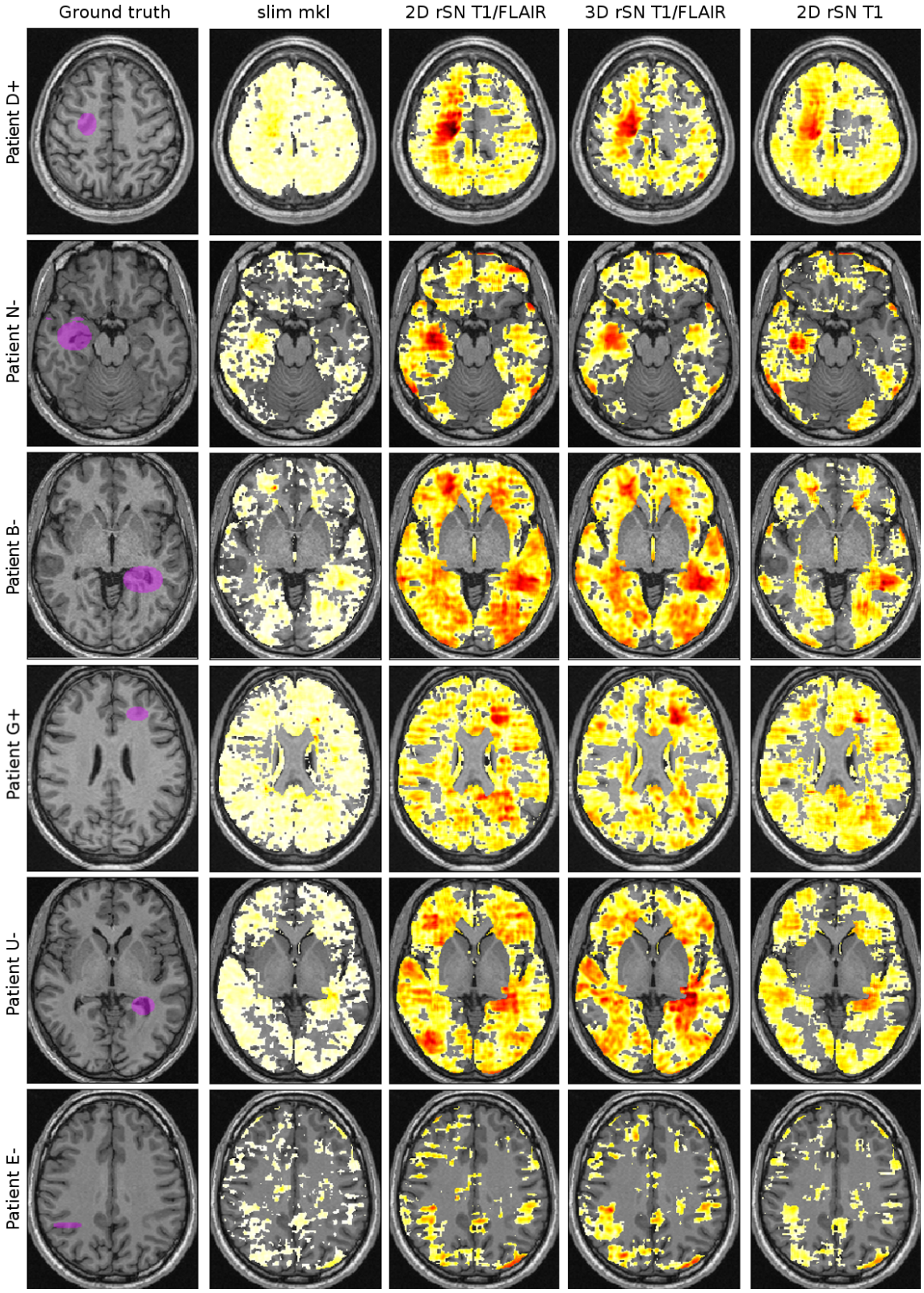


Figure 9.3: Visualization of the CAD detections. First column: original slice centered at the lesion highlighted in a purple circle. From second to the fifth column: the normalized output maps obtained with intermediate fusion; multichannel 2D rSN with $\alpha = 0.5$; multichannel 3D rSN with $\alpha = 0.75$; mono-modal rSN on T1w MRI. Darker shades correspond to more negative score and, hence, the detected anomalies. Note the differences in the contrast in different settings. Intermediate fusion results in a rather low contrast between the lesions and the rest of the images.

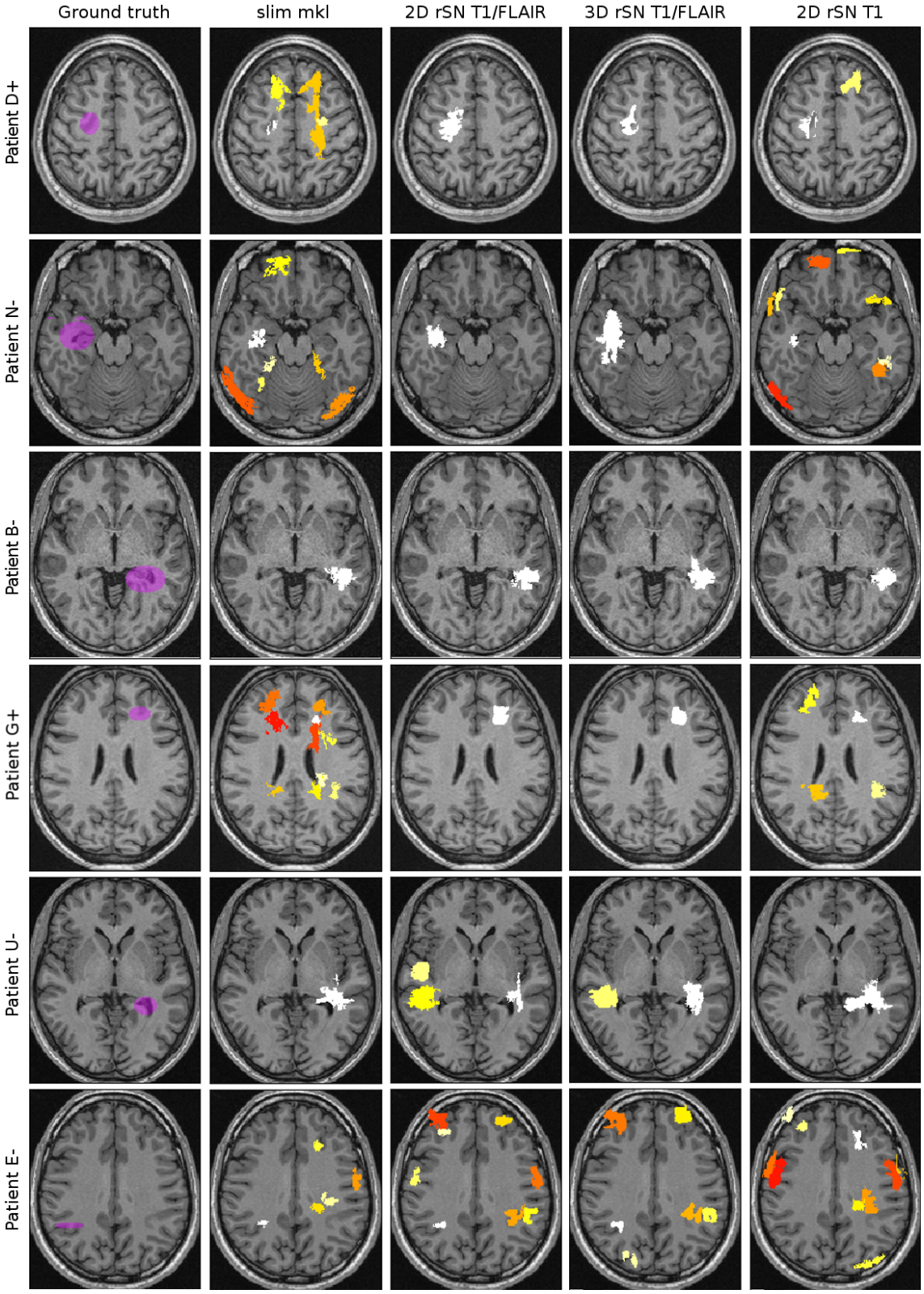


Figure 9.4: Visualization of the CAD detections. The cluster maps show the minimum number of FP clusters allowing to detect the lesion, when it is detected, and top 10 clusters, when it is not. Some clusters' projections may overlap (so visually their number might be underestimated). First column: original slice centered at the lesion highlighted in a purple circle; second to fifth column: the cluster maps obtained with intermediate fusion; multichannel 2D rSN with $\alpha = 0.5$; multichannel 3D rSN with $\alpha = 0.75$; mono-modal rSN on T1w MRI. Note the difference in the number of FP clusters needed before the true detection emerges. This corresponds to the rank of a true detection.

epilepsy detection, current studies do not commonly employ multimodality data. Those that do, do not necessarily combine multimodality features but rather do individual analysis per feature. Moreover, the existing CAD systems do not integrate raw multimodality data but the features extracted from different modalities. The only exception to the observed pattern is the approach proposed in [Gill et al., 2018] (that combined raw T1-w and FLAIR MRI data in a supervised convolutional neural network based framework). The improvement established in our CAD system, when considering both T1-weighted and FLAIR MRI data, only shows that the integration of multiple imaging modalities should be an important aspect of dedicated CAD systems.

Eventually, we turn into placing the performance achieved with the proposed CAD system into the grid of the existing works. Some of these studies use manually designed features characterizing cortical malformations based on surface based morphometry (SBM) [Thesen et al., 2011, Hong et al., 2014, Ahmed et al., 2016]. Others associate these morphometric features to the intensity anomalies in T1w MRI mainly caused by heterotopia lesions [El Azami et al., 2016, Gill et al., 2017]. Our method seeks to find more complex features in an unsupervised manner in order to identify lesions with unknown signatures. Naturally, such an approach, when applied to a specific pathology, is likely to produce more false positive detections.

Although a fair comparison with published results is difficult because of the differences in the patient groups, results reported in table 9.1 (62% sensitivity for 8-9 false positives per scan) are of the same order as those reported in recent studies for the difficult task of automated detection in MRI negative patients. Indeed, the system proposed in [Ahmed et al., 2016] based on SBM features coupled with semi-supervised hierarchical conditional random fields achieves between 52% and 70% sensitivity (depending on the feature) on a sample of 20 T1 weighted MRI negative patients among the top 10 detections per scan. In [El Azami et al., 2016], a CAD system 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 when evaluated on a small cohort of 10 T1w MRI negative patients. The recent supervised approach in [Gill et al., 2018], that combines raw T1-w and FLAIR MRI data in a supervised convolutional neural network, achieves between 83-91% sensitivity; the system, however, was trained on a significantly higher number of cases. The most important difference, however, lays in a common pattern among all the existing methods - targeting a specific cause of epilepsy. Most frequently, the pathology of interest is FCD, and particularly FCD type II. This category of the epilepsy causes is the most likely one to have recognizable markers on neuroimaging. The data set, considered in our study and described in section 5.2, mainly consists of challenging, purely MRI negative cases where the patients were considered normal over multiple visual examinations and retrospective studies. For most of them, the histopathological analysis did not reveal any

clear characteristics (e.g. FCD). Since these are the true challenges among patients diagnosed with epilepsy, we find that the proposed CAD system meets the expectations and achieves reasonable results. The performance, however, could be improved further.
Our CAD system 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 have not appeared 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 lesion. 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.

Chapter 10

Epilepsy lesion detection on PET/MR images

In the previous chapters we explored CAD systems for automated epilepsy lesion detection on MR images. In chapter 3 we presented and discussed the existing approaches for such systems. Chapter 4 presented our approach to the problem of epilepsy lesion detection on MRI. In chapter 6, we introduced various unsupervised deep architectures, including our own configuration of siamese networks, to be used as representation learning mechanisms in a framework that casts subtle brain abnormality localization task as a voxel-level outlier detection problem. We performed an evaluation of the proposed framework trained on T1-weighted MR images on a set of patients with confirmed epilepsy lesions. In chapter 9 we extended the proposed framework to integrate both T1-weighted and FLAIR MRI modalities.

In this chapter, we make an attempt to explore the potential of PET imaging, as a complementary modality to MRI, for an automated detection of subtle epilepsy lesions. PET imaging is not a routine clinical exam for the evaluation of drug resistant epilepsy patients. The PHRC (programme hospitalier de recherche clinique) research project initiated by our main collaborator J. Jung, thus, aims to evaluate the impact of this modality on epilepsy detection. The patients selected in the scope of this project had therefore multiparametric MRI and PET exams. It is, however, more difficult to set up a data set of healthy subjects with PET acquisitions, as described in section 5.2. As a consequence, the healthy controls who had PET exams are less in number than those who had T1-weighted and FLAIR imaging in our data set.

In this chapter, we propose to handle the problem of incomplete data through synthesizing PET images from the corresponding MRI acquisitions. The synthetic PET images and

the real ones are later exploited in the CAD system described in chapter 5 and the contribution of the generated PET data is shown via the quantitative evaluation of the system performance.

10.1 Number of training examples: limitation

As described in section 5.2, the number of healthy subjects whose PET acquisitions are available for the problem is only 35. One of the key points, therefore, becomes the question of how much the small number of data points would limit the performance of the proposed framework. To this end, we evaluate the CAD system trained on the available PET images on 2 post-surgical images of patients, not included in our data set, shown on fig. 10.1. The post-surgical PET scans carry large evident abnormalities in the resected zones which are clearly visible by bare eye and should be easily identified by the proposed system. The representations are learnt with the simplest architecture so far - a convolutional autoencoder, with the same architecture and training routine that was presented in section 7.3.1 for T1-weighted MRI. The oc-SVM design is as described in section 7.3.2. Fig. 10.1 illustrates the generated raw score maps for the post-surgical scans. As it can be seen, the distribution of the oc-SVM output scores across all the voxels assigns anomalousness to almost everywhere. There is no striking difference in the resected zone which constitutes to a failure of the system, trained on 35 data points per voxel, to recognize the large abnormalities present in the images. The impact of the small number of subjects on the representation learning component is an aspect that should be investigated. We, however, hypothesize that such a behaviour is due to the insufficient number of data points to learn the normality of each voxel, taking place in the second step of the CAD system i.e. oc-SVM model learning per voxel. It is, therefore, our objective to explore strategies to increase the number of PET training samples. Our strategy is to synthesize PET data from the corresponding MRI acquisitions and leverage the synthetic PET data in the oc-SVM learning.

