N°d'ordre NNT :

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 23/09/2016, par :
MERIEM EL AZAMI

Computer aided diagnosis of epilepsy lesions based on multivariate and multimodality data analysis

---

Devant le jury composé de :

Rakotomamonjy, Alain	Professeur des universités	Université de Rouen	Rapporteur
Rueckert, Daniel	Professeur des universités	Imperial College London	Rapporteur
Canu, Stéphane	Professeur des universités	INSA-ROUEN	Examinateur
Hammers, Alexander	Professeur des universités	King's College London	Examinateur
Friboulet, Denis	Professeur des universités	INSA-LYON	Directeur de thèse
Lartizien, Carole	Chargé de recherche	CNRS	Co-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 Bat Blaise Pascal 3e etage 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 69626 Villeurbanne cedex directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec : M.C. HAVGOUDOUKIAN Ecole-Doctorale.eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60.97 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Sec : Safia AIT CHALAL Bat Darwin - UCB Lyon 1 04.72.43.28.91 Insa : H. CHARLES Safia.ait-chalal@univ-lyon1.fr	Mme Gudrun BORNETTE CNRS UMR 5023 LEHNA Université Claude Bernard Lyon 1 Bât Forel 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cédex Tél : 06.07.53.89.13 e2m2@ univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://www.ediss-lyon.fr** Sec : Safia AIT CHALAL Hôpital Louis Pradel - Bron 04 72 68 49 09 Insa : M. LAGARDE Safia.ait-chalal@univ-lyon1.fr	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 avenue Jean Capelle INSA de Lyon 696621 Villeurbanne Tél : 04.72.68.49.09 Fax :04 72 68 49 16 Emmanuelle.canet@univ-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** **http://infomaths.univ-lyon1.fr** Sec :Renée EL MELHEM Bat Blaise Pascal 3e etage infomaths@univ-lyon1.fr	Mme Sylvie CALABRETTO LIRIS – INSA de Lyon Bat Blaise Pascal 7 avenue Jean Capelle 69622 VILLEURBANNE Cedex Tél : 04.72. 43. 80. 46 Fax 04 72 43 16 87 Sylvie.calabretto@insa-lyon.fr
Matériaux	**MATERIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Saint Exupéry Ed.materiaux@insa-lyon.fr	M. Jean-Yves BUFFIERE INSA de Lyon MATEIS Bâtiment Saint Exupéry 7 avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72.43 71.70 Fax 04 72 43 85 28 Ed.materiaux@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** **http://mega.universite-lyon.fr** Sec : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Saint Exupéry mega@insa-lyon.fr	M. Philippe BOISSE INSA de Lyon Laboratoire LAMCOS Bâtiment Jacquard 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72 .43.71.70 Fax : 04 72 43 72 37 Philippe.boisse@insa-lyon.fr
ScSo	**ScSo*** **http://recherche.univ-lyon2.fr/scso/** Sec : Viviane POLSINELLI Brigitte DUBOIS Insa : J.Y. TOUSSAINT viviane.polsinelli@univ-lyon2.fr	Mme Isabelle VON BUELTZINGLOEWEN Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Tél : 04.78.77.23.86 Fax : 04.37.28.04.48

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

Abstract

One third of patients suffering from epilepsy are resistant to medication. For these patients, surgical removal of the epileptogenic zone offers the possibility of a cure. Surgery success relies heavily on the accurate localization of the epileptogenic zone. The analysis of neuroimaging data such as magnetic resonance imaging (MRI) and positron emission tomography (PET) is increasingly used in the pre-surgical work-up of patients and may offer an alternative to the invasive reference of Stereo-electro-encephalography (SEEG) monitoring. To assist clinicians in screening these lesions, we developed a computer aided diagnosis system (CAD) based on a multivariate data analysis approach. Our first contribution was to formulate the problem of epileptogenic lesion detection as an outlier detection problem. The main motivation for this formulation was to avoid the dependence on labelled data and the class imbalance inherent to this detection task. The proposed system builds upon the one class support vector machines (OC-SVM) classifier. OC-SVM was trained using features extracted from MRI scans of healthy control subjects, allowing a voxelwise assessment of the deviation of a test subject pattern from the learned patterns. System performance was evaluated using realistic simulations of challenging detection tasks as well as clinical data of patients with intractable epilepsy. The outlier detection framework was further extended to take into account the specificities of neuroimaging data and the detection task at hand. Three original contributions in the domain of outlier detection algorithms were proposed and evaluated based on referenced databases (UCI databases) and on the MRI epilepsy database, when available. First, to handle the presence of noise in the training data, we proposed a reformulation of the support vector data description (SVDD) method, a variant of the OC-SVM method, based on the l_0-pseudo-norm. We demonstrated that the resulting l_0-SVDD problem can be solved using an iterative procedure providing data specific weighting terms. Second, to deal with the multi-parametric nature of the neuroimaging data, an optimal late fusion strategy for combining multiple base one-class classifiers was investigated. The late fusion approach consisted in building local models associated each with a single MR sequence (DTI, Flair..) and then combining their output based on an original score combination. This approach was shown to outperform standard early fusion approach based on a single global model learned using features extracted from all MR sequences. Finally, to help with score interpretation, threshold selection and score combination, we proposed to transform the score outputs of the outlier detection algorithm into well calibrated probabilities. A two steps strategy was proposed. We first generalized the SVDD method by reformulating the associated problem to estimate nested probability level-sets and then used a calibration function to convert the outputted scores into well-calibrated probability estimates. Two calibration functions were considered: the sigmoid function classically used in binary classification problems, and the generalized extreme value distribution, much more suited for long-tailed probability distributions that can be encountered in the context of outlier detection.

Résumé étendu

Environ 150.000 personnes souffrent en France d'une épilepsie partielle réfractaire à tous les médicaments. La chirurgie, qui constitue aujourd'hui le meilleur recours thérapeutique 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 (fluorodéoxyglucose) 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.

Pour assister les cliniciens dans leur tâche diagnostique, nous avons développé un système d'aide au diagnostic (CAD) reposant sur l'analyse multivariée de données d'imagerie. Compte tenu de la difficulté relative à la constitution de bases de données annotées et équilibrées entre classes, notre première contribution a été de placer l'étude dans le cadre méthodologique de la détection du changement. L'algorithme du séparateur à vaste marge adapté à ce cadre-là (OC-SVM) a été utilisé pour apprendre, à partir de cartes multi-paramétriques extraites d'IRM T1 de sujets normaux, un modèle prédictif caractérisant la normalité à l'échelle du voxel. Le modèle OC-SVM a été entraîné en utilisant des caractéristiques extraites des images IRM de témoins sains. Le modèle permet ensuite de faire ressortir, dans les images de patients, les zones cérébrales suspectes s'écartant de la normalité. Les performances de ce premier système CAD ont été évaluées sur des lésions simulées ainsi que sur une base de données de patients. Pour prendre en compte la nature complexe des données d'imagerie ainsi que celle de la tâche de détection de lésions épileptogènes, nous avons proposé d'étendre le schéma initialement proposé. Trois contributions dans le contexte de détection du changement ont été apportées et leurs performances ont été évaluées en utilisant des bases de données référencées (base de données UCI) ainsi que des données d'imagerie lorsque cela était possible.