10.2 Cross-modality synthesis in medical imaging

In many medical applications, it is common to have missing data. In particular, in multimodal imaging studies, for one reason or another, some subjects may happen to have an impartial set of multimodal data acquisitions. Discarding these subjects, especially given that the overall number of participants in medical problems is low, may result in underperforming models, as showcased above. One modern approach to account for the missing data is the synthesis of missing modalities from those that are present. This problem has been tackled in many recent studies.

[Nie et al., 2017] developed a method for synthesizing CT images from MRI, for brain and pelvic data sets. The method uses GANs to generate CT patches from MR image patches

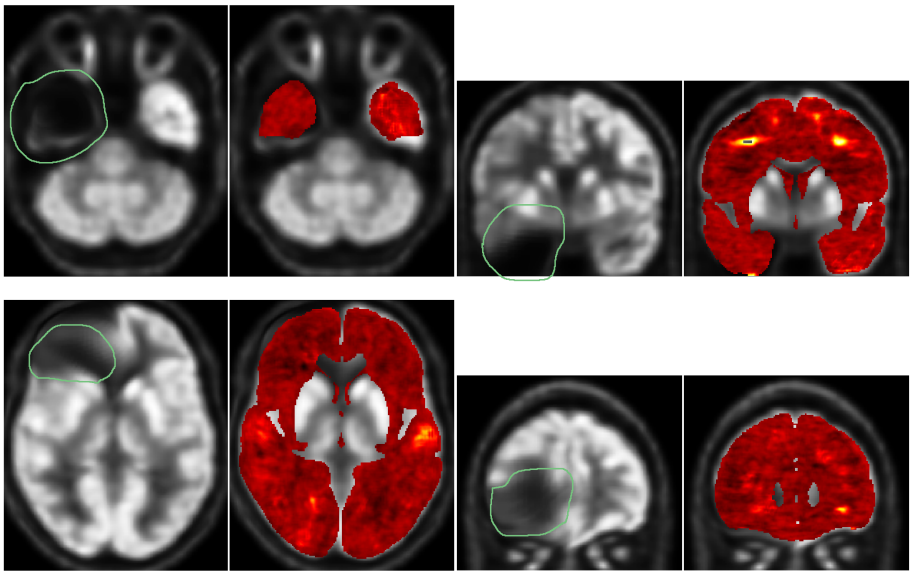


Figure 10.1: CAD system output on post-surgical scans of two patients, showcasing evident anomalies in the resected zones. Each row corresponds to a patient, from left to right - transverse slice centered at the resected area, transverse slice of the system output, coronal slice centered at the resected zone and a coronal slice of the system output. The resected areas are contoured in green. Darker colors on the output maps correspond to more negative values. From the almost uniform distribution of oc-SVM scores among the voxels, it is clear that most voxels were seen as anomalous. There is no striking difference in the anomalousness of the resected zone. These cases showcase the failure of the system to recognize obvious anomalies.

and further integrates the GANs into an auto-context model (ACM) in order to refine the generated images. This method, however, requires an access to a sufficient amount of paired CT and MRI training data. In [Wolterink et al., 2017], CT image synthesis from MRI is performed on unpaired data, using a cycleGAN [Zhu et al., 2017a]. CycleGAN is a variation of GANs enhanced with a bidirectional generation of images from one domain to the other and a constraint on the consistency between an image and its reconstruction through the *cycle* of two generators. CycleGAN is again modified in the MRI to CT image synthesis task in [Xiang et al., 2018] and in [Yang et al., 2018] where a constraint on the structural consistency between the synthetic and real images is integrated into the framework.

Several studies performed synthetic image generation while evaluating the synthesized images in an auxiliary supervised task. In [Li et al., 2014], PET image patches were generated from MRI patches with a 3D convolutional neural network in the first step and later used for 3 classification settings, discriminating Alzheimer's disease patients and patients with mild cognitive impairment from healthy controls. [Ben-Cohen et al., 2018] proposed a conditional GAN, where liver PET images are generated from CT input images through a convolutional network while the discriminator seeks to distinguish real CT-PET pairs from CT-generated PET pairs. Additionally, the reconstruction error of the generated PET images is added to the global loss. The synthesized images were further

evaluated in a lesion detection task, showing an improvement in the average false positive rate. [Chartsias et al., 2017] performed cardiac MRI synthesis from CT images using a cycleGAN and then evaluated the advantage of the synthesized images in a U-Net like segmentation network. [Zhang et al., 2018] proposed a framework based on the cycleGAN where the complementary modalities, i.e. CT and MRI, are used to synthesize examples of one another and are evaluated on a cardiac image segmentation task. In addition to the cycleGAN loss, the method imposes a term assuring the consistency of the segmentation of the imputed images and the ground truth label map for cardio-vascular diseases. This variation imposes a certain consistency between the synthesized and real images. Similarly, [Pan et al., 2018] used a 3D cycleGAN to synthesize PET images from MRI and later used them in a supervised method classifying Alzheimer patients versus healthy controls. [Jiang et al., 2018] proposed a cycleGAN based framework modified in order to generate MR images from CT scans by preserving the tumors of lung cancer patients. The synthesize MR images are later used in a U-Net for lung tumor segmentation.
The mentioned studies treat image synthesis problem rather as a pre-processing step, either for standalone image generation or to be used later in an independent medical task. Some recent studies have proposed methods that perform modality imputation, at the same accounting for the eventual task at hand, such as segmentation. Employing a similar strategy, [Orbes-Arteaga et al., 2018] developed a method that combines a module generating FLAIR images from T1-weighted brain MRI and another module that aims at segmenting white matter hypointensities. In [Huo et al., 2018], the authors propose a cycleGAN based abdomen CT to MRI synthesis component, together with a segmentation component on the real and generated CT. To employ such a strategy, labeled ground truth references are required which makes it difficult to use the approach in unsupervised settings.

Among the works mentioned above, many solely aim to produce realistic looking synthetic images, evaluating the quality of the approaches through a quantitative analysis of the results (with such metrics as mean absolute error and peak-signal-to-noise-ratio), with no particular medical task at hand. The second group of methods either evaluates the quality of the generated images on an auxiliary task, or develops a method, explicitly accounting for the eventual medical application.
We are interested in the second category of approaches that couple an image synthesis component to a specific task. As it can be seen from the recent studies above, the impact of the synthesized data is usually evaluated in supervised contexts. Eventually, we are interested in evaluating the influence of synthetic images in the task of anomaly detection.

10.3 Data description

10.3.1 Original PET-MRI data set

The experiments below are based on the data set described in details in section 5.2. In particular, the T1-weighted MRI and PET acquisitions of 35 healthy controls and 19 patients (patients T^- and U^- did not have PET exams) with confirmed epilepsy lesions are considered. Additionally, 40 T1-weighted MR images of healthy controls are available. The T1-weighted and PET images were co-registered and normalized to the MNI space, as described in section 5.2.4.

10.3.2 MRI to PET synthesis with U-Net

U-Net architecture introduced in [Ronneberger et al., 2015] has made a major contribution in the application of deep architectures on various medical problems. Originally, the network was proposed for a segmentation problem, yielding an impressive performance even when very few training examples are available. [Çiçek et al., 2016] proposed a 3D version of the original U-Net. The main characteristic of the U-Net architecture is the contracting path, a sequence of convolutional and max pooling layers, typical to modern convolutional networks, and an expansive path, a series of up-sampling and up-convolution layers. There are skip connections introduced between the layers of the two paths which results in the layers of the expansive paths receiving at input not only the output of the previous layer but also the output of the symmetrically located layer in the contracting path. The advantage of such connections is that the spatial information lost during consecutive maxpooling operations is eventually recovered. Additionally, there are two dropout layers at the end of the contracting path.

We consider exploiting a U-Net-like architecture in order to generate PET images from MR images fed as input to the network. In other words, an architecture will be given a set of MR images and its weights will be optimized so as to produce PET images as close as possible to the corresponding PET scans of the input images. Naturally, this approach requires a set of paired MRI-PET images.

The architecture applied for MRI to PET synthesis is shown on fig. 10.2. The 2D transverse slices of all the MRI-PET pairs of healthy controls were gathered from all 35 healthy controls. The slices were cropped to 160 x 160, normalized to 0-1 interval at image level and fed into the network. The input to the network, thus, consists in MRI images slices, with the ground truth for the output being the corresponding PET slices. The network was optimized with the Adam optimizer with a learning rate of 0.0001, using the mean square error between the output of the network and the corresponding ground truth PET slice as the loss function. The network was trained for at least 25 epochs and, then, early stopping was implemented based on the performance of 3 healthy controls left out as a validation set. We performed data augmentation by generating transformations by vertically flipping

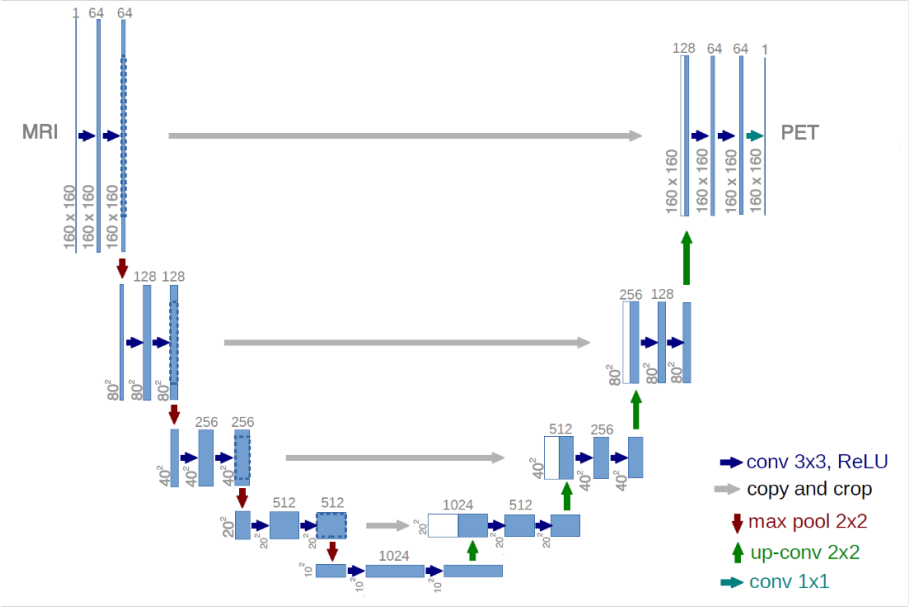


Figure 10.2: The U-Net architecture exploited to synthesize PET images from MRI scans.

and rotating the images within a range of 15 degrees. We did not explore more complex data augmentation strategies. Examples of PET images generated with the network are illustrated on fig. 10.3. As it can be seen, the synthesized images look realistic; however, they do not reproduce the original PET slices with a high fidelity. Since this experiment is only at its preliminary stage, we have not tested other approaches that could result in images of a better quality.

10.3.3 Hybrid PET-MRI data set

After performing MRI to PET synthesis, we are left with a hybrid data set consisting of real MRI data and real and synthesized PET data. Precisely, 35 healthy controls have real MRI and PET acquisitions while the remaining 40 controls have real MRIs and synthetic PET images, generated from the real MRI acquisitions via the U-Net-like architecture described in the previous section. We will leverage this mixed data set to evaluate the potential of the synthesized PET images in the CAD system developed in this work.

10.4 Experiments

The experiments below explore the potential of synthesized data for the problem of outlier detection. To this extent, we leverage the hybrid data set obtained through pulling together the real MRI and real/synthetic PET images. The CAD system is identical to its description given in chapter 5. The representation learning component, implemented with an unsupervised deep architecture, is trained on real inputs only i.e. the patches extracted from the 35 healthy subjects with both MRI and PET acquisitions. The introduction of the synthetic PET data takes place within the second component of the system - per

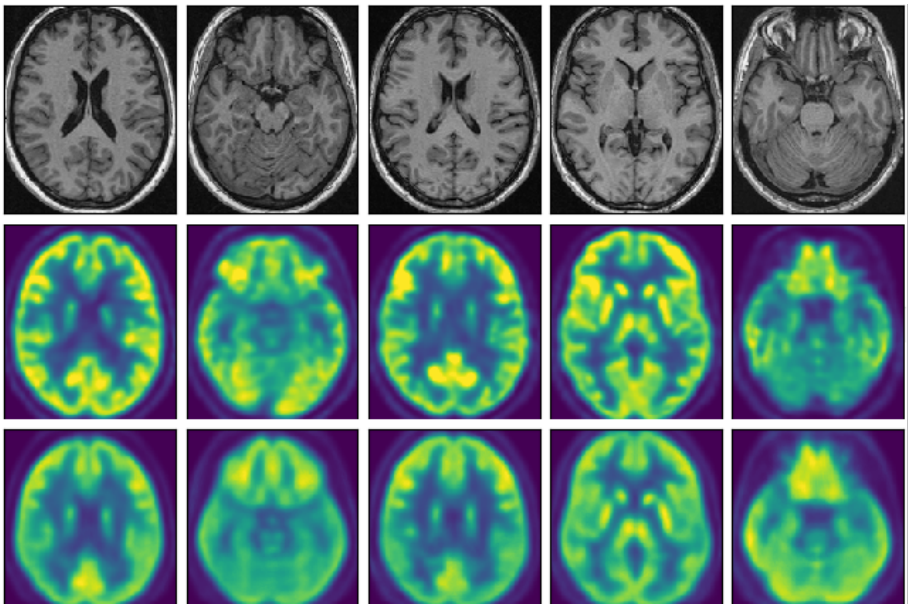


Figure 10.3: Examples of PET images synthesized with U-Net. Top row: original MRI transverse slices, middle row: original PET transverse slices, bottom row: synthetic PET transverse slices.

voxel outlier detection with oc-SVM models. The main motivation behind this setup is to evaluate the synthetic data in the context of outlier detection. In the future, the analysis should be extended to the representation learning component as well.
Eventually, the following configurations of the CAD system are evaluated and compared.