Tout d'abord, un nouveau schéma de détection plus robuste à la présence de bruit d'étiquetage dans la base de données d'apprentissage a été proposé. Ce schéma constitue une reformulation de l'algorithme SVDD, une variante de l'algorithme OC-SVM. Dans cette reformulation, une pénalité l_0 a été utilisée dans le terme d'attache aux données de la fonction coût à minimiser. Nous avons démontré que le nouveau problème d'optimisation L_0-SVDD peut être résolu de façon itérative, permettant d'obtenir un terme de pondération associé à chaque donnée d'apprentissage. Ensuite, nous avons proposé une stratégie de fusion de données qui permet d'intégrer des données d'imagerie multi-modales ou multi-séquences. Deux types de fusion ont été considérés : fusion précoce et fusion tardive. D'une part, la fusion précoce correspond à la fusion au niveau des caractéristiques qui revient à construire un modèle prédictif global à partir de toutes les caractéristiques extraites des données multi-séquences. D'autre part, la fusion tardive correspond à construire plusieurs modèles prédictifs et de fusionner leurs sorties. Dans le contexte de détection de l'épilepsie, nous avons proposé de combiner plusieurs classifieurs OC-SVM associés chacun à une séquence IRM (T1, FLAIR et DTI). Enfin, pour permettre une meilleure interprétation de la sortie du système CAD et son intégration dans le bilan pré-opératoire global, nous avons proposé une méthode de conversion de la sortie du CAD en probabilités. Cette méthode repose sur deux étapes. Dans une première étape, une généralisation de l'algorithme SVDD a été proposée pour l'estimation de courbes de niveaux de probabilités dotées d'une structure hiérarchique. Ensuite, la deuxième étape consiste à convertir les scores en sortie de cette généralisation en probabilités. Deux fonctions de calibration ont été considérées : la sigmoïde utilisée classiquement dans le contexte de classification binaire, et la loi d'extremum généralisée qui convient mieux à l'estimation des distributions avec longue traîne, fréquentes dans le contexte de détection de changement.

Synthèse par chapitre

Introduction générale

L'épilepsie est une maladie neurologique pouvant atteindre le système nerveux. Elle se caractérise par des crises imprédictibles causées par des décharges brusques dans l'activité électrique (influx nerveux) du cerveau. Dans la majorité des cas, les causes de ces crises sont inconnues. On compte, environ, 65 millions de personnes souffrants d'épilepsie dans le monde. Dans 70% des cas d'épilepsie, les médicaments antiépileptiques peuvent être utilisés afin de contrôler les crises [Nagae *et al.* (2016)]. Le reste des situations correspond à des patients qui ne répondent pas à ces traitements. On parle, alors, d'épilepsie réfractaire. L'ablation chirurgicale de la région épileptogène constitue un espoir de guérison pour les personnes atteintes d'épilepsie réfractaire. Cependant, l'issue de la chirurgie dépend essentiellement de la capacité des médecins praticiens à repérer précisément la zone épileptogène. Le bilan pré-chirurgical préparatoire à l'ablation chirurgicale nécessite l'analyse des électroencéphalogrammes (EEG) et des données obtenues à partir des techniques de neuroimagerie. Ceci a pour but de produire des localisations approximatives des régions suspectes d'être épileptogènes. Enfin, l'électroencéphalographie intracrânienne (SEEG) est utilisée pour confirmer la localisation des zones épileptogènes [Duncan *et al.* (2016)]. Grâce aux avancées récentes des techniques de neuroimagerie telles que l'imagerie par résonance magnétique (IRM) et la tomographie par émission de positons (PET), l'analyse des données obtenues à partir de ces techniques joue un rôle de plus en plus important dans le protocole du bilan pré-chirurgical, et aide à la restriction des zones qui seront étudiées par la SEEG, examen très invasif.

Les méthodes de reconnaissance de formes ont été utilisées, avec succès, pour aider les médecins praticiens à détecter les anomalies et améliorer le diagnostic de différentes maladies à partir de données de neuroimagerie [Norman *et al.* (2006), Klöppel *et al.* (2008), Orrù *et al.* (2012), Gray *et al.* (2013)]. Notamment, les techniques de diagnostic assisté par ordinateur (CAD) ont été largement utilisées durant ces dernières années, afin d'aider à la détection des lésions épileptogènes. Un intérêt particulier a été porté à l'utilisation des techniques CAD pour les données IRM [Antel *et al.* (2003), Duchesne *et al.* (2006), Keihaninejad *et al.* (2012), Hong *et al.* (2014), Ahmed *et al.* (2015)].

Contributions

Dans cette thèse, nous avons conçu un système CAD basé sur une analyse multivariée et multimodale. Le cadre que nous avons proposé permet d'extraire les caractéristiques discriminatoires à partir des données de neuroimagerie (différentes modalités et/ou séquences). Cette étape intervient, habituellement, dans le bilan préchirurgical des patients souffrant d'une épilepsie réfractaire. Pour chaque patient, le système CAD proposé produit une carte de groupements de voxels (cluster) étiquetés montrant les zones qui présentent des comportements qu'on pourrait associer à des zones épileptogènes. La technique proposée a été évaluée sur des données cliniques qui nous ont été fournies par l'Hôpital Neurologique de Lyon via nos collaborateurs Pr. F. Mauguière, Dr. J. Jung et Dr. A. Hammers.

La conception d'un tel système CAD nécessite la manipulation de données de neuroimagerie qui présente divers défis. En particulier, les données acquises sont bruitées, de grande dimension et viennent de différentes modalités. Aussi, il faut noter le manque de données de patients annotées et le déséquilibre de classes qu'on retrouve dans la tâche

de détection. L'une des contributions principales de notre travail est la prise en compte des spécificités des données de neuroimagerie et le développement de moyens efficaces pour faire face aux défis que nous aurons identifiés à l'avance. Notre première contribution était de formuler le problème de la détection des lésions épileptogènes comme un problème de détection de changement afin d'éviter la dépendance aux données des patients étiquetés et le déséquilibre des classes qui est inhérent au problème de détection que nous considérons. Le système CAD conçu est basé sur une extension du classifieur Séparateur à Vaste Marge (SVM) adapté au cas d'une seule classe -one-class (OC-SVM). Le classifieur OC-SVM a été appris à l'échelle du voxel afin de contourner la grande dimensionnalité des données et permettre une localisation précise des zones épileptogènes. Une extension du cadre a été proposée dans le but de contourner les autres défis. Ainsi, le bruit présent sur les données a été contourné grâce à une reformulation de l'algorithme support vector data description (SVDD). Cette reformulation considère un coût l_0 (coût 0-1) au lieu du coût Hinge qui est utilisé dans la formulation initiale. Afin de minimiser la fonction objectif qui résulte de la nouvelle formulation L_0-SVDD, une procédure itérative, permettant de calculer des poids de façon adaptative dans la fonctionnelle à minimiser, était introduite. Pour faire face à la nature multimodale des données de neuroimagerie, une stratégie de fusion optimale est recherchée. Deux stratégies de fusion sont considérées : une première stratégie, précoce, consistant à construire un seul modèle prédictif et une deuxième, tardive, qui consiste à apprendre des classifieurs OC-SVM associés à chacune des séquences IRM disponibles et de combiner ensuite leurs sorties. Enfin, pour aider à l'interprétation des scores obtenus nous avons proposé une technique qui permet de convertir les résultats de l'algorithme de SVDD en probabilités calibrées. À cet effet, deux généralisations de l'algorithme de SVDD et deux fonctions de calibration sont proposés.