1. PET-only CAD system with
    - per voxel oc-SVMs trained on the real PET data of 35 healthy controls
    - per voxel oc-SVMs trained on the real PET data of 35 healthy controls and 40 synthetic volumes
2. MRI-PET CAD system with
    - per voxel oc-SVMs trained on the real MRI-PET data of 35 healthy controls
    - per voxel oc-SVMs trained on the real MRI of 75 healthy controls and real PET data of 35 healthy controls and 40 synthetic PET volumes

In both settings, we aim at exploring the difference in performance when the synthetic data is introduced.

10.4.1 Baseline architecture for representation learning

As the representation learning component of the CAD system, we will reuse the baseline architecture introduced in the proposed CAD system for epilepsy detection on T1-weighted MRI in section 7.3.1. Namely, we will consider a convolutional autoencoder shown on fig. 10.4. In chapter 7 and, further, in chapter 9, we have shown that the regularized siamese

network and Wasserstein autoencoder outperform the simple convolutional autoencoder. This experiment, however, is only at its preliminary stage and, thus, we consider the simplest suitable choice, expecting to choose better architectures in the future work.
The input to the CAE network consists in 15 x 15 patches extracted from the images with a stride of 8. This architecture is exploited in two settings - PET-only monochannel and T1-weighted MRI/PET multichannel scenarios. In both cases, only the *real* image acquisitions were given to the network at input. In the first case the number of channels at input is equal to 1 and to 2 in the second case. The encoding path consists of 3 hidden layers with kernel size 3x3 where only the first layer is followed by a max pooling layer. The decoding path is designed in a similar fashion. We used ReLU activation function in all the layers except for the last one where sigmoid is used. Similarly to the setup in section 7.3.1, this network was trained to optimize the mean squared error of the input patches and the corresponding reconstructions output by the network, using Adam optimization algorithm with learning rate=0.001 and momentum=0.5.

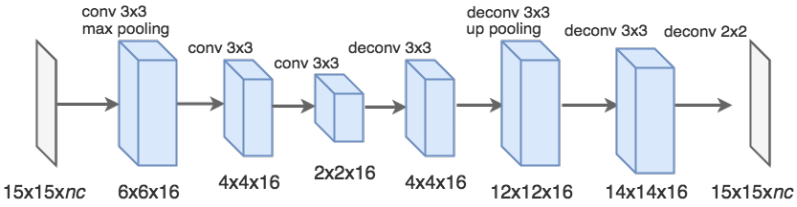


Figure 10.4: Stacked convolutional autoencoder architecture (CAE). nc at input denotes the number of channels that depends on the setting. For the monomodal scenario, $nc = 1$ and for the multimodal scenario $nc = 2$.

10.4.2 Outlier detection and post-processing

The per voxel outlier detection step is identical to the setup described in section 7.3.2. Each voxel v_i is associated with a oc-SVM classifier C_i which is trained on the matrix $M_i = [\mathbf{z_{i1}}, ..., \mathbf{z_{in}}]$ where $\mathbf{z_{ij}}$ is the representation vector corresponding to the (multimodal) patch(es) centered at v_i of subject j and n is the number of subjects.
The RBF kernel was chosen for each classifier C_i. For each oc-SVM individually, we chose to set γ to the median of the standardized Euclidean pairwise distances of the corresponding matrix M_i. The parameter ν, the upper bound of the fraction of allowed outliers in the oc-SVM formulation 5.1, was set to 0.03.
For each voxel v_i, the corresponding oc-SVM model C_i outputs the score for the voxel, i.e. the distance to the found optimal hyperplane, corresponding to

$$score(v_i) \leftarrow \mathbf{w}^* \cdot \phi(\mathbf{z_i}) - \rho^*$$

where $\mathbf{w}^*$ and ρ^* define the optimal hyperplane, as explained in section 5.1.3. Eventually, all voxel distance scores combined together yield the *distance map* D_p for the given patient

p.

The post-processing of the output D_p maps for the patients was carried out as described in section 7.3.3. For a given patient, the raw output map is first normalized with a voxel-level score standard deviation map. The eventual distance map is obtained by averaging the original raw and normalized maps.

Next, all the voxel score values of the distance map are pooled together into a histogram which was then approximated by a non-parametric distribution using a kernel density estimator. The approximated patient's distance score distribution is then thresholded at some pre-chosen *p-value* and a 26-connectivity rule is applied to identify the connected components. These components are referred to as *clusters*. For each patient individually, the distance map is thresholded at the *p-value* that produces at most 15 clusters.

Finally, we rank the detected clusters according to a ranking criterion that privileges large clusters with low average score values. Eventually, the topmost 10 detections are kept.

10.5 Results

In order to estimate the contribution of synthesized PET data, we evaluated the described CAD system on real data on and on real and synthesized data.

Fig. 10.5 illustrates the performance of the CAD system in monomodal and multimodal settings with and without synthesized PET data. The synthesized PET data were only introduced in the oc-SVM learning stage. This allows us to evaluate to contribution of the synthesized data in the outlier detection context alone. As can be deduced from figure 10.5, in both monomodal and multimodal CAD systems the synthesized PET data have improved the sensitivity.

Table 10.1 summarizes the performance of the CAD system corresponding to the maximum sensitivity for the experiments above. The positive contribution of the synthesized PET images is evident in both monomodal and multimodal settings. For PET-only CAD system, the maximum sensitivity increases from 2/19 to 8/19 while for multimodal T1/PET CAD system, the sensitivity changes from 4/19 to 7/19, when synthetic PET data is included into the system.

However, in the PET-only CAD setting, the lesions detected with and without synthetic PET data do not coincide. The only 2 patients (J^- and L^-) detected with the system trained on the real data only, are missed when the synthetic data is introduced. This means that the outlier detection models built around a few data points may occasionally do better, by learning tighter boundaries and isolating outliers. When more data points are available, the normality boundary would tend to expand and accommodate more points, occasionally including some outliers. The quantitative results, however, suggest that more representative points result in a more realistic boundary and, thus, more outliers are recognized.

In the scope of this experiment, evaluating the impact of synthetic PET data, we could also compare the monomal PET-only CAD system with the multimodal T1/PET system. Table 10.1 also gives an idea on this aspect. When trained on real data, considering the T1 MRI modality increases the performance from 2/19 to 4/19 maximum sensitivity. However, there is not a single lesion detected in both settings. Same comparison including synthetic PET data actually reveals a slight decrease in sensitivity when T1-weighted MRi is considered as well, by going from 8/19 to 7/19. These results may indicate that the combination of T1-weighted MRI and PET images in a multichannel architecture may not be the best fusion strategy for these two modalities. Indeed, T1-w MRI and PET images have quite different structures. A better fusion strategy should be considered in the future work.

Fig. 10.6 illustrates the obtained normalized score maps output by the corresponding oc-SVM models trained on 1) real PET data, 2) real and synthetic PET data, 3) real T1-w MRI / real PET data and 4) real T1-w MRI / real + synthetic PET data. As it can be seen the real PET-only CAD system results in rather uniformly anomalous output map. This only supports the intuition that a small number of training data points results in a restricted model of normality. When adding synthetic data and, thus, augmenting the number of training points, the output maps change drastically (middle column). Same observation takes place in the multimodal setting (last two columns). Post-processing the maps by applying the routine described in 10.4.2, results in cluster maps illustrated on fig. 10.7. The figure depicts the maximum intensity projections of the found detections onto a transverse slice centered at the lesion. When a lesion is detected, FPs, corresponding to its rank, are shown. When a lesion is not detected, top 10 clusters are shown.

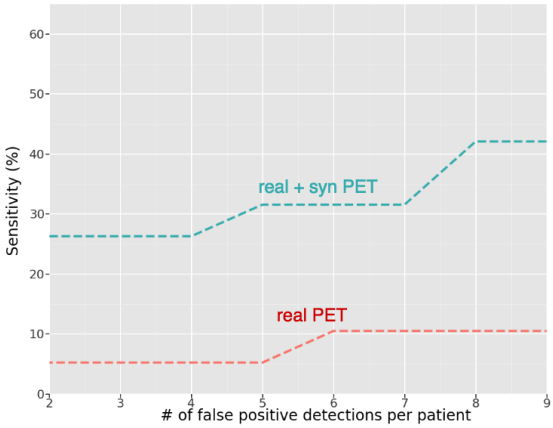
10.6 Conclusion

In the recent works, many strategies have been proposed for cross-modality data synthesis in medical imaging. When coupled with a particular medical task, the recent approaches showed an improvement observed in the systems when synthesized data is exploited versus when it is not. The exploitation of synthesized data in the task of outlier detection remains, however, rather unexplored. In this chapter we made an attempt to leverage a straightforward strategy to generate PET images from T1-weighted MRI. Further, we evaluated the performance of the proposed CAD system when real versus real and synthesized data were used in the outlier detection stage.

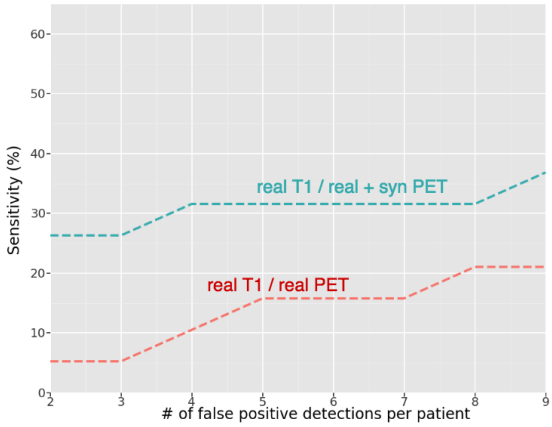
The results obtained in this experiment show a convincing improvement of the CAD system performance when the training data set in the outlier detection stage is enhanced with synthetic data points. This takes place in both PET-only and multimodal T1w / PET CAD systems. Indeed, when adding generated PET data in the PET-only CAD system,

Patient	**Lesion location**	real PET	real + syn. PET	real T1/PET	real T1/ real + syn. PET
Patient A^-	Insula R	✗	✓(9)	✗	✗
Patient B^-	Temporal Lobe L	✗	✓(2)	✓(5)	✗
Patient C^-	Hippocampus R	✗	✗	✗	✗
Patient D^+	Superior frontal gyrus R	✗	✓(2)	✗	✗
Patient E^-	Inferiolateral remainder of parietal lobe R	✗	✗	✗	✗
Patient F^-	Hippocampus L, parahippocampus L	✗	✗	✗	✗
Patient G^-	Middle frontal gyrus L	✗	✓(6)	✗	✓(1)
Patient H^-	Superior frontal gyrus R	✗	✓(1)	✗	✗
Patient I^-	Hippocampus L, parahippocampus L	✗	✗	✗	✗
Patient J^-	Precentral gyrus R	✓(1)	✗	✗	✓(1)
Patient K^-	Superior temporal gyrus R	✗	✗	✗	✗
Patient L^-	Middle frontal gyrus R	✓(7)	✗	✗	✗
Patient M^-	Anterior temporal lobe R	✗	✗	✓(9)	✓(5)
Patient N^-	Anterior temporal lobe R	✗	✓(9)	✗	✗
Patient O^-	Middle frontal gyrus L	✗	✓(2)	✓(2)	✓(1)
Patient P^-	Hippocampus R	✗	✗	✗	✗
Patient Q^-	Lateral remainder of occipital lobe L	✗	✓(2)	✓(4)	✓(1)
Patient R^-	Orbital gyrus R	✗	✗	✗	✓(2)
Patient S^-	Hippocampus R	✗	✗	✗	✓(10)
Overall # of detections		**2**	**8**	**4**	**7**

Table 10.1: Comparative results of different configurations of the CAD system at patient level. For each patient, column 2 reports the lesion location while columns 3 to 6 indicate, for each CAD setting, if the lesion was detected (✓) or missed (✗), as well as the rank of the true detection inside parentheses.



(a) fROC curve of the CAD system in the monomodal setting where the oc-SVM models for outlier detection were trained on real PET data (red) versus real and synthesized PET data (green). The number of data points for each oc-SVM is 35 and 75, without and with synthesized data, respectively.



(b) fROC curve of the CAD system in the multimodal setting where the oc-SVM models for outlier detection were trained on real PET data (red) versus real and synthesized PET data (green). The T1-weighted data is always real. The number of data points for each oc-SVM is 35 and 75, without and with synthesized data, respectively.

Figure 10.5: The contribution of the synthesized PET data in monomodal and multimodal CAD systems. In both cases, the integration of the synthesized PET data improves the sensitivity of the system.

the maximum sensitivity increases from 2/19 to 8/19. This indicates the potential of synthetic data in the considered context of outlier detection. In our experiment, we made only a preliminary attempt to synthesize PET images from MRI. The obtained synthetic PET images, though realistic, could be improved further in terms of quality. To this end, other methods should be considered, such as the CycleGAN architecture which showed promising performance in the state-of-the-art studies on image imputation. Synthetic data of a superior quality could largely improve the performance of our CAD system. Our experiment, however, gives a preliminary idea on the future exploitation of PET data in automated epilepsy lesion detection systems. As we have seen in chapter 3, PET imaging is only occasionally explored in the existing CAD systems.