Organisation du manuscrit

Le manuscrit est composé de trois parties. Dans la première partie, nous commençons par donner une description de l'épilepsie réfractaire et du protocole du bilan pré-chirurgical dans le Chap. 1. Nous donnons ensuite, dans le Chap. 2, une vue d'ensemble sur la structure, les choix et les étapes de conception d'un système CAD. Dans ce contexte, nous passons en revue l'état de l'art des systèmes CAD qui ont été proposés dans le cadre de l'épilepsie réfractaire. Nous concluons la partie I par une analyse approfondie de la tâche de diagnostic étant donnés les entrées disponibles et les résultats souhaités. Dans la deuxième partie, nous commençons d'abord par décrire, dans Chap. 4, les algorithmes OC-SVM et SVDD, qui sont les deux algorithmes d'apprentissage au cœur du système CAD proposé. Nous donnons ensuite au Chap. 5 une description détaillée d'un premier système CAD ainsi qu'une évaluation de ses performances à la fois sur des simulations réalistes et des données cliniques. Dans la troisième partie, nous proposons trois extensions du système décrit dans la partie II afin de faire face aux autres défis. Nous proposons tout d'abord au Chap. 6 une extension de la méthodologie SVDD au cas d'observations incertaines. Ceci permet de gérer la présence du bruit lié à l'étiquetage. Nous étudions ensuite au Chap. 7 une stratégie optimale de fusion pour combiner plusieurs séquences IRM. Enfin, le Chap. 8, est dédié à une méthode de conversion des résultats du système CAD proposé en probabilités calibrées. Enfin, des conclusions générales et des perspectives de ce travail sont présentées à la fin du manuscrit.

Chapitre 1 : Épilepsie pharmaco-résistante

Description

L'épilepsie est une maladie neurologique qui touche environ 1% de la population mondiale [Nagae *et al.* (2016)]. Ses manifestations sont variées et complexes. Elle implique l'ensemble du cerveau et peut aboutir à des états aussi bien bénins que très sévères. Malgré cette diversité, les maladies épileptiques ont toutes un point commun : la survenue de crises à caractère soudain, repérées par les données cliniques et électroencéphalographiques.

La recherche médicale a permis le développement d'une pharmacopée permettant de contrôler les crises pour environ 70% des cas. Les autres cas correspondent à ce qu'on appelle les épilepsies pharmaco-résistantes. L'épilepsie du lobe temporal et les dysplasies corticales focales (FCD), un sous-type de malformations corticales du développement, sont les causes les plus fréquentes de ce type d'épilepsie [Taylor *et al.* (1971)]. Les FCD représentent environ 25% des cas d'épilepsie focale et il s'agit essentiellement d'anomalies corticales. Si au cours du développement, la migration des neurones à partir des zones sous-ventriculaires jusqu'au cortex est perturbée d'autres types d'anomalies peuvent apparaître. Par exemple, les hétérotopies focales sont caractérisées par la présence de neurones en dehors du cortex et dispersés dans la substance blanche.

La détection de ces malformations corticales du développement est complexe car celles-ci présentent une grande hétérogénéité en termes de type de malformations (trois sous types pour les FCDs, les hétérotopies focales, hétérotopies sous forme de bandes . . .), de localisation dans le cerveau, et de taille de lésion.

Diagnostic

Pour les patients souffrant d'épilepsie pharmaco-résistante, une exérèse chirurgicale des zones épileptogènes (ZEs) peut conduire à une libération des crises [Fauser *et al.* (2004), Krsek *et al.* (2009), Lerner *et al.* (2009)]. L'étape cruciale pour cette approche est évidemment la localisation précise des zones épileptogènes au cours d'un bilan pré-chirurgical. En effet cela permet de planifier l'intervention chirurgicale et garantir que celle-ci n'introduira pas de déficits fonctionnels ou cognitifs handicapants [Krsek *et al.* (2009), Téllez-Zenteno *et al.* (2010)].

Ce bilan est très complexe et peut être divisé en deux phases [Duncan *et al.* (2016)] :

- la phase I consiste en l'analyse d'enregistrements vidéo-EEG (électroencéphalographie) et des examens de neuroimagerie tels que l'IRM (imagerie par résonance magnétique) structurelle et la TEP (tomographie par émission de positons).

- la phase II consiste en l'analyse d'enregistrements stéréo-EEG. Durant cet examen des électrodes EEG sont implantées en profondeur afin d'atteindre depuis la surface du crâne les régions suspectes repérées lors des précédents examens (vidéo-EEG, IRM, TEP, ...).

L'analyse croisée de tous ces examens par des experts permet la localisation de la ZE. À la suite de cela, l'équipe multidisciplinaire élabore un plan de chirurgie et décide si l'opération est envisageable compte tenu de la localisation de la ZE.

Neuroimagerie de l'épilepsie L'imagerie cérébrale (dite aussi neuroimagerie) désigne l'ensemble des techniques issues de l'imagerie médicale qui permettent d'observer le cerveau.

- La neuroimagerie fonctionnelle cherche à caractériser le cerveau en action. Une utilisation usuelle de ces méthodes consiste à faire effectuer une tâche cognitive à un individu et à mesurer le signal produit par l'activité cérébrale. Suivant les techniques et les outils mathématiques employés, il est possible de retrouver, avec plus ou moins de précision, quelle région du cerveau était particulièrement active et à quel moment de la tâche cognitive. L'imagerie par résonance magnétique fonctionnelle (IRMf), la tomographie par émission de positons (TEP), l'électroencéphalographie (EEG) et la magnétoencéphalographie (MEG) sont les outils principaux pour obtenir une image fonctionnelle du cerveau.
- La neuroimagerie structurelle permet d'identifier, localiser et mesurer les différentes parties de l'anatomie du système nerveux central. Dans la pratique médicale clinique, elle permet d'identifier, de localiser et caractériser (par son extension par exemple) une lésion cérébrale. Cette caractérisation peut avoir une visée diagnostique. Elle peut aussi permettre la programmation d'une intervention chirurgicale. L'IRM anatomique est l'outil principal pour obtenir une image structurelle du cerveau.

EEG et vidéo-EEG L'électroencéphalographie (EEG) est une méthode qui permet l'enregistrement de l'activité électrique du cerveau. L'EEG est une méthode non invasive qui repose sur le placement d'un certain nombre d'électrodes. Ces électrodes sont placées à la surface du crâne loin des sources électriques. Par conséquent, les enregistrements EEG ont une faible résolution spatiale (5 à 9 cm). Les enregistrements vidéo-EEG permettent en même temps de suivre le comportement du patient à la survenue des crises et d'enregistrer l'activité électrique du cerveau.

En analysant les données EEG et vidéo-EEG, les épileptologues sont capables d'identifier les crises qui sont en lien avec l'épilepsie, de caractériser l'activité électrique du cerveau, de déterminer s'il s'agit de crises focales ou généralisées et de donner une localisation approximative de la zone épileptogène. Les résultats de cette analyse sont en général confrontés aux données d'imagerie qui elles offrent une meilleure résolution spatiale.