Another aspect in this experiment is the combination of T1-weighted MRI and PET images into a single framework. As the main objective of the experiments in this chapter was to evaluate the contribution of synthetic PET data for outlier detection, we have made only a limited attempt to combine the modalities as input channels, similarly to the early fusion performed for T1-weighted and FLAIR MRI data in chapter 9. It may not be the optimal strategy for T1-weighted and PET modalities due to significant differences in the information contained in those acquisitions. Further strategies should be explored in the future. Moreover, additional architectures should be considered so as to find a better one serving as a representation learning component. Among the considered choices, Wasserstein autoencoder and the regularized siamese network should be included, following their superior

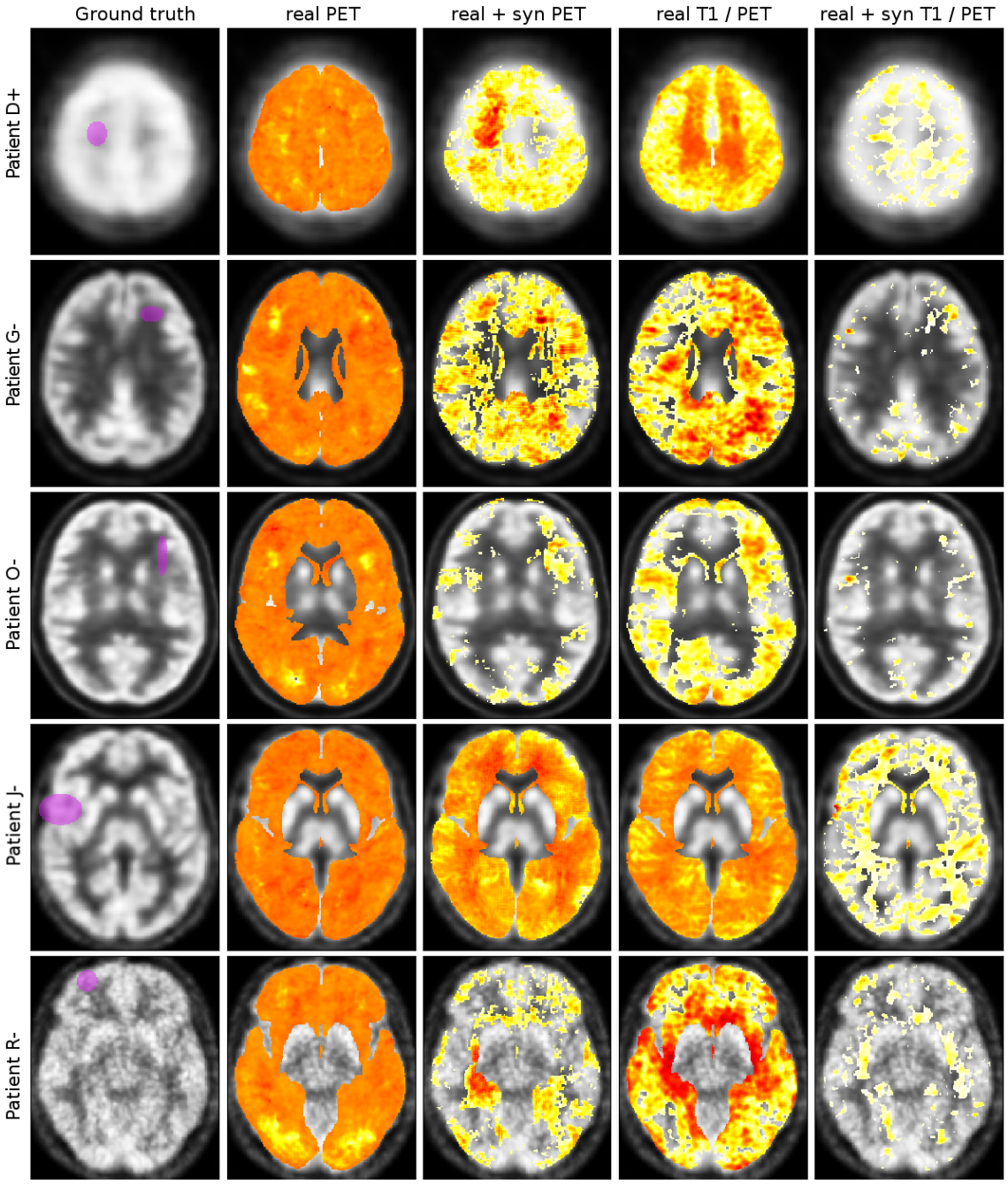


Figure 10.6: Visualization of the CAD detections. First column: original slice centered at the lesion highlighted in a purple circle. From second to last column: the normalized output maps obtained with a CAD system based on real PET data; real + synthetic PET data; real T1 / real PET data; real T1 / real + synthetic PET data. Darker shades correspond to more negative scores, and thus, anomalous regions.

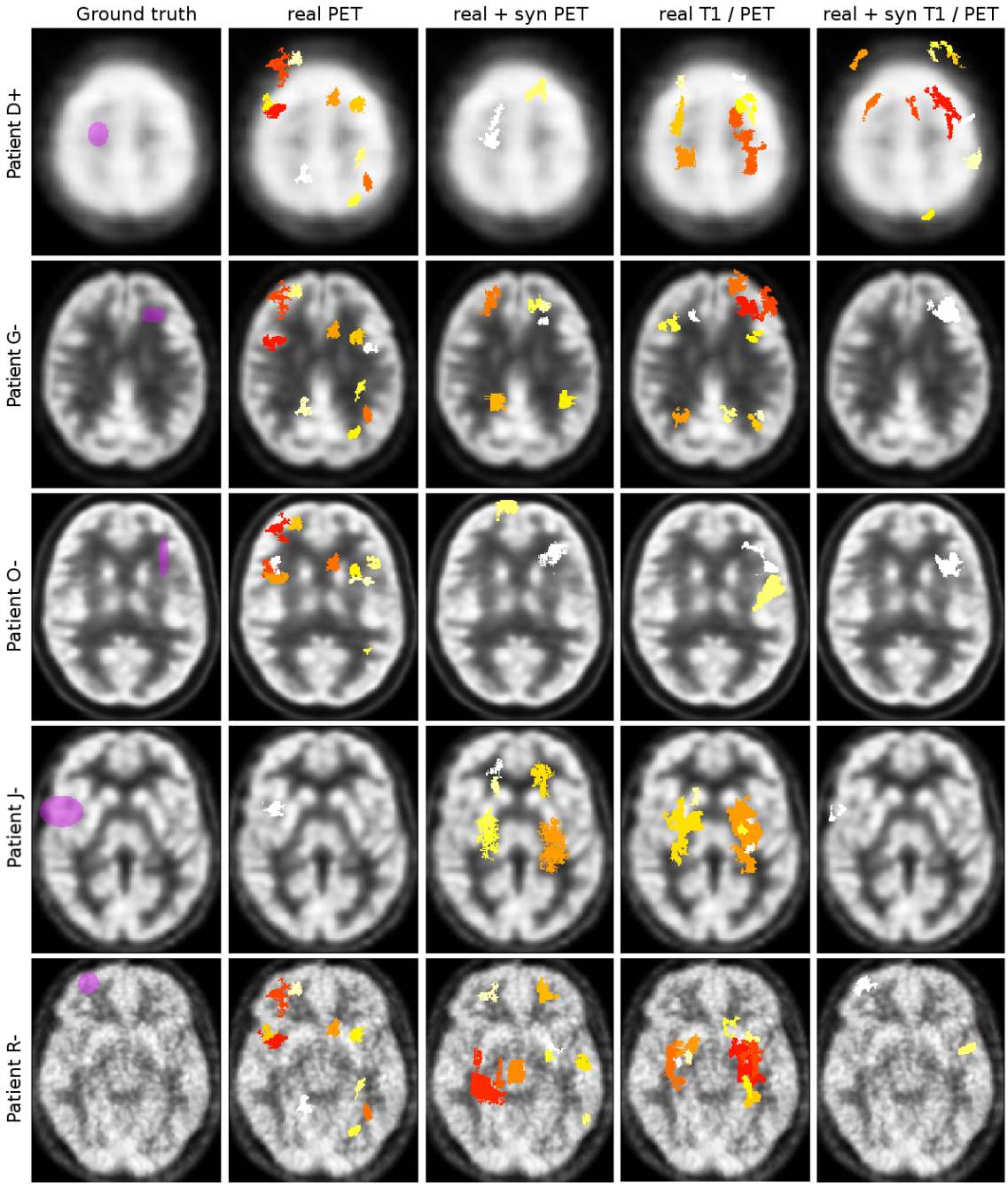


Figure 10.7: Visualization of the CAD detections. The cluster maps show the minimum number of false positive clusters allowing to detect the lesion (in other words, its rank), when it is detected, and top 10 clusters, when it is not. Some clusters' projections may overlap so visually their number might be underestimated. In reality, the clusters are distributed across the 3D brain volume, so the projections may sometimes seem to appear outside the scalp. First column: original slice centered at the lesion highlighted in a purple circle; From second to last column: the cluster maps obtained with a CAD system based on real PET data; real + synthetic PET data; real T1 / real PET data; real T1 / real + synthetic PET data. Note the difference in the number of clusters needed before the true detection emerges.

performance established in the experiments in chapters 7 and 9, corresponding to the T1-w MRI CAD system and the multimodal T1-w/FLAIR MRI CAD system, respectively.

Conclusion and perspectives

This work made an attempt to develop a computer aided diagnosis (CAD) system for automated detection of subtle abnormalities on brain imaging. The clinical application of the proposed CAD system consists in the detection of epilepsy lesions, particularly, in patients considered normal over routine visual examination of the MRI scans.

We started by presenting a comprehensive overview of the existing CAD systems for epilepsy detection, pointing out the main limitations. We gave our considerations and constraints on the problem at hand, with respect to the limited number of data and the lack of accurately annotated lesions available for training, and formalized an entirely unsupervised CAD system. Such a system is based on the concept of per voxel outlier detection, by learning the normality model of each voxel in the brain. The previous strategy proposed by [El Azami et al., 2016] in the scope of this project, followed the tendency of the current CAD systems for epilepsy detection and employed clinically guided features. The main disadvantage of such approaches is that they operate upon handcrafted features, mimicking the current clinical intuition on the appearance of epilepsy lesions. Moreover, such features typically limit the systems to a single epilepsy cause or category (FCD).

Our first contribution in this work was to propose various unsupervised deep architectures that could produce relevant representations, so as to replace the narrow range of handcrafted features with a more optimal and generic one. Eventually, we proposed a novel configuration of siamese networks that seem to be particularly adapted to the context of outlier detection, exploited throughout this work. We evaluated the proposed unsupervised representation learning strategy within the adopted unsupervised CAD system on epilepsy lesion detection on T1-weighted MRI. The considered data set consists of 21 epilepsy patients, with 18 MRI-negatives. We showed the superior performance, achieved with the features learnt with the proposed deep architecture, compared to the same CAD system employing handcrafted features. Moreover, we compared the overall CAD system against the currently common SPM analysis approach, based on per-voxel mass univariate GLM analysis. The comparison revealed the advantages of the proposed CAD system.

The evaluation of the CAD system on T1-weighted MRI resulted in maximum sensitivity between 42-48%, depending on the representation learning model. This gives a room for improvement. It could be achieved by considering additional imaging modalities which offer complementary information on the pathology at hand. Indeed, in pre-surgical evaluation it is common to consult various medical imaging modalities in order to have a comprehensive understanding of the pathology. In the current CAD systems, however, multimodality data has not been explored extensively. Our second contribution, therefore, consists in extending the considered CAD system to accommodate multimodality imaging data, by proposing relevant data fusion strategies. As such, we considered two options. The first one consists in early fusion by combining the available modalities as input channels to the representation learning models. This approach amounts to learning common representations for all input modalities. The second strategy consists in learning

representations per modality individually and later combine the learnt modality-specific representations through the multiple kernel learning paradigm. In this case, a combination of the representations is sough to separate better the normal data points. When comparing these two strategies within the CAD system on T1-weighted / FLAIR MRI, we showed the promising performance of the multiple kernel learning strategy and emphasized the superior sensitivity achieved with the early fusion approach. Eventually, on T1-weighted and FLAIR multimodal analysis, the CAD system achieved around 62% sensitivity.
Our last contribution presents our exploratory efforts towards the integration of PET data for epilepsy detection. PET data has been considered only occasionally in the current epilepsy detection CADs. The clinical studies, however, show that PET imaging improves the epilepsy lesion detection rate over visual inspection. In order to exploit the available PET data within the proposed CAD system, we implemented the CAD system on PET-only data and a combination of T1-weighted MRI and PET imaging. However, the number of healthy subjects who had PET exams is only 35. The proposed CAD system built on such a small number of cases would not perform at its best. We therefore considered first synthesizing PET images from the T1-weighted MRI and then integrating the generated PET data in the outlier detection stage of the CAD. The results, obtained in the scope of this preliminary experiment, clearly show that the synthetic data improves the performance of the CAD system in both PET-only and combined T1-w/PET settings.

Future work

Our main contribution in this work was to propose representation learning strategies, to be coupled with per voxel oc-SVM models within the adopted CAD system framework. The CAD system itself may be improved in the future work with respect to various aspects. The concept of learning a oc-SVM model per voxel may be optimized by considering groups of homogeneous voxels together and building a oc-SVM model per group. This would imply proposing a relevant method of parcelizing homogeneous voxels together. Following this intuition, we have conducted a number of experiments, by employing a few clustering algorithms. We have noticed that the overall sensitivity of the system rather degrades which is likely to be a result of an inappropriate choice of the number of clusters. An exhaustive search over a possible number should be considered which will require an appropriate criterion to be introduced so as to choose the most beneficial parcelization of the voxels. This, eventually, demands additional effort.
The second aspect in the CAD design that could be improved in the post-processing routine. As we have stated throughout this work, the unsupervised representation learning stage produces representations that are not specific or discriminating to any particular pathology. The system, therefore, detects all the abnormalities identified with respect to

the healthy control population used for training. This tends to result in a rather elevated number of false positive detections, when the CAD system is applied to detect a particular pathology. To this end, the post-processing stage could be designed to eliminate the false positives with respect to the particular problem at hand. This will amount to deciding upon relevant criteria to retain some detections and eliminate others. Such rules could also be designed to eliminate false detections stemming from imaging artifacts or registration discrepancies. These rules could be implemented through an additional neural architecture trained to discriminate true and false detections. Eventually, problem-specific supervised data, if available, could be introduced in order to drive an efficient discrimination of true and false positive detections.