Imagerie par résonance magnétique (IRM) Le principe de l'IRM anatomique consiste à réaliser des images du cerveau grâce aux atomes d'hydrogène qu'il contient. Placés dans un puissant champ magnétique externe B, les atomes d'hydrogène, qui possèdent un moment dipolaire magnétique dû au spin d'un nucléon non apparié, s'orientent dans la même direction. Une faible magnétisation M apparaît dans la direction de B. Les atomes sont alors excités par des ondes radio (RF, de 10Hz à 100 MHz) durant une très courte période afin que la magnétisation M tourne de 90 degrés ou 180 degrés (ils sont mis en résonance). A l'arrêt de l'impulsion RF le système retourne à l'équilibre selon une cinétique de diffusion longitudinale et transversale. Différentes séquences IRM sont alors définies selon les paramètres définissant l'impulsion ainsi que les temps de répétition de celle-ci. Un protocole optimisé pour la détection des lésions épileptogènes a été proposé par la ligue mondiale de lutte contre l'épilepsie (http://www.ilae.org). Ce protocole préconise l'utilisation des séquences IRM T1 et T2 FLAIR avec une épaisseur de coupe minimale (*e.g.* millimétrique). Ces deux types de séquences permettent de mettre en avant des caractéristiques des lésions épileptogènes. Parmi celles-ci on trouve [Barkovich and Kuzniecky (1996), Besson *et al.* (2008), Bernasconi and Bernasconi (2015)] :

- l'épaississement du manteau cortical dans 50-90% des cas.
- une limite floue entre la matière grise (MG) et la matière blanche (MB) dans 60-80% des cas.

- une malformation de l'organisation gyrale.
- présence d'un signal hyper-intense en FLAIR dans 71-100% des cas.

Dans le contexte de l'imagerie de l'épilepsie, l'IRM anatomique est la méthode d'imagerie la plus utilisée. Cependant, lorsque la ZE n'est pas identifiée sur l'IRM (IRM négative), d'autres examens d'imagerie sont envisagés et en particulier la TEP et la MEG.

EEG de profondeur et Stéréo-EEG Si les techniques non-invasives (imagerie et enregistrements EEG) ne permettent pas l'identification précise de l'endroit de départ des crises et donc une planification de l'exérèse chirurgicale, un examen stéréo-EEG est alors envisagé. Cet examen constitue aujourd'hui l'examen de référence pour la localisation des ZE et est nécessaire dans 20 à 30% des cas [David *et al.* (2011)]. L'EEG de profondeur permet d'améliorer la résolution spatiale de l'EEG jusqu'à 1 cm. En particulier, en stéréo-EEG (SEEG) des électrodes (0.8 mm de diamètre et contenant 5 à 18 contacts d'enregistrement de 2 mm de long et séparés par des parties isolantes) peuvent être insérées en profondeur pour atteindre des zones cérébrales (*e.g.* les structures internes telles que l'amygdale ou l'hippocampe) et enregistrer leurs activités électriques. La principale limitation de la SEEG est son mauvais échantillonnage spatiale. En effet, seul un nombre limité d'électrodes peut être utilisé pour explorer le cerveau. Il est ainsi primordial de se reposer sur les résultats de la phase I du bilan pré-chirurgical pour formuler des hypothèses localisatrices et réduire l'étendu de la zone à explorer par les électrodes. Ces techniques d'exploration sont invasives et nécessitent un geste chirurgical qui comporte des risques.

Performances diagnostiques

Les techniques utilisées lors du bilan pré-opératoire permettent d'avoir des performances de détection très différentes. Cette performance dépend de l'acquisition (*e.g.* l'utilisation ou non d'un protocole optimisé pour la détection de l'épilepsie) et de l'expérience du radiologue et/ou neurologue chargé d'interpréter les données. Par exemple, la sensibilité de détection en utilisant l'IRM est de 39% lorsqu'un protocole non optimisé est utilisé et que les images IRM sont interprétées par un non expert. Lorsque les images sont interprétées par un expert cette sensibilité est de 50% et si en plus un protocole optimisé est utilisé celle-ci atteint les 90% [Von Oertzen *et al.* (2002), Duncan *et al.* (2016)]. La sensibilité dans la détection des lésions épileptogènes dépend aussi du type de la lésion. L'IRM est négative dans 20 à 30% des cas pour l'épilepsie du lobe temporal et dans 34% des cas pour les épilepsies de type FCD. Le succès de la chirurgie (patient libre de crises) est très corrélé avec la détection d'une lésion sur l'IRM ou en histologie (épilepsie lésionnelle). Pour les épilepsies du lobe temporal, le pronostic suite à la chirurgie est positif dans 69% des cas pour les épilepsies lésionnelles et dans 45% des cas pour les épilepsies non lésionnelles. Pour les épilepsies extra-temporales, ce pronostic est de 66% pour les épilepsies lésionnelles et de 34% pour les épilepsies non lésionnelles [Téllez-Zenteno *et al.* (2010)].

Chapitre 2 : Systèmes experts pour l'épilepsie

Les avancées récentes dans les techniques d'imagerie et l'identification de descripteurs permettant la détection des lésions épileptogènes ont permis d'améliorer le pronostic pour les patients souffrant d'épilepsie réfractaire aux médicaments. Ce pronostic reste cependant assez faible et est fortement corrélé à la détection d'une lésion sur les images médicales et en particulier sur l'IRM. Il dépend aussi des techniques d'imagerie utilisées et du niveau d'expérience du radiologue qui doit interpréter l'ensemble des données d'imagerie afin de localiser la zone épileptogène.

Les techniques de traitement de l'image ont été appliquées avec succès aux données de neuroimagerie afin d'assister les cliniciens dans leur tâche diagnostique [Norman *et al.* (2006), Klöppel *et al.* (2008), Orrù *et al.* (2012), Gray *et al.* (2013)]. En particulier, le développement de systèmes d'aide au diagnostic (CAD) est devenu un sujet de recherche prolifique. Ces systèmes peuvent servir à identifier des zones suspectes dans l'image et d'attirer l'attention des cliniciens sur ces zones ou alors au classement des images médicales dans plusieurs catégories (*e.g.* patient versus témoin sain). Comparé à l'analyse visuelle, le diagnostic assisté par ordinateur est en général plus rapide et permet d'obtenir des résultats plus reproductibles car celui-ci est moins subjectif et permet de réduire les variations inter- et intra-cliniciens.

Structure d'un système CAD

Malgré une grande diversité dans les domaines d'application, les systèmes CAD reposant sur l'imagerie partagent en général une architecture commune. Ils sont principalement composés de quatre étapes :

- **Pré-traitement des images :** l'objectif est d'améliorer la qualité des images. Cela comprend le débruitage d'images, la correction d'artefacts, le recalage spatial et la normalisation (mise à l'échelle) des images afin de permettre la comparaison d'images obtenues sous différentes conditions.

- **Définition du niveau de classification :** cela revient à définir la notion de "objet" qui sera ensuite classé. Pour un CAD reposant sur l'analyse de données d'imagerie, un objet peut correspondre à un voxel de l'image, une région d'intérêt (ROI) dans l'image ou l'image entière. Le choix d'un niveau pour la classification permet souvent de réduire la taille des données à analyser et permet dans certains cas de construire des modèles locaux (régionaux) simples. Souvent, il y a cependant un compromis à faire entre la simplicité de la représentation, la sensibilité spatiale souhaitée et la capacité à trouver des caractéristiques permettant de discriminer entre les différentes classes d'objets. On peut distinguer trois niveaux de classification, classification au niveau du patient, classification au niveau ROI et la classification au niveau du voxel. La classification au niveau du patient correspond à faire une analyse à l'échelle du patient, avec comme objectif de distinguer différents groupes de sujets, *e.g.* patients versus témoins sains. La classification au niveau du patient ne permet pas une localisation précise d'anomalies. La classification au niveau ROI permet une meilleure localisation des anomalies et permet l'apprentissage d'un modèle local spécifique à une ROI donnée. La classification au niveau du voxel consiste à décider pour chaque voxel s'il s'agit d'un voxel sain ou pathologique. Pour une image test, la sortie en général correspond à une carte indiquant des groupes de voxels pathologiques.

- **Extraction et sélection de caractéristiques :** il s'agit de calculer d'autres mesures à partir des images brutes ou pré-traitées. Les mesures extraites contiennent idéalement plus d'informations que les images brutes et sont non redondantes. Selon le niveau de classification choisi, ces caractéristiques sont assemblées dans un vecteur nommé vecteur de caractéristiques qui permet d'avoir une description synthétique d'un objet. Des connaissances à priori sur les différentes classes peuvent être utilisées pour guider l'extraction de caractéristiques discriminantes à partir des images initiales. Si aucun à priori sur les classes n'est connu, un grand nombre de caractéristiques peut être extrait. Elles peuvent être de plusieurs types : statistiques, fréquentielles, descripteurs de texture [Haralick *et al.* (1973)], descripteurs de forme et de bords [Lowe (1999), Dalal and Triggs (2005)]. Un des inconvénients majeurs de ce type de caractéristiques est le fait qu'un nombre limité de transformations est utilisé pour déduire les caractéristiques (*i.e.* les filtres, les descripteurs de forme...) et que celles-ci ne prennent pas en compte la nature des données brutes. Pour palier à cela, de nouvelles méthodes d'apprentissage de représentations ont récemment été développées. Parmi elles, on peut citer les méthodes d'apprentissage par dictionnaires [Mairal *et al.* (2009)] et les méthodes basées sur des architectures de type réseaux de neurones convolutifs profonds [Bengio *et al.* (2013)].

  Apprendre un modèle à partir d'observations ayant un grand vecteur de caractéristiques nécessite d'avoir une base d'apprentissage qui contient beaucoup d'exemples d'apprentissage. En pratique, un petit nombre d'observations est disponible pour apprendre le modèle. Dans ce cas, le problème d'apprentissage est dit mal posé. Pour éviter d'apprendre un modèle trop complexe et qui ne permet pas d'obtenir de bonnes performances lorsqu'il est testé sur de nouvelles observations (*i.e.* un système qui ne se généralise pas), une étape supplémentaire de sélection de caractéristiques est requise. L'objectif de cette étape est de réduire le nombre de caractéristiques en éliminant les caractéristiques non-informatives pour ne garder que les caractéristiques discriminantes et qui améliorent les performances de classification. Diverses méthodes de sélection de caractéristiques ont été proposées dans le contexte des systèmes CAD pour la neuroimagerie [Mwangi *et al.* (2014)].

- **Classification :** consiste en l'utilisation des caractéristiques extraites afin d'apprendre un modèle permettant d'attribuer à chaque objet une étiquette qui correspond à sa classe d'appartenance. Le modèle est en général appris sur des données d'apprentissage et utilisé ensuite pour déduire les étiquettes de nouvelles observations qui n'ont pas servi à l'apprentissage du modèle. De façon plus formelle, si $\mathbf{x}_i$, $\mathbf{x}_i \in \mathbb{R}^p$ est une observation, un algorithme de classification vise à estimer une fonction f :

$$\begin{array}{rccc} f: & \mathcal{X} = \mathbb{R}^p & \longrightarrow & \{1, \ldots, K\} \\ & \mathbf{x}_i & \longmapsto & y_i, \end{array} \tag{1}$$

  qui associe à $\mathbf{x}_i$ l'étiquette y_i. Cette fonction est en général obtenue comme solution d'un problème d'optimisation étant donnée une base de données d'apprentissage $\mathcal{X}^{tr}$. On peut distinguer entre deux principales techniques de classification : la classification supervisée et la classification non supervisée. On parle de classification supervisée lorsque la base de données d'apprentissage $\mathcal{X}^{tr}$ est constituée d'un nombre n observations $\mathbf{x}_i$ d'étiquette connue y_i. Lorsque la base d'apprentissage contient uniquement des observations $\mathbf{x}_i$ d'étiquette inconnue, on parle dans ce cas de classification non supervisée ou de partitionnement. Très souvent, le type de la fonction f est fixé à priori et l'apprentissage revient à estimer les paramètres $\mathbf{w}$

du modèle f choisi. Les paramètres optimaux $\mathbf{w}^*$ sont obtenus en minimisant une certaine fonction de perte $\mathcal{E}$ qui évalue la qualité des étiquettes prédites (classification) ou des groupes formés (partitionnement). Plusieurs choix sont possibles pour cette fonction de perte, notamment le coût 0-1, le coût Hinge ou encore l'erreur quadratique moyenne. Étant donné que nous ne disposons que d'un nombre fini d'observations pour trouver les paramètres optimaux, la fonction de perte est approximée par l'erreur empirique $\mathcal{E}_{\text{emp}}$ aussi appelée terme d'attache aux données qui correspond à l'erreur commise pour les données d'apprentissage. Pour éviter d'obtenir un modèle f trop complexe et qui se généralise mal (sur-apprentissage), un terme de régularisation aussi appelé erreur structurelle $\mathcal{E}_{\text{struct}}$ est aussi minimisé. Ainsi la fonction totale de perte à minimiser est la somme de deux termes :

$$\mathcal{E}(f, \mathbf{w}, \mathcal{X}^{tr}) = \mathcal{E}_{\text{emp}}(f, \mathbf{w}, \mathcal{X}^{tr}) + \lambda \mathcal{E}_{\text{struct}}(f, \mathbf{w}).$$

L'importance relative de ces deux termes est réglée via un paramètre de régularisation λ défini par l'utilisateur. Plusieurs algorithmes ont été proposés pour les problèmes de classification supervisée et non supervisée. Ces algorithmes diffèrent dans la définition des différents termes d'erreur, le choix de la fonction f et dans la résolution du problème d'optimisation [Xu and Wunsch (2005), Kotsiantis *et al.* (2007)].