The third aspect is the integration of spatial information into the representation learning component. This amounts to incorporating a vector, encoding the spatial localization of patches, into the input to the considered architectures. Following this idea, we have tested a number of possibilities which occasionally resulted in a slightly superior performance. The gain, however, was not significant. It is our understanding that such a strategy requires a careful choice of representation of the spatial information that would actually improve the learnt representations.

Finally, we considered training the CAD system in an end-to-end fashion. One possible strategy we have explored was to integrate the spatial information into the representation learning system and identify outliers as those examples whose representations, conditioned on their spatial localization, differ largely from those of the healthy controls. To this end, for each patch, we considered introducing a vector of its distance to a number of fixed points in the brain volume. This approach, however, resulted in a low sensitivity, by producing patient-control group deviation maps, practically constant everywhere. It is our understanding that such a spatial representation is not an effective information to condition the learnt representations on. For one, the vector is continuous, it depends on the number of fixed reference points, and the number of different conditional vectors is quite large. A different strategy, resulting in conditional vectors with different properties, perhaps, might be more appropriate.

Regarding the multimodality representation learning, implemented through a multichannel neural network in this work, could be performed through a better optimized architecture. We have shown the gain in sensitivity when both T1-weighted and FLAIR MRI modalities were combined. However, in a few cases, the lesions detected with one of the modalities, were missed by their combination. This suggests that a superior sensitivity can be achieved, perhaps with a more appropriate architecture combining the two data sources.

Our last chapter explored the PET imaging in the context of epilepsy detection. We focused on the synthesis of missing PET acquisitions from the corresponding MR images and performed a preliminary evaluation of the CAD performance, with and without synthetic

data. Naturally, other cross-modality image synthesis methods could be applied as presented in section 10.2. In particular, the cycleGAN architecture seems to be a promising option and should be considered in the future.

The next aspect in the scope of this experiment is that, when evaluating the CAD system on the combination of T1-weighted MRI and PET images, we have not considered other combination strategies, including MKL or a deep network, combining the modality-specific information at some later stage. MRI and PET acquisitions, having different appearances, might benefit from a different strategy, taking into account the specifics of both modalities. Eventually, a multimodal MRI-PET CAD system is yet to be studied. Moreover, an optimal combination of T1-weighted, FLAIR MRI and PET imaging should be considered in a CAD system, for a more comprehensive analysis.

Our final observation concerns the application of the overall CAD system in medical applications. As stated when presenting the CAD system, the proposed framework is not tailored to any pathology in particular. It is, therefore, of a great practical interest to apply the proposed CAD to a number of neuropathologies in order to evaluate its potential and limitations. Eventually, introducing the system into an early pre-surgical evaluation phase may prove to be useful to the clinicians. In the scope of our collaboration with J. Jung, we have started to implement the entire pipeline through the VIP platform, and shortly in the future the CAD system will be available in the clinical setting.

Publication list

- Zara Alaverdyan, Julien Jung, Romain Bouet and Carole Lartizien. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. *Medical Image Analysis (MEDIA)* [submitted]

- Zara Alaverdyan, Jiazheng Chai and Carole Lartizien. Unsupervised feature learning for outlier detection with stacked convolutional autoencoders, siamese networks and Wasserstein autoencoders: Application to epilepsy detection. *4th International Workshop on Deep Learning in Medical Image Analysis (DLMIA) held in conjunction with MICCAI, Granada, Spain, 2018.*

- Zara Alaverdyan, Julien Jung, Romain Bouet and Carole Lartizien. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. *International conference on Medical Imaging with Deep Learning (MIDL), Amsterdam, Netherlands, 2018.*

- Zara Alaverdyan and Carole Lartizien. Feature extraction with regularized siamese networks for outlier detection: application to epilepsy lesion detection. *Conférence sur l'apprentissage automatique (CAp 2017), Grenoble, France, 2017.*

Appendix

Chapter A

Alternative input patch size

In section 7.3.1 we discussed the architectures designed as feature extracting components in the proposed CAD for subtle anomaly detection on brain images. We argued that the choice of the input patch size to those architectures should not be very large since representations learnt on large patches may not capture the necessary local characteristics distinguishing subtle abnormalities. To illustrate this point we consider an alternative architecture with a patch size of 31 x 31 at input, illustrated on fig. A.1. The dimension of the representations in the middle layer, when flattened, is the same as before (64). randomly chosen patches, together with their reconstructions obtained with this network, are shown on fig. A.2. As it can be seen the quality of the reconstruction is rather convincing. When exploiting this architecture as a feature extraction mechanism for the pipeline described in chapter 7.1, following the same design choices as in chapter 7, the patient output score maps clearly missed even the most obvious lesions. Fig. A.3 presents a comparison of typical output maps obtained with the system trained on the patch size 31. As it can be noticed, the CAD system with representations learnt on large patches does not succeed at capturing sufficiently well the abnormalities around the true lesion. The scores around the lesion are only slightly, if at all, different from the rest of the image. The showcased example carries a visually remarkable, though subtle, lesion. Other patients in our data set, mostly being *MRI-negative*, have even more subtle lesions which were not recognized by the system trained on large patches. The proposed CAD system with representations learnt on larger patches might be relevant to the contexts where more obvious and large anomalies are sought. Eventually, more elaborate networks may be designed to preserve the local characteristics of the normal subjects. Future work might explore such alternatives.

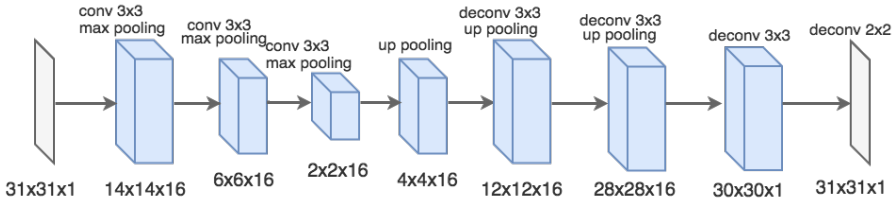


Figure A.1: An experimental architecture for 31 x 31 patches at input.

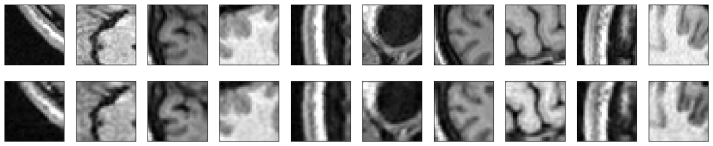


Figure A.2: 10 random 31 x 31 patches (top row) together with their corresponding reconstructions obtained with a convolutional autoencoder with input size 31 x 31.

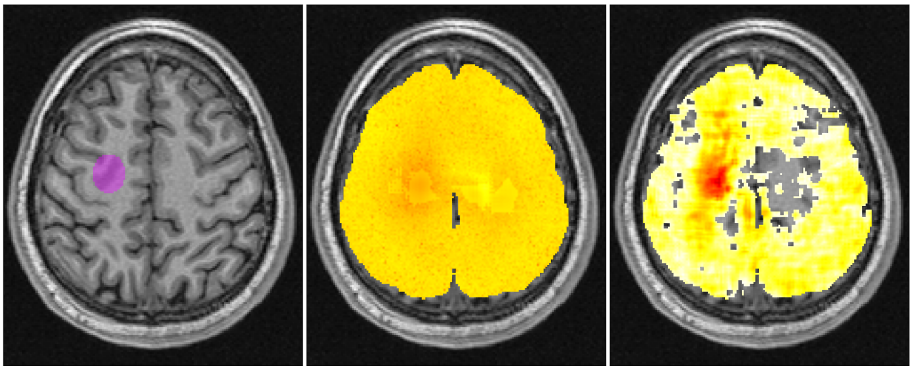


Figure A.3: Comparison of the output maps obtained with the proposed CAD with convolutional autoencoders of different input patch sizes. Left: Original image slice centered at a typical visually remarkable lesion, highlighted in purple. Center: The output map obtained with the system on patch size 31. Right: The output map obtained with the system on patch size 15. Notice the complete failure of the system trained on large patches to capture any anomalousness around the lesion.

Bibliography

[Adler et al., 2017] Adler, S., Wagstyl, K., Gunny, R., Ronan, L., Carmichael, D., Cross, J. H., Fletcher, P. C., and Baldeweg, T. (2017). Novel surface features for automated detection of focal cortical dysplasias in paediatric epilepsy. *NeuroImage: Clinical*, 14:18–27.

[Aggarwal and Yu, 2001] Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. In *ACM Sigmod Record*, volume 30, pages 37–46. ACM.

[Ahmed et al., 2015] Ahmed, B., Brodley, C. E., Blackmon, K. E., Kuzniecky, R., Barash, G., Carlson, C., Quinn, B. T., Doyle, W., French, J., Devinsky, O., et al. (2015). Cortical feature analysis and machine learning improves detection of "mri-negative" focal cortical dysplasia. *Epilepsy & Behavior*, 48:21–28.

[Ahmed et al., 2014] Ahmed, B., Thesen, T., Blackmon, K., Zhao, Y., Devinsky, O., Kuzniecky, R., and Brodley, C. (2014). Hierarchical conditional random fields for outlier detection: an application to detecting epileptogenic cortical malformations. In *International Conference on Machine Learning*, pages 1080–1088.

[Ahmed et al., 2016] Ahmed, B., Thesen, T., Blackmon, K. E., Kuzniekcy, R., Devinsky, O., and Brodley, C. E. (2016). Decrypting cryptogenic epilepsy: semi-supervised hierarchical conditional random fields for detecting cortical lesions in mri-negative patients. *The Journal of Machine Learning Research*, 17(1):3885–3914.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[Alarcon et al., 2006] 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(4):474–480.

[Aleskerov et al., 1997] Aleskerov, E., Freisleben, B., and Rao, B. (1997). Cardwatch: A neural network based database mining system for credit card fraud detection. In *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*, pages 220–226. IEEE.

[Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[An and Cho, 2015] An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center, Tech. Rep.*

[Antel et al., 2003] Antel, S. B., Collins, D. L., Bernasconi, N., Andermann, F., Shinghal, R., Kearney, R. E., Arnold, D. L., and Bernasconi, A. (2003). Automated detection of focal cortical dysplasia lesions using computational models of their mri characteristics and texture analysis. *Neuroimage*, 19(4):1748–1759.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223.

[Ashburner, 2009] Ashburner, J. (2009). Computational anatomy with the spm software. *Magnetic resonance imaging*, 27(8):1163–1174.

[Ashburner and Friston, 2000] Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821.

[Ashburner and Friston, 2005] Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.

[Asman and Landman, 2013] Asman, A. J. and Landman, B. A. (2013). Non-local statistical label fusion for multi-atlas segmentation. *Medical image analysis*, 17(2):194–208.

[Bach et al., 2004] Bach, F. R., Lanckriet, G. R., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM.

[Bagadia et al., 2011] Bagadia, A., Purandare, H., Misra, B. K., and Gupta, S. (2011). Application of magnetic resonance tractography in the perioperative planning of patients with eloquent region intra-axial brain lesions. *Journal of Clinical Neuroscience*, 18(5):633–639.

[Barkovich and Kuzniecky, 1996] Barkovich, A. J. and Kuzniecky, R. I. (1996). Neuroimaging of focal malformations of cortical development. *Journal of Clinical Neurophysiology*, 13(6):481–494.

[Barson et al., 1996] Barson, P., Field, S., Davey, N., McAskie, G., and Frank, R. (1996). The detection of fraud in mobile phone networks. *Neural Network World*, 6(4):477–484.

[Baur et al., 2017] Baur, C., Albarqouni, S., and Navab, N. (2017). Semi-supervised deep learning for fully convolutional networks. In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2017)*, pages 311–319, Cham. Springer International Publishing.

[Bay et al., 2006] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

[Bell et al., 2009] 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(9):2053–2060.

[Ben-Cohen et al., 2018] Ben-Cohen, A., Klang, E., Raskin, S. P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M. M., and Greenspan, H. (2018). Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection. *arXiv preprint arXiv:1802.07846*.

[Ben-Gal, 2005] Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer.

[Bernasconi et al., 2011] Bernasconi, A., Bernasconi, N., Bernhardt, B. C., and Schrader, D. (2011). Advances in mri for'cryptogenic'epilepsies. *Nature reviews neurology*, 7(2):99.

[Bernasconi and Bernasconi, 2015] Bernasconi, N. and Bernasconi, A. (2015). *MRI-negative epilepsy: evaluation and surgical management*, pages 16–27. Cambridge University Press.

[Bertinetto et al., 2016] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer.

[Besson et al., 2008] Besson, P., Bernasconi, N., Colliot, O., Evans, A., and Bernasconi, A. (2008). Surface-based texture and morphological analysis detects subtle cortical dysplasia. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 645–652. Springer.

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

[Bien et al., 2009] Bien, C. G., Szinay, M., Wagner, J., Clusmann, H., Becker, A. J., and Urbach, H. (2009). Characteristics and surgical outcomes of patients with refractory magnetic resonance imaging–negative epilepsies. *Archives of Neurology*, 66(12):1491–1499.

[Blümcke and Spreafico, 2011] Blümcke, I. and Spreafico, R. (2011). An international consensus classification for focal cortical dysplasias. *The Lancet Neurology*, 10(1):26–27.

[Blümcke et al., 2013] Blümcke, I., Thom, M., Aronica, E., Armstrong, D. D., Bartolomei, F., Bernasconi, A., Bernasconi, N., Bien, C. G., Cendes, F., Coras, R., et al. (2013). International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: a task force report from the ilae commission on diagnostic methods. *Epilepsia*, 54(7):1315–1329.

[Blümcke et al., 2011] Blümcke, I., Thom, M., Aronica, E., Armstrong, D. D., Vinters, H. V., Palmini, A., Jacques, T. S., Avanzini, G., Barkovich, A. J., Battaglia, G., et al. (2011). The clinicopathologic spectrum of focal cortical dysplasias: A consensus classification proposed by an ad hoc task force of the ilae diagnostic methods commission 1. *Epilepsia*, 52(1):158–174.

[Bortsova et al., 2017] Bortsova, G., van Tulder, G., Dubost, F., Peng, T., Navab, N., van der Lugt, A., Bos, D., and 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*, pages 356–364. Springer.

[Bos et al., 2014] 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., and Ikram, M. A. (2014). Intracranial carotid artery atherosclerosis and the risk of stroke in whites: the rotterdam study. *JAMA neurology*, 71(4):405–411.

[Bos et al., 2012] 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., and Ikram, M. A. (2012). Atherosclerotic calcification relates to cognitive function and to brain changes on magnetic resonance imaging. *Alzheimer's & Dementia*, 8(5):S104–S111.

[Bounsiar and Madden, 2014] Bounsiar, A. and Madden, M. G. (2014). Kernels for one-class support vector machines. In *Information Science and Applications (ICISA), 2014 International Conference on*, pages 1–4. IEEE.

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

[Brauckhoff et al., 2009] Brauckhoff, D., Salamatian, K., and May, M. (2009). Applying pca for traffic anomaly detection: Problems and solutions. In *INFOCOM 2009, IEEE*, pages 2866–2870. IEEE.

[Breiman, 2017] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[Breiman and Spector, 1992] Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, pages 291–319.

[Breunig et al., 2000] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM.

[Bromley et al., 1993] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *IJPRAI*, 7(4):669–688.

[Brosch et al., 2016] Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., and 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(5):1229–1239.

[Bruggemann et al., 2007] Bruggemann, J. M., Wilke, M., Som, S. S., Bye, A. M., Bleasel, A., and 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 Res*, 77(2-3):93–101.

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

[Cai et al., 2016] Cai, Y., Landis, M., Laidley, D. T., Kornecki, A., Lum, A., and Li, S. (2016). Multi-modal vertebrae recognition using transformed deep convolution network. *Computerized Medical Imaging and Graphics*, 51:11–19.

[Candès and Recht, 2009] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

[Cantor-Rivera et al., 2015] Cantor-Rivera, D., Khan, A. R., Goubran, M., Mirsattari, S. M., and Peters, T. M. (2015). Detection of temporal lobe epilepsy using support vector machines in multi-parametric quantitative MR imaging. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 41:14–28.

[Caputo et al., 2002] Caputo, B., Sim, K., Furesjo, F., and 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*, volume 2002.

[Carne et al., 2004] Carne, R., O'brien, T., Kilpatrick, C., MacGregor, L., Hicks, R., Murphy, M., Bowden, S., Kaye, A., and Cook, M. (2004). Mri-negative pet-positive temporal lobe epilepsy: a distinct surgically remediable syndrome. *Brain*, 127(10):2276–2285.

[Celebi et al., 2013] Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, 40(1):200–210.

[Chandola et al., 2009] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

[Chartsias et al., 2017] Chartsias, A., Joyce, T., Dharmakumar, R., and Tsaftaris, S. A. (2017). Adversarial image synthesis for unpaired multi-modal cardiac data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 3–13. Springer.

[Chassoux et al., 2010] Chassoux, F., Rodrigo, S., Semah, F., Beuvon, F., Landre, E., Devaux, B., Turak, B., Mellerio, C., Meder, J.-F., Roux, F.-X., et al. (2010). Fdg-pet improves surgical outcome in negative mri taylor-type focal cortical dysplasias. *Neurology*, 75(24):2168–2175.

[Chen and Salman, 2011] Chen, K. and Salman, A. (2011). Extracting speaker-specific information with a regularized siamese deep network. In *Advances in Neural Information Processing Systems*, pages 298–306.

[Chen et al., 2008] Chen, Q., Lui, S., Li, C.-X., Jiang, L.-J., Ou-Yang, L., Tang, H.-H., Shang, H.-F., Huang, X.-Q., Gong, Q.-Y., and Zhou, D. (2008). Mri-negative refractory partial epilepsy: role for diffusion tensor imaging in high field mri. *Epilepsy research*, 80(1):83–89.

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

[Chickering, 2002] Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

[Choi et al., 2005] Choi, S. W., Lee, C., Lee, J.-M., Park, J. H., and Lee, I.-B. (2005). Fault detection and identification of nonlinear processes based on kernel pca. *Chemometrics and intelligent laboratory systems*, 75(1):55–67.

[Chopra et al., 2005] Chopra, S., Hadsell, R., and 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*, volume 1, pages 539–546. IEEE.

[Çiçek et al., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer.

[Colliot et al., 2006] Colliot, O., Bernasconi, N., Khalili, N., Antel, S. B., Naessens, V., and Bernasconi, A. (2006). Individual voxel-based analysis of gray matter in focal cortical dysplasia. *Neuroimage*, 29(1):162–71.

[Colombo et al., 2012] Colombo, N., Tassi, L., Deleo, F., Citterio, A., Bramerio, M., Mai, R., Sartori, I., Cardinale, F., Russo, G. L., and Spreafico, R. (2012). Focal cortical dysplasia type iia and iib: Mri aspects in 118 cases proven by histopathology. *Neuroradiology*, 54(10):1065–1077.

[Concha et al., 2012] Concha, L., Kim, H., Bernasconi, A., Bernhardt, B. C., and Bernasconi, N. (2012). Spatial patterns of water diffusion along white matter tracts in temporal lobe epilepsy. *Neurology*, 79(5):455–462.

[Cover and Hart, 1967] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

[Cui et al., 2014] Cui, Z., Chang, H., Shan, S., Zhong, B., and Chen, X. (2014). Deep network cascade for image super-resolution. In *European Conference on Computer Vision*, pages 49–64. Springer.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

[Dale et al., 1999] Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194.

[Das et al., 2016] Das, A., Yenala, H., Chinnakotla, M., and Shrivastava, M. (2016). Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 378–387.

[de Vos et al., 2016] de Vos, B. D., Wolterink, J. M., de Jong, P. A., Viergever, M. A., and Išgum, I. (2016). 2d image classification for 3d anatomy localization: employing deep convolutional neural networks. In *Medical Imaging 2016: Image Processing*, volume 9784, page 97841Y. International Society for Optics and Photonics.

[Deng et al., 2010] Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A.-r., and Hinton, G. (2010). Binary coding of speech spectrograms using a deep auto-encoder. In *Eleventh Annual Conference of the International Speech Communication Association*.

[Dheeba et al., 2014] Dheeba, J., Singh, N. A., and Selvi, S. T. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of biomedical informatics*, 49:45–52.

[Diaz and Hollmén, 2002] Diaz, I. and Hollmén, J. (2002). Residual generation and visualization for understanding novel process conditions. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, volume 3, pages 2070–2075. IEEE.

[Doi, 2007] Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211.

[Domingos, 2000] Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238.

[Donahue et al., 2016] Donahue, J., Krähenbühl, P., and Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

[Dong et al., 2014] Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer.

[Dou et al., 2017] Dou, Q., Chen, H., Yu, L., Qin, J., and Heng, P.-A. (2017). Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567.

[Dou et al., 2016] Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V. C., Shi, L., and Heng, P.-A. (2016). Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195.

[Dubost et al., 2017] Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W. J., Vernooij, M., and 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., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention (MICCAI 2017)*, pages 214–221, Cham. Springer International Publishing.

[Duchesne et al., 2006a] Duchesne, S., Bernasconi, N., Bernasconi, A., and Collins, D. (2006a). Mr-based neurological disease classification methodology: Application to lateralization of seizure focus in temporal lobe epilepsy. *Neuroimage*, 29(2):557–566.

[Duchesne et al., 2006b] Duchesne, S., Bernasconi, N., Bernasconi, A., and Collins, D. L. (2006b). MR-based neurological disease classification methodology: application to lateralization of seizure focus in temporal lobe epilepsy. *Neuroimage*, 29(2):557–66.

[Dumoulin et al., 2016] Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. (2016). Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

[Duncan et al., 2016] Duncan, J. S., Winston, G. P., Koepp, M. J., and Ourselin, S. (2016). Brain imaging in the assessment for epilepsy surgery. *The Lancet Neurology*, 15(4):420–433.

[Efron and Tibshirani, 1997] Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

[El Azami et al., 2016] El Azami, M., Hammers, A., Jung, J., Costes, N., Bouet, R., and Lartizien, C. (2016). Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem. *PloS one*, 11(9):e0161498.

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

[Fan et al., 2005] Fan, R.-E., Chen, P.-H., and Lin, C.-J. (2005). Working set selection using second order information for training support vector machines. *Journal of machine learning research*, 6(Dec):1889–1918.

[Filippi et al., 2016] 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., and Barkhof, F. (2016). Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. *Lancet Neurol*, 15(3):292–303.

[Fisher et al., 2014] Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., Engel Jr, J., Forsgren, L., French, J. A., Glynn, M., et al. (2014). Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482.

[Focke et al., 2008] Focke, N. K., Symms, M. R., Burdett, J. L., and Duncan, J. S. (2008). Voxel-based analysis of whole brain flair at 3t detects focal cortical dysplasia. *Epilepsia*, 49(5):786–793.

[Focke et al., 2012] Focke, N. K., Yogarajah, M., Symms, M. R., Gruber, O., Paulus, W., and Duncan, J. S. (2012). Automated mr image classification in temporal lobe epilepsy. *Neuroimage*, 59(1):356–362.

[Fonseca et al., 2012] Fonseca, V. C., Yasuda, C. L., Tedeschi, G. G., Betting, L. E., and Cendes, F. (2012). White matter abnormalities in patients with focal cortical dysplasia revealed by diffusion tensor imaging analysis in a voxelwise approach. *Frontiers in neurology*, 3:121.

[Fotin et al., 2016] Fotin, S. V., Yin, Y., Haldankar, H., Hoffmeister, J. W., and Periaswamy, S. (2016). Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97850X. International Society for Optics and Photonics.

[Fukushima and Miyake, 1982] Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

[Gehring et al., 2013] Gehring, J., Miao, Y., Metze, F., and Waibel, A. (2013). Extracting deep bottleneck features using stacked auto-encoders. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3377–3381. IEEE.

[Ghafoorian et al., 2017a] Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., Keizer, K., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., et al. (2017a). Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, 14:391–399.

[Ghafoorian et al., 2017b] Ghafoorian, M., Karssemeijer, N., Heskes, T., Uden, I. W., Sanchez, C. I., Litjens, G., Leeuw, F.-E., Ginneken, B., Marchiori, E., and Platel, B. (2017b). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):5110.

[Ghafoorian et al., 2016] Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uder, I., de Leeuw, F.-E., Marchiori, E., van Ginneken, B., and Platel, B. (2016). Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1414–1417. IEEE.

[Gill et al., 2017] Gill, R. S., Hong, S.-J., Fadaie, F., Caldairou, B., Bernhardt, B., Bernasconi, N., and 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*, pages 349–356. Springer.

[Gill et al., 2018] 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., et al. (2018). Deep convolutional networks for automated detection of epileptogenic brain malformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 490–497. Springer.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

[Goetz et al., 2014] Goetz, M., Weber, C., Bloecher, J., Stieltjes, B., Meinzer, H.-P., and Maier-Hein, K. (2014). Extremely randomized trees based brain tumor segmentation. *Proceeding of BRATS challenge-MICCAI*, pages 006–011.

[Gönen and Alpaydın, 2011] Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[Grubbs, 1969] Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

[Guerrini et al., 2003] Guerrini, R., Sicca, F., and Parmeggiani, L. (2003). Epilepsy and malformations of the cerebral cortex. *Epileptic Disorders*, 5(2):9–26.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

[Hadsell et al., 2006] Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE.

[Hammers, 2015] Hammers, A. (2015). Pet in mri-negative refractory focal epilepsy. *MRI-negative epilepsy: evaluation and surgical management*, page 28.

[Hammers et al., 2003] Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., and Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human brain mapping*, 19(4):224–247.

[Han et al., 2015] Han, X., Leung, T., Jia, Y., Sukthankar, R., and Berg, A. C. (2015). Matchnet: Unifying feature and metric learning for patch-based matching. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3279–3286. IEEE.

[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., et al. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning new york. *NY: Springer*.

[Hastie et al., 1999] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., and Botstein, D. (1999). Imputing missing data for gene expression arrays.

[Havaei et al., 2017] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31.

[Havaei et al., 2016] Havaei, M., Guizard, N., Chapados, N., and Bengio, Y. (2016). Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer.

[Hawkins, 1980] Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

[He et al., 2006] He, Z., Deng, S., Xu, X., and Huang, J. Z. (2006). A fast greedy algorithm for outlier mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 567–576. Springer.

[Heckemann et al., 2006] Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126.

[Hinrichs et al., 2011] Hinrichs, C., Singh, V., Xu, G., Johnson, S. C., Initiative, A. D. N., et al. (2011). Predictive markers for ad in a multi-modality framework: an analysis of mci progression in the adni population. *Neuroimage*, 55(2):574–589.

[Hinton and Zemel, 1994] Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10.

[Hirose et al., 2017] Hirose, N., Sadeghian, A., Goebel, P., and Savarese, S. (2017). To go or not to go? a near unsupervised learning approach for robot navigation. *arXiv preprint arXiv:1709.05439*.

[Hodge and Austin, 2004] Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126.

[Hong et al., 2014] Hong, S.-J., Kim, H., Schrader, D., Bernasconi, N., Bernhardt, B. C., and Bernasconi, A. (2014). Automated detection of cortical dysplasia type ii in mri-negative epilepsy. *Neurology*, 83(1):48–55.

[Hua et al., 2015] Hua, K.-L., Hsu, C.-H., Hidayati, S. C., Cheng, W.-H., and Chen, Y.-J. (2015). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, 8.