En pratique, dans certains domaines d'application, il est parfois très difficile d'avoir accès à une base d'apprentissage avec suffisamment d'observations dans chacune des classes. Les observations de la classe pathologique (étiquetées ou non étiquetées) sont souvent plus coûteuses à obtenir et par conséquent la classe pathologique est souvent sous-échantillonnée par rapport aux autres classes présentes dans la base d'apprentissage. Dans ce cas, on parle de déséquilibre entre classes. Les algorithmes de classification standards (binaire ou multi-classes) ne permettent pas en général d'obtenir un bon modèle prédictif en cas de déséquilibre entre classes. Une alternative dans ce cas est de considérer des méthodes de détection de changement. Pour ce type de méthode, l'objectif est d'apprendre une description d'une classe cible (*i.e.* $\mathcal{X}^{tr}$ composé uniquement d'observations appartenant à la même classe). Toute observation n'appartenant pas à cette classe est alors détectée comme une nouveauté ou une anomalie compte tenu de la classe cible apprise. Comme pour les méthodes de classification, différents algorithmes ont été proposés dans le cadre de détection de changement [Chandola *et al.* (2009), A.F. Pimentel *et al.* (2014)].

Évaluation d'un système CAD

Plusieurs métriques peuvent être utilisées pour évaluer la performance d'un système d'aide au diagnostic. Cette performance dépend de différents facteurs. Les facteurs principaux sont [Petrick *et al.* (2013)] :

1. le bon partitionnement des données en une première base, la base de sélection du modèle, qui sert à estimer les paramètres optimaux du modèle et une deuxième base, la base de test, indépendante de la première qui sert uniquement à tester et évaluer les performances du modèle. Seul un nombre fini d'observations est disponible, par conséquent, trouver le bon partitionnement soulève souvent la problématique de la représentativité de la base de données d'apprentissage de la vraie distribution des classes. Des méthodes de partitionnement peuvent être utilisées. Elles consistent à partitionner la base de sélection des paramètres en base d'apprentissage et en base de validation. Si la règle de partitionnement est fixée par avance, on parle dans ce

cas de validation croisée (k-fold), si la règle correspond à effectuer plusieurs tirages aléatoires, on parle alors de bootstrap.

2. la définition de la vérité terrain pour les observations de la base de données de test. Cette définition est en général obtenue via un consensus d'experts ou en se basant sur les résultats d'un test de référence. Elle dépend aussi du niveau de classification choisi. Par exemple, pour une classification au niveau ROI, il est aussi nécessaire de définir un ensemble de règles (*e.g.* sur la localisation et/ou sur le recoupement) qui permettent de distinguer les détections correspondant à des vrais positifs ou à des faux positifs.

3. la définition de métriques d'évaluation (*e.g.* sensibilité, spécificité et courbe ROC). Ces métriques dépendent de la définition de la vérité terrain. Ces métriques sont aussi utilisées pour optimiser les hyper-paramètres (paramètres à définir par l'utilisateur) de l'algorithme de classification en utilisant la base de données de sélection des paramètres optimaux.

Les systèmes CAD pour l'épilepsie

Dans le contexte de la détection de l'épilepsie pharmaco-résistante, différents systèmes CAD ont été proposés dans la littérature. Pour la détection des épilepsies du lobe temporal, des méthodes de classification par patient ont été utilisées soit pour distinguer entre les patients et les témoins sains soit pour la latéralisation du foyer épileptogène (épilepsie du lobe temporal droit versus gauche) [Duchesne *et al.* (2006), Focke *et al.* (2012), Keihaninejad *et al.* (2012)]. Pour la détection de dysplasies corticales focales, la plupart des méthodes proposent une classification au niveau du voxel permettant de distinguer les voxels pathologiques de ceux correspondant à du tissu sain. Ces méthodes nécessitent le recalage des images des patients et des témoins afin de garantir la correspondance voxel à voxel [Bernasconi *et al.* (2001), Huppertz *et al.* (2005), Colliot *et al.* (2006), Thesen *et al.* (2011)]. Dans trois études, un niveau de classification intermédiaire a été considéré. En effet, [Besson *et al.* (2008), Hong *et al.* (2014), Ahmed *et al.* (2015)] proposent d'abord d'extraire la surface corticale et de la représenter par un maillage dont les sommets sont ensuite utilisés pour apprendre un modèle prédictif.

En ce qui concerne l'étape d'extraction de caractéristiques, des descripteurs qui modélisent les à priori cliniques ont été utilisés. Parmi cela, on trouve l'intensité du signal IRM T1 et FLAIR [Duchesne *et al.* (2006), Focke *et al.* (2012), Riney *et al.* (2012), Cantor-Rivera *et al.* (2015)], une mesure de l'épaisseur corticale [Antel *et al.* (2003), Srivastava *et al.* (2005), Thesen *et al.* (2011)] et la caractérisation de la limite entre la substance blanche et la substance grise (via un calcul de gradient) [Antel *et al.* (2003), Thesen *et al.* (2011), Ahmed *et al.* (2015)]. Les descripteurs purement image (*e.g.* texture, descripteurs de forme) ont rarement été considérés.

Enfin les méthodes de classification automatique n'ont pas été très utilisées. [Focke *et al.* (2012), Keihaninejad *et al.* (2012)] ont proposé d'utiliser un séparateur à vaste marge (SVM) pour la latéralisation des épilepsies du lobe temporal. [Antel *et al.* (2003)] ont proposé l'utilisation de deux classifieurs de Bayes en cascade pour détecter des lésions de type FCD. Dans les autres études, des méthodes statistiques de régression ont été utilisées et notamment le modèle linéaire [Srivastava *et al.* (2005), Bruggemann *et al.* (2007), Chassoux *et al.* (2010), Thesen *et al.* (2011)], l'analyse discriminante linéaire et la régression logistique [Hong *et al.* (2014), Ahmed *et al.* (2015)].

Les performances des différents systèmes CAD ont été évaluées par validation croisée.

Les métriques utilisées sont le taux de bonnes classifications (le nombre de vrais positifs et de vrais négatifs) et la mesures de sensibilité et spécificité. Dans la plupart des cas, la vérité terrain a été construite en prenant en compte l'avis d'un expert ou le résultat de l'histologie.

Chapitre 3 : Analyse du problème

Objectifs et verrous

L'objectif de ce projet est le design d'un système d'aide au diagnostic qui puisse extraire des caractéristiques discriminantes parmi des données multi-modales, composées des différentes modalités et/ou séquences d'images habituellement utilisées pour l'évaluation pré-chirurgicale des patients atteints d'épilepsie réfractaire. Pour chaque patient test, la forme souhaitée pour les réponses du système d'aide au diagnostic est celle d'une carte de groupements étiquetés, qui souligne les zones suspectes du cerveau laissant apparaître des anomalies associées à l'épilepsie réfractaire. Pour une meilleure interprétation de cette réponse, les étiquettes des groupements doivent représenter des scores de suspicion bien calibrés, et idéalement, correspondre à la probabilité que le groupement présente une anomalie liée à l'épilepsie.

Un autre aspect important qui doit être pris en compte dans l'analyse du problème est la base de données disponible pour construire et évaluer le système d'aide au diagnostic.

- **Données de patients :** dans ce projet, nous avons accès à des données de patients non étiquetées, provenant principalement de l'hôpital neurologique de Lyon, par le biais du Dr. A. Hammers (un premier groupe de cinq patients), ou en tant que partie prenante d'un projet de recherche PHRC (programme hospitalier de recherche clinique) en cours, initié par le Pr. F. Maugière et le Dr. J. Jung (un second groupe de patients). Ce programme de recherche a pour but d'évaluer l'intérêt d'utiliser une base de données composées d'images du cerveau multi-modales pour l'évaluation pré-chirurgicale de l'épilepsie réfractaire. La base de données cliniques associée devrait à terme contenir cent patients atteints d'épilepsie réfractaire ; cependant, les données de douze patients seulement étaient disponibles au moment de cette thèse. Tous les patients seront soumis systématiquement à un protocole d'imagerie comprenant à la fois la TEP au FDG et l'IRM (T1, FLAIR et de diffusion).

- **Données de témoins sains :** au travers de notre collaboration avec le CERMEP, infrastructure d'imagerie utilisée pour obtenir les données des patients, nous avons accès à trois bases de données de témoins sains. Dans la première base de données, le protocole d'imagerie correspond à celui utilisé pour le premier groupe de cinq patients souffrant d'épilepsie, et contient seulement les images IRM T1 de 37 sujets sains. Les deux autres bases de données de témoins sains ont été acquises au cours du programme de recherche PHRC. L'une d'elles est composée d'images IRM (T1, FLAIR et de diffusion) de 40 témoins sains. L'autre contient les données de 35 témoins sains qui ont été soumis aux examens IRM (T1, FLAIR et de diffusion) et TEP au FDG. Pour ces deux bases de données, les protocoles d'imagerie correspondent à ceux utilisés pour le second groupe de patients atteints d'épilepsie.

Étant données ces spécifications, plusieurs verrous peuvent être identifiés :

1. un déséquilibre de classes : c'est le résultat d'une disponibilité conséquente de données de témoins sains comparée à la disponibilité de données de patients étiquetées. Les données de patients étiquetées sont très coûteuses à l'acquisition car elles nécessitent un étiquetage manuel des observations pathologiques par un expert, fondé sur les données d'imagerie ou sur l'histologie. Le déséquilibre de classes est encore accentué par la taille réduite des anomalies liées à l'épilepsie réfractaire, comparée au volume du cerveau. En particulier, en ce qui concerne les lésions focales, leur

taille en voxels ne dépasse pas 1% du volume total du cerveau (*e.g.* 1,5 million de voxels cubiques millimétriques).

2. du bruit d'étiquetage : cela provient principalement de la subjectivité dans la délimitation manuelle de la zone épileptogénique, même après examen des données de neuroimagerie, en prenant en compte la zone opérée. En général, cette zone opérée peut aussi inclure du tissu sain. L'étiquetage de la zone opérée complète comme pathologique résulterait donc en un étiquetage incorrect de ce tissu sain situé dans la zone opérée.

3. des données de grande dimensionnalité : le manque de connaissances à priori sur la localisation potentielle des anomalies épileptogènes cibles rend très difficile une éventuelle restriction du domaine d'analyse. En raison d'une résolution relativement élevée des données de neuroimagerie utilisées par les protocoles optimisés pour la détection de liaisons épileptogènes, analyser complètement le cerveau nécessiterait de gérer des données de grandes dimensions, et peut être très coûteux.

4. des données multi-modales : au cours du bilan pré-chirurgical, les patients atteints d'épilepsie réfractaire sont soumis à plusieurs examens d'imagerie. L'IRM (T1, FLAIR) et la TEP sont les plus fréquents. La définition de la zone épileptogène repose sur l'examen de toutes les données de neuroimagerie disponibles. Chaque modalité ou séquence possède ses propres caractéristiques (dimension, résolution spatiale, contraste), qui aident à capturer les anomalies liées à l'épilepsie, et fournissent des informations complémentaires qui facilitent la localisation des zones épileptogènes. La prise en compte de toutes les modalités nécessite en général des approches multivariées.

5. du bruit sur les caractéristiques : cela peut être le résultat d'artefacts présents dans les images brutes, ou d'erreurs lors du prétraitement des images et dans les étapes d'extraction de caractéristiques, par exemple, pendant le recalage et la segmentation. La difficulté ici est de trouver un moyen de réduire l'impact de données d'apprentissage bruitées sur le modèle appris.

Certaines des méthodes présentées dans la revue de l'état de l'art présentée au Chap. 2 s'attaquent à des problèmes mentionnés ci-dessus. Dans [Antel *et al.* (2003), Hong *et al.* (2014), Ahmed *et al.* (2015)], des approches supervisées sont utilisées. Le déséquilibre de classes est traité en utilisant toutes les observations pathologiques disponibles, et en proposant une stratégie d'échantillonnage des données pour sélectionner un sous-ensemble des observations de la classe non pathologique [Antel *et al.* (2003)], ou bien en sélectionnant aléatoirement un échantillon des observations saines, de même taille que la classe pathologique [Hong *et al.* (2014), Ahmed *et al.* (2015)]. Dans les autres études, le déséquilibre de classes est évité en utilisant des modèles statistiques paramétriques et le modèle linéaire qui a seulement besoin des données d'une classe pour inférer les paramètres du modèle. Dans les paramètres de classification supervisée [Hong *et al.* (2014), Ahmed *et al.* (2015)], le bruit sur les étiquettes est traité en restreignant la définition de la vérité terrain. Des observations pathologiques sont extraites par la définition d'un masque qui restreint la zone opérée aux zones qui présentent aussi une caractéristique de texture ou une mesure d'épaisseur corticale anormale, selon un seuil donné, alors que les observations non pathologiques sont extraites seulement de sujets de contrôle sains. Dans [Antel *et al.* (2003), Hong *et al.* (2014), Ahmed *et al.* (2015)], la grande dimensionnalité des données de neuroimagerie est traitée en extrayant la surface corticale et en adoptant un schéma

de classification voxel par voxel. Les stratégies d'échantillonnage adoptées pour réduire le déséquilibre de classes aident aussi pour réduire le nombre total d'observations (*e.g.* 10 000 à 50 000 sommets au lieu de 1,5 million).

Il faut noter cependant que les deux derniers verrous, la nature multi-modale des données et le bruit sur les caractéristiques, ne sont pas spécifiquement traités par les méthodes de l'état de l'art et qu'aucune méthode combinant toutes les séquences et modalités d'imagerie n'a été étudiée.

Nos contributions

Notre première contribution est la formulation du problème de détection de lésion épileptogène comme un problème de détection de changement. Cette formulation était principalement motivée par le manque général de données de patients étiquetées, et en particulier dans les bases de données mises à disposition par nos collaborateurs experts. Cela réduit aussi le besoin de traiter le premier verrou (*i.e.* le déséquilibre de classes), discuté plus haut. Estimer un modèle en utilisant un algorithme de détection de changement (ou classification à une seule classe) nécessite seulement des observations provenant d'une classe, et évite d'obtenir un modèle biaisé par la classe majoritaire. Dans notre cas, le modèle est entraîné sur des observations provenant d'images de témoins sains. Ces images sont toutes passées en revue au préalable, et des témoins présentant des anomalies structurelles significatives ont été exclus de la base de données de témoins sains. Parmi tous les algorithmes de détection de changement, nous avons choisi les séparateurs à vaste marge, adaptés au cas d'une seule classe (OC-SVM), et sa variante, l'algorithme support vector data description (SVDD). Ces classifieurs sont entraînés à l'échelle du voxel, pour permettre de gérer les données de neuroimagerie qui sont typiquement de grande dimensionnalité, afin de garantir une localisation précise de la zone épileptogène, et d'éviter l'estimation de modèles complexes liés à la complexité anatomique du cerveau. Les caractéristiques utilisées pour apprendre le modèle ont été extraites exclusivement d'images IRM T1. Le système d'aide au diagnostique a été évalué et validé à la fois sur des données simulées réalistes, et des données de patients provenant des deux bases de données décrites plus haut. Pour réduire le coût calculatoire associé à l'apprentissage d'un modèle par voxel, le système d'aide au diagnostique a été implanté sur une architecture de calculs distribués, et déployé sur une grille de calcul. Cette première contribution est détaillée dans le Chap. 4 et le Chap. 5.