[Huafu and Hai, 2004] Huafu, C. and Hai, N. (2004). Detection the character wave in epileptic eeg by wavelet. *Journal of Electronic Science and Technology*, 2(1):69–71.

[Huang et al., 2007] Huang, L., Nguyen, X., Garofalakis, M., Jordan, M. I., Joseph, A., and Taft, N. (2007). In-network pca and anomaly detection. In *Advances in Neural Information Processing Systems*, pages 617–624.

[Huang et al., 2018] Huang, S., Huang, D., and Zhou, X. (2018). Learning multimodal deep representations for crowd anomaly event detection. *Mathematical Problems in Engineering*, 2018.

[Huo et al., 2018] Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R. G., and Landman, B. A. (2018). Adversarial synthesis learning enables segmentation without target modality ground truth. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1217–1220. IEEE.

[Huppertz et al., 2005] Huppertz, H.-J., Grimm, C., Fauser, S., Kassubek, J., Mader, I., Hochmuth, A., Spreer, J., and 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(1-2):35–50.

[Huppertz et al., 2011] Huppertz, H.-J., Wagner, J., Weber, B., House, P., and Urbach, H. (2011). Automated quantitative flair analysis in hippocampal sclerosis. *Epilepsy research*, 97(1-2):146–156.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[Isgum et al., 2009] Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M. A., and Van Ginneken, B. (2009). Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in ct scans. *IEEE transactions on medical imaging*, 28(7):1000–1010.

[Jain et al., 1999] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

[Jakoby et al., 2011] Jakoby, B., Bercier, Y., Conti, M., Casey, M., Bendriem, B., and Townsend, D. (2011). Physical and clinical performance of the mct time-of-flight pet/ct scanner. *Physics in Medicine & Biology*, 56(8):2375.

[Jensen, 1996] Jensen, F. V. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London.

[Jiang et al., 2018] Jiang, J., Hu, Y.-C., Tyagi, N., Zhang, P., Rimner, A., Mageras, G. S., Deasy, J. O., and Veeraraghavan, H. (2018). Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 777–785. Springer.

[Jin et al., 2018] Jin, B., Krishnan, B., Adler, S., Wagstyl, K., Hu, W., Jones, S., Najm, I., Alexopoulos, A., Zhang, K., Zhang, J., Ding, M., Wang, S., and Wang, Z. I. (2018). Automated detection of focal cortical dysplasia type ii with surface-based magnetic resonance imaging postprocessing and machine learning. *Epilepsia*, 59(5):982–992.

[Johnson and Wichern, 1992] Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*. Prentice Hall.

[Jolliffe, 2011] Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer.

[Juhász et al., 2000] Juhász, C., Chugani, D., Muzik, O., Watson, C., Shah, J., Shah, A., and Chugani, H. (2000). Electroclinical correlates of flumazenil and fluorodeoxyglucose pet abnormalities in lesional epilepsy. *Neurology*, 55(6):825–835.

[Kabat and Król, 2012] Kabat, J. and Król, P. (2012). Focal cortical dysplasia–review. *Polish journal of radiology*, 77(2):35.

[Kabir et al., 2007] Kabir, Y., Dojat, M., Scherrer, B., Forbes, F., and Garbay, C. (2007). Multimodal mri segmentation of ischemic stroke lesions. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 1595–1598. IEEE.

[Kamnitsas et al., 2016] Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., Rueckert, D., and Glocker, B. (2016). Deepmedic for brain tumor segmentation. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 138–149. Springer.

[Kamnitsas et al., 2017] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78.

[Karpathy et al., 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

[Keihaninejad et al., 2012] Keihaninejad, S., Heckemann, R. A., Gousias, I. S., Hajnal, J. V., Duncan, J. S., Aljabar, P., Rueckert, D., and Hammers, A. (2012). Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic mri segmentation. *PloS one*, 7(4):e33096.

[Keller et al., 2007] Keller, S. S., Cresswell, P., Denby, C., Wieshmann, U., Eldridge, P., Baker, G., and Roberts, N. (2007). Persistent seizures following left temporal lobe surgery are associated with posterior and bilateral structural and functional brain abnormalities. *Epilepsy research*, 74(2-3):131–139.

[Kent, 1983] Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173.

[Kim et al., 2011] Kim, Y. H., Kang, H.-C., Kim, D.-S., Kim, S. H., Shim, K.-W., Kim, H. D., and Lee, J. S. (2011). Neuroimaging in identifying focal cortical dysplasia and prognostic factors in pediatric and adolescent epilepsy surgery. *Epilepsia*, 52(4):722–727.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kini et al., 2016] Kini, L. G., Gee, J. C., and Litt, B. (2016). Computational analysis in epilepsy neuroimaging: a survey of features and methods. *NeuroImage: Clinical*, 11:515–529.

[Kiran et al., 2018] Kiran, B. R., Thomas, D. M., and Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2):36.

[Kisilev et al., 2016] Kisilev, P., Sason, E., Barkan, E., and Hashoul, S. (2016). Medical image description using multi-task-loss cnn. In *Deep Learning and Data Labeling for Medical Applications*, pages 121–129. Springer.

[Kleesiek et al., 2014] Kleesiek, J., Biller, A., Urban, G., Kothe, U., Bendszus, M., and Hamprecht, F. (2014). Ilastik for multi-modal brain tumor segmentation. *Proceedings MICCAI BraTS (Brain Tumor Segmentation Challenge)*, pages 12–17.

[Koch et al., 2015] Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2.

[Koepp and Woermann, 2005] Koepp, M. J. and Woermann, F. G. (2005). Imaging structure and function in refractory focal epilepsy. *The Lancet Neurology*, 4(1):42–53.

[Kong et al., 2016] Kong, B., Zhan, Y., Shin, M., Denny, T., and Zhang, S. (2016). Recognizing end-diastole and end-systole frames via deep temporal regression network. In *International conference on medical image computing and computer-assisted intervention*, pages 264–272. Springer.

[Kooi et al., 2017] Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., and Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312.

[Kotsiantis et al., 2007] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24.

[Kriegel et al., 2009] Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009). Loop: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1649–1652. ACM.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

[Kumar et al., 2015] Kumar, D., Wong, A., and Clausi, D. A. (2015). Lung nodule classification using deep features in ct images. In *Computer and Robot Vision (CRV), 2015 12th Conference on*, pages 133–138. IEEE.

[Kumar et al., 2010] Kumar, S. P., Sriraam, N., Benakop, P., and Jinaga, B. (2010). Entropies based detection of epileptic seizures with artificial neural network classifiers. *Expert Systems with Applications*, 37(4):3284–3291.

[Kwan and Brodie, 2000] Kwan, P. and Brodie, M. J. (2000). Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319.

[Lamusuo et al., 2001] Lamusuo, S., Jutila, L., Ylinen, A., Kälviäinen, R., Mervaala, E., Haaparanta, M., Jääskeläinen, S., Partanen, K., Vapalahti, M., and Rinne, J. (2001). [18f] fdg-pet reveals temporal hypometabolism in patients with temporal lobe epilepsy even when quantitative mri and histopathological analysis show only mild hippocampal damage. *Archives of neurology*, 58(6):933–939.

[Laurikkala et al., 2000] Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., and Kavsek, B. (2000). Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, volume 1, pages 20–24.

[Lazarevic et al., 2003] Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., and Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 25–36. SIAM.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[Ledig et al., 2017] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4.

[Lee and Park, 2008] Lee, J.-S. and Park, C. H. (2008). Adaptive decision fusion for audio-visual speech recognition. In *Speech Recognition*. InTech.

[Lee et al., 2004] Lee, S.-K., Kim, D. I., Mori, S., Kim, J., Kim, H. D., Heo, K., and Lee, B. I. (2004). Diffusion tensor mri visualizes decreased subcortical fiber connectivity in focal cortical dysplasia. *Neuroimage*, 22(4):1826–1829.

[Lerner et al., 2009] Lerner, J. T., Salamon, N., Hauptman, J. S., Velasco, T. R., Hemb, M., Wu, J. Y., Sankar, R., Donald Shields, W., Engel Jr, J., Fried, I., et al. (2009). Assessment and surgical outcomes for mild type i and severe type ii cortical dysplasia: a critical review and the ucla experience. *Epilepsia*, 50(6):1310–1335.

[Li et al., 2014] Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., and Ji, S. (2014). Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–312. Springer.

[Litjens et al., 2014] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H. (2014). Computer-aided detection of prostate cancer in mri. *IEEE transactions on medical imaging*, 33(5):1083–1092.

[Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

[Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

[Loosli and Aboubacar, 2017] Loosli, G. and Aboubacar, H. (2017). Using svdd in simplemkl for 3d-shapes filtering. *arXiv preprint arXiv:1712.02658*.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee.

[Lu et al., 2013] Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440.

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Makhzani et al., 2015] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

[Manjunath and Ma, 1996] Manjunath, B. S. and Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):837–842.

[Mao et al., 2016] Mao, X., Shen, C., and Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810.

[Masci et al., 2011] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer.

[Mazziotta et al., 2001] 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(1412):1293–1322.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

[McLachlan and Krishnan, 2007] McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

[Menze et al., 2015] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993.

[Metz, 1986] Metz, C. E. (1986). Roc methodology in radiologic imaging. *Investigative radiology*, 21(9):720–733.

[Mourão-Miranda et al., 2011] Mourão-Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., Williams, S. C., Shawe-Taylor, J., and Brammer, M. (2011). Patient classification as an outlier detection problem: an application of the one-class support vector machine. *Neuroimage*, 58(3):793–804.

[Munawar et al., 2017a] Munawar, A., Vinayavekhin, P., and De Magistris, G. (2017a). Limiting the reconstruction capability of generative neural network using negative learning. *arXiv preprint arXiv:1708.08985*.

[Munawar et al., 2017b] Munawar, A., Vinayavekhin, P., and 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*, pages 1017–1025. IEEE.

[Murtagh and Contreras, 2011] Murtagh, F. and Contreras, P. (2011). Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121*.

[Murthy, 1998] Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, 2(4):345–389.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[Niaf et al., 2014] Niaf, E., Flamary, R., Rakotomamonjy, A., Rouviere, O., and Lartizien, C. (2014). Svm with feature selection and smooth prediction in images: Application to cad of prostate cancer. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 2246–2250. IEEE.

[Niaf et al., 2012] Niaf, E., Rouvière, O., Mège-Lechevallier, F., Bratan, F., and Lartizien, C. (2012). Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric mri. *Physics in Medicine & Biology*, 57(12):3833.

[Nie et al., 2017] Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., and Shen, D. (2017). Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer.

[Noble et al., 2004] Noble, W. S. et al. (2004). Support vector machine applications in computational biology. *Kernel methods in computational biology*, 71:92.

[Orbes-Arteaga et al., 2018] Orbes-Arteaga, M., Cardoso, M. J., Sørensen, L., Modat, M., Ourselin, S., Nielsen, M., and Pai, A. (2018). Simultaneous synthesis of flair and segmentation of white matter hypointensities from t1 mris. *arXiv preprint arXiv:1808.06519*.

[Orosco et al., 2013] Orosco, L., Correa, A. G., and Laciar, E. (2013). a survey of performance and techniques for automatic epilepsy detection. *Journal of Medical and Biological Engineering*, 33(6):526–537.

[Pan et al., 2018] Pan, Y., Liu, M., Lian, C., Zhou, T., Xia, Y., and Shen, D. (2018). Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimer's disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 455–463. Springer.

[Park et al., 2004] Park, S. H., Goo, J. M., and Jo, C.-H. (2004). Receiver operating characteristic (roc) curve: practical review for radiologists. *Korean Journal of Radiology*, 5(1):11–18.

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *BMVC*, volume 1, page 6.

[Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

[Pavlidis et al., 2002] Pavlidis, P., Weston, J., Cai, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *Journal of computational biology*, 9:401–411.

[Payer et al., 2016] Payer, C., Štern, D., Bischof, H., and Urschler, M. (2016). Regressing heatmaps for multiple landmark localization using cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–238. Springer.

[Pedregosa et al., 2011] 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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pereira et al., 2015] Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2015). Deep convolutional neural networks for the segmentation of gliomas in multi-sequence mri. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 131–143. Springer.

[Petrick et al., 2013] Petrick, N., Sahiner, B., Armato, S. G., Bert, A., Correale, L., Delsanto, S., Freedman, M. T., Fryd, D., Gur, D., Hadjiiski, L., et al. (2013). Evaluation of computer-aided detection and diagnosis systems. *Medical physics*, 40(8).

[Pimentel et al., 2014] Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

[Poria et al., 2016] Poria, S., Cambria, E., Howard, N., Huang, G.-B., and Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

[Poudel et al., 2016] Poudel, R. P., Lamata, P., and Montana, G. (2016). Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 83–94. Springer.

[Pustina et al., 2015] Pustina, D., Avants, B., Sperling, M., Gorniak, R., He, X., Doucet, G., Barnett, P., Mintzer, S., Sharan, A., and Tracy, J. (2015). Predicting the laterality of temporal lobe epilepsy from pet, mri, and dti: A multimodal study. *NeuroImage: clinical*, 9:20–31.

[Radford et al., 2015] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[Rajavel and Sathidevi, 2015] Rajavel, R. and Sathidevi, P. (2015). Optimum integration weight for decision fusion audio–visual speech recognition. *International Journal of Computational Science and Engineering*, 10(1-2):145–154.

[Rakotomamonjy et al., 2008] Rakotomamonjy, A., Bach, F. R., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research*, 9(Nov):2491–2521.

[Rathore et al., 2014] Rathore, C., Dickson, J. C., Teotónio, R., Ell, P., and Duncan, J. S. (2014). The utility of 18f-fluorodeoxyglucose pet (fdg pet) in epilepsy surgery. *Epilepsy research*, 108(8):1306–1314.

[Reed et al., 2016] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

[Riney et al., 2012] Riney, C. J., Chong, W. K., Clark, C. A., and Cross, J. H. (2012). Voxel based morphometry of flair mri in children with intractable focal epilepsy: implications for surgical intervention. *European journal of radiology*, 81(6):1299–1305.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

[Rosenow and Lüders, 2001] Rosenow, F. and Lüders, H. (2001). Presurgical evaluation of epilepsy. *Brain*, 124(9):1683–1700.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Sabuncu et al., 2010] Sabuncu, M. R., Yeo, B. T., Van Leemput, K., Fischl, B., and Golland, P. (2010). A generative model for image segmentation based on label fusion. *IEEE transactions on medical imaging*, 29(10):1714–1729.

[Saini and Dutta, 2017] Saini, J. and Dutta, M. (2017). An extensive review on development of eeg-based computer-aided diagnosis systems for epilepsy detection. *Network: Computation in Neural Systems*, 28(1):1–27.

[Salamon et al., 2008] Salamon, N., Kung, J., Shaw, S., Koo, J., Koh, S., Wu, J., Lerner, J., Sankar, R., Shields, W., Engel, J., et al. (2008). Fdg-pet/mri coregistration improves detection of cortical dysplasia in patients with epilepsy. *Neurology*, 71(20):1594–1601.

[Salimans et al., 2016] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

[Schlegl et al., 2017] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. Information Processing in Medical Imaging, pages 146–157. Springer International Publishing.

[Schneider, 2001] Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, 14(5):853–871.

[Schölkopf et al., 2001] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

[Schölkopf et al., 1997] Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer.

[Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Seeböck et al., 2016] Seeböck, P., Waldstein, S., Klimscha, S., Gerendas, B. S., Donner, R., Schlegl, T., Schmidt-Erfurth, U., and Langs, G. (2016). Identifying and categorizing anomalies in retinal imaging data. *arXiv preprint arXiv:1612.00686*.

[Shah et al., 2018] Shah, M. P., Merchant, S., and Awate, S. P. (2018). Abnormality detection using deep neural networks with robust quasi-norm autoencoding and semi-supervised learning. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 568–572. IEEE.

[Shen et al., 2017] Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248.

[Shen et al., 2015] Shen, W., Zhou, M., Yang, F., Yang, C., and Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer.

[Simo-Serra et al., 2015] Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., and Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 118–126. IEEE.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Sonnenburg et al., 2006a] Sonnenburg, S., Rätsch, G., and Schäfer, C. (2006a). A general and efficient multiple kernel learning algorithm. In *Advances in neural information processing systems*, pages 1273–1280.

[Sonnenburg et al., 2006b] Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006b). Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(Jul):1531–1565.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[Srivastava et al., 2005] Srivastava, S., Maes, F., Vandermeulen, D., Van Paesschen, W., Dupont, P., and Suetens, P. (2005). Feature-based statistical analysis of structural MR data for automatic detection of focal cortical dysplastic lesions. *Neuroimage*, 27(2):253–66.

[Strumia et al., 2012] Strumia, M., Ramantani, G., Mader, I., Henning, J., Bai, L., and Hadjidemetriou, S. (2012). Analysis of structural mri data for the localisation of focal cortical dysplasia in epilepsy. In *Workshop on Clinical Image-Based Procedures*, pages 25–32. Springer.

[Subasi et al., 2005] Subasi, A., Kiymik, M. K., Alkan, A., and Koklukaya, E. (2005). Neural network classification of eeg signals by using ar with mle preprocessing for epileptic seizure detection. *Mathematical and Computational Applications*, 10(1):57–70.

[Suk et al., 2014] Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N., et al. (2014). Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582.

[Suk and Shen, 2013] Suk, H.-I. and Shen, D. (2013). Deep learning-based feature representation for ad/mci classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 583–590. Springer.

[Sun, 2013] Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038.

[Szegedy et al., 2017] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

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

[Tax and Duin, 2004] Tax, D. M. and Duin, R. P. (2004). Support vector data description. *Machine learning*, 54(1):45–66.

[Téllez-Zenteno et al., 2010] Téllez-Zenteno, J. F., Ronquillo, L. H., Moien-Afshari, F., and Wiebe, S. (2010). Surgical outcomes in lesional and non-lesional epilepsy: a systematic review and meta-analysis. *Epilepsy research*, 89(2-3):310–318.

[Thesen et al., 2011] 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(2):e16430.

[Thivard et al., 2006] Thivard, L., Adam, C., Hasboun, D., Clémenceau, S., Dezamis, E., Lehéricy, S., Dormont, D., Chiras, J., Baulac, M., and Dupont, S. (2006). Interictal diffusion MRI in partial epilepsies explored with intracerebral electrodes. *Brain : a journal of neurology*, 129(Pt 2):375–85.

[Thivard et al., 2011] Thivard, L., Bouilleret, V., Chassoux, F., Adam, C., Dormont, D., Baulac, M., Semah, F., and Dupont, S. (2011). Diffusion tensor imaging can localize the epileptogenic zone in nonlesional extra-temporal refractory epilepsies when [18F]FDG-PET is not contributive. *Epilepsy Research*, 97(1-2):170–182.

[Thung et al., 2017] Thung, K.-H., Yap, P.-T., and Shen, D. (2017). Multi-stage diagnosis of alzheimer's disease with incomplete multimodal data via multi-task deep learning. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 160–168. Springer.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

[Tolstikhin et al., 2017] Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.

[Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

[Tsamardinos et al., 2006] Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

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

[van Ginneken et al., 2011] van Ginneken, B., Schaefer-Prokop, C. M., and Prokop, M. (2011). Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–732.

[Vapnik, 1995] Vapnik, V. (1995). The nature of statistical learning. *Theory*.

[Vapnik, 1999] Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

[Varma and Simon, 2006] Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

[Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

[Wagner et al., 2011] Wagner, J., Weber, B., Urbach, H., Elger, C. E., and Huppertz, H.-J. (2011). Morphometric mri analysis improves detection of focal cortical dysplasia type ii. *Brain*, 134(10):2844–2854.

[Wang et al., 2016] Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

[Wang et al., 2015] Wang, F., Kang, L., and Li, Y. (2015). Sketch-based 3d shape retrieval using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1875–1883. IEEE.

[Wang et al., 2013] Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623.

[Ward Jr, 1963] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[Wardlaw, 2008] Wardlaw, J. M. (2008). What is a lacune?

[Wardlaw et al., 2013] 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., and Dichgans, M. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol*, 12(8):822–38.

[Wiebe et al., 2001] Wiebe, S., Blume, W. T., Girvin, J. P., and Eliasziw, M. (2001). A randomized, controlled trial of surgery for temporal-lobe epilepsy. *New England Journal of Medicine*, 345(5):311–318.

[Willmann et al., 2007] Willmann, O., Wennberg, R., May, T., Woermann, F., and Pohlmann-Eden, B. (2007). The contribution of 18f-fdg pet in preoperative epilepsy surgery evaluation for patients with temporal lobe epilepsy: a meta-analysis. *Seizure*, 16(6):509–520.

[Wolterink et al., 2017] Wolterink, J. M., Dinkla, A. M., Savenije, M. H., Seevinck, P. R., van den Berg, C. A., and Išgum, I. (2017). Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer.

[Wu et al., 2016] Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90.

[Xiang et al., 2018] Xiang, L., Li, Y., Lin, W., Wang, Q., and Shen, D. (2018). Unpaired deep cross-modality synthesis with fast training. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 155–164. Springer.

[Xie et al., 2012] Xie, J., Xu, L., and Chen, E. (2012). Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349.

[Xing et al., 2016] Xing, C., Ma, L., and Yang, X. (2016). Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. *Journal of Sensors*, 2016.

[Xu et al., 2013] Xu, C., Tao, D., and Xu, C. (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

[Xu et al., 2015] Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*.

[Xu et al., 2018] Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., et al. (2018). Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 187–196. International World Wide Web Conferences Steering Committee.

[Xu and C. Wunsch II, 2005] Xu, R. and C. Wunsch II, D. (2005). Survey of clustering algorithms. 16:645 – 678.

[Yang et al., 2011] Yang, C.-A., Kaveh, M., and Erickson, B. J. (2011). Automated detection of focal cortical dysplasia lesions on t1-weighted mri using volume-based distributional features. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 865–870. IEEE.

[Yang et al., 2018] Yang, H., Sun, J., Carass, A., Zhao, C., Lee, J., Xu, Z., and Prince, J. (2018). Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 174–182. Springer.

[Yih et al., 2011] Yih, W.-t., Toutanova, K., Platt, J. C., and Meek, C. (2011). Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256. Association for Computational Linguistics.

[Yu et al., 2017] Yu, L., Yang, X., Chen, H., Qin, J., and Heng, P.-A. (2017). Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, pages 66–72.

[Yuan et al., 2012] Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., Initiative, A. D. N., et al. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632.

[Zagoruyko and Komodakis, 2015] Zagoruyko, S. and Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4353–4361. IEEE.

[Zandi et al., 2010] Zandi, A. S., Javidan, M., Dumont, G. A., and Tafreshi, R. (2010). Automated real-time epileptic seizure detection in scalp eeg recordings using an algorithm based on wavelet packet transform. *IEEE Transactions on Biomedical Engineering*, 57(7):1639–1651.

[Zeghidour et al., 2016] Zeghidour, N., Synnaeve, G., Usunier, N., and Dupoux, E. (2016). Joint learning of speaker and phonetic similarities with siamese networks. In *INTERSPEECH*, pages 1295–1299.

[Zenati et al., 2018] Zenati, H., Foo, C. S., Lecouat, B., Manek, G., and Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection. In *ICLR workshop*.

[Zeng et al., 2017] Zeng, K., Yu, J., Wang, R., Li, C., and Tao, D. (2017). Coupled deep autoencoder for single image super-resolution. *IEEE transactions on cybernetics*, 47(1):27–37.

[Zhang et al., 2011] Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A. D. N., et al. (2011). Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867.

[Zhang et al., 2018] Zhang, Z., Yang, L., and Zheng, Y. (2018). Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251.

[Zhao et al., 2017] Zhao, J., Xie, X., Xu, X., and Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54.

[Zheng et al., 2016] Zheng, L., Duffner, S., Idrissi, K., Garcia, C., and Baskurt, A. (2016). Siamese multi-layer perceptrons for dimensionality reduction and face identification. *Multimedia Tools and Applications*, 75(9):5055–5073.

[Zhu et al., 2017a] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*.

[Zhu et al., 2017b] Zhu, Q., Du, B., Turkbey, B., Choyke, P. L., and Yan, P. (2017b). Deeply-supervised cnn for prostate segmentation. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 178–184. IEEE.

[Zhu et al., 2017c] Zhu, X., Thung, K.-H., Adeli, E., Zhang, Y., and Shen, D. (2017c). Maximum mean discrepancy based multiple kernel learning for incomplete multimodality neuroimaging data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–80. Springer.

[Zikic et al., 2014] Zikic, D., Ioannou, Y., Brown, M., and Criminisi, A. (2014). Segmentation of brain tumor tissues with convolutional neural networks. *Proceedings MICCAI-BRATS*, pages 36–39.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : ALAVERDYAN DATE de SOUTENANCE : 18/01/2019

Prénoms : Zaruhi

TITRE : Unsupervised representation learning for anomaly detection on neuroimaging. Application to epilepsy lesion detection on brain MRI

NATURE : Doctorat Numéro d'ordre : 2019LYSEI005

Ecole doctorale : Electronique, Electrotechnique, Automatique (EEA)

Spécialité :Traitement du signal et de l'image

RESUME :

One third of epilepsy patients are diagnosed with medically refractory epilepsy where seizures cannot be controlled by pharmacotherapy. For such patients, surgical resection of the epileptogenic zone may offer a seizure-free life. The success of such surgeries largely depends on the accuracy of the epileptogenic zone localization. Neuroimaging, including magnetic resonance imaging (MRI) and positron emission tomography (PET), has been increasingly considered in the pre-surgical examination routine.
This work represents one attempt to develop a computer aided diagnosis system for epileptogenic lesion detection based on multimodal neuroimaging data. The adopted approach, first introduced in Azami et al., 2016, consists in casting the lesion detection task as a per-voxel outlier detection problem, based on training a one-class SVM model for each voxel in the brain on a set of healthy controls.
The main focus of this work is to design representation learning mechanisms, capturing the most discriminant information from multimodality imaging. Manual features might not be the most optimal ones for the task at hand. Our first contribution consists in proposing various unsupervised neural architectures as potential feature extracting mechanisms and, eventually, introducing a novel configuration of siamese networks, to be plugged into the outlier detection context. The proposed system, evaluated on a set of T1-weighted MRI of epilepsy patients, showed a promising performance but a room for improvement as well. To this end, we considered extending the CAD system so as to accommodate multimodality data which offers complementary information on the problem at hand. Our second contribution, therefore, consists in proposing strategies to combine representations of different imaging modalities into a single framework for anomaly detection. The extended system showed a significant improvement on the task of epilepsy lesion detection on T1-weighted and FLAIR images. Our last contribution focuses on the integration of PET data into the system. Given the small number of PET images available, we propose to synthetize PET data from the corresponding MRI acquisitions and show an improved performance of the system when synthesized images are used in addition to the real ones.

MOTS-CLÉS : unsupervised representation learning, siamese networks, outlier detection, autoencoders, one class SVM, computer aided diagnosis, epilepsy

Laboratoire (s) de recherche : Centre de recherche en acqusition et traitement de l'image pour la santé

Directeur de thèse: Carole LARTIZIEN

Président de jury :

Composition du jury :

Cardoso, Jorge M.	Professeur des Universités	King's College London	Rapporteur
Mateus, Diana	Professeur des Universités	Ecole Centrale Nantes	Rapporteur
Bonnet-Loosli, Gaëlle	Maître de Conférences	Université Clermont Auvergne	Examinatrice
Fromont, Elisa	Professeur des Universités	Université de Rennes 1	Examinatrice
Jung, Julien	Praticien Hospitalier	CHU de Lyon	Examinateur
Lartizien, Carole	Directrice de recherche	CNRS	Directrice de thèse