Dans ce premier système d'aide au diagnostique, les verrous 2 et 4 n'étaient pas particulièrement pris en compte. Le cadre proposé a donc été étendu. Pour gérer la présence de bruit dans les données d'apprentissage (bruit sur l'étiquetage), nous proposons une reformulation de l'algorithme support vector data description (SVDD), dans lequel le terme de pénalité L_1 dans le problème primal a été remplacé par un terme de pénalité L_0. Nous montrons que le problème résultant, L_0-SVDD, peut être résolu grâce à une procédure itérative fournissant des poids spécifiques aux données. Les détails liés à cette contribution sont donnés dans le Chap. 6.

Au Chap. 7, une stratégie de fusion optimale pour combiner des classifieurs OC-SVM/SVDD est étudiée pour traiter le caractère multi-modal des données de neuroimagerie. Deux niveaux de fusion des données ont été considérés. L'approche de fusion précoce consiste à construire un unique modèle global qui utilise les caractéristiques extraites de toutes les modalités d'imagerie; la fusion tardive consiste à apprendre des modèles locaux associés chacun avec une seule modalité d'imagerie, puis à combiner leurs résultats. Dans nos expériences, nous avons essayé de fusionner les informations extraites des trois

séquences d'IRM : T1, FLAIR et de diffusion. Les résultats du système d'aide au diagnostique ont été validés par des examens SEEG, et suggèrent que la meilleure stratégie pour la détection soit la fusion tardive.

Finalement, dans le Chap. 8, nous proposons de transformer les résultats du système d'aide au diagnostique en des probabilités bien calibrées, pour faciliter l'interprétation des scores, le choix du seuil, et la combinaison des scores. Une stratégie en deux étapes est proposée. Nous commençons par généraliser la méthode SVDD en reformulant le problème associé pour estimer des courbes de niveaux de probabilités dotées d'une structure hiérarchique, puis nous utilisons une fonction de calibration pour convertir les scores en sortie en probabilités. Deux fonctions de calibration ont été considérées : la sigmoïde, utilisée classiquement dans le contexte de classification binaire, et la loi d'extremum généralisée, qui convient mieux à l'estimation des distributions avec longue traîne, fréquentes dans le contexte de détection de changement. La sélection d'hyper-paramètres est réalisée en optimisant la qualité des niveaux de probabilités par validation croisée. L'optimisation de la qualité des probabilités estimées au lieu des performances de détection (*e.g.* précision, aire sous la courbe ROC) présente l'avantage considérable de ne pas nécessiter d'exemples provenant de la seconde classe pour sélectionner les paramètres du modèle.

Chapitre 4 : Méthodes de détection de changement

La problématique de détection de changement consiste à définir une description qui caractérise une classe cible donnée. Le but est ensuite de détecter si une observation test est conforme (acceptée) à cette description ou non. La différence majeure avec les méthodes de classification binaire est que la base d'apprentissage est constituée uniquement d'observations qui appartiennent à la classe cible. Les méthodes de détection du changement ont été utilisées dans plusieurs domaines d'application, parmi eux : la détection de fraude et d'intrusion, la détection d'anomalie dans le contexte médical et la reconnaissance de formes dans des images. Dans ce chapitre on s'intéresse à deux algorithmes de détection de changement : l'algorithme one-class support vecteur machine (OC-SVM) [Schölkopf *et al.* (2001)], une adaptation de l'algorithme de séparateurs à vaste marge (SVM) au problème de classification à une seule classe, et l'algorithme support vector data description (SVDD) [Tax and Duin (2004)], une de ses variantes. Le choix de ces deux algorithmes a été motivé par leur capacité à fournir des descriptions flexibles sans faire d'hypothèses restrictives sur la distribution des observations faisant partie de la classe cible. Ils sont aussi moins coûteux en termes de capacité de stockage et de temps de calcul requis lors de la phase de test. Enfin, les deux approches permettent de contrôler l'erreur de classification commise pour les observations de la classe cible par le biais d'un paramètre défini par l'utilisateur.

L'algorithme OC-SVM

Principe Cet algorithme a été introduit par [Schölkopf *et al.* (2001)]. Il s'agit d'une adaptation des SVM classiques [Vapnik (1998)] au problème de classification à une seule classe ("single class classification problems") ou encore au problème de détection de nouveauté.

Le principe de la méthode est d'abord de projeter les données d'apprentissage dans un espace de dimension plus grande (appelé espace des caractéristiques) par l'intermédiaire d'une fonction noyau ϕ. Cette fonction est associée à un noyau K qui définit le produit scalaire dans l'espace d'arrivée. Ce noyau est choisi de sorte à garantir la séparabilité entre les données d'apprentissage après projection et l'origine de l'espace d'arrivée. Dans l'espace d'arrivée, le problème est alors formulé comme un problème de classification de type SVM à deux classes où l'origine est considérée comme le seul élément de la deuxième classe. De même que dans le SVM classique, on cherche parmi une infinité d'hyperplans séparateurs celui qui maximise la marge entre les deux classes.

Formulation dans le primal [Schölkopf *et al.* (2001)] propose de formuler le problème associé au OC-SVM comme un problème de minimisation d'une fonction de perte constituée d'un terme de régularisation L_2 et d'un coût de type Hinge. La fonction de perte est :

$$\begin{aligned} \mathcal{E}_{\text{OC-SVM}}(\mathbf{w}, \rho, \mathcal{X}^{tr}) &= \mathcal{E}_{\text{struct}} + \lambda\, \mathcal{E}_{\text{emp}} \\ &= \tfrac{1}{2}\|\mathbf{w}\|^2 - \rho + \tfrac{1}{\nu n} \sum_{i=1}^{n} max(0, -((\phi(\mathbf{x}_i) \cdot \mathbf{w}) - \rho)), \end{aligned}$$

Des variables ressorts ξ_i sont ensuite introduites pour relâcher les contraintes sur les exemples d'apprentissage. Le problème de minimisation associé à l'algorithme OC-SVM peut

alors s'écrire sous la forme :

$$\begin{cases} \min\limits_{\mathbf{w},\rho,\boldsymbol{\xi}} & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum\limits_{i=1}^{n}\xi_i \\ \text{avec} & (\phi(\mathbf{x}_i)\cdot\mathbf{w}) \geq \rho - \xi_i, \qquad i \in [1,n] \\ \text{et} & \xi_i \geq 0, \qquad\qquad\qquad\quad i \in [1,n], \end{cases}$$

où $\frac{1}{\nu n}$ correspond au paramètre de régularisation qui permet de faire un compromis entre la complexité du modèle (erreur structurelle) et le nombre d'erreurs (erreur empirique). La fonction de décision du OC-SVM est donnée par : $f_{OC-SVM}(\mathbf{x}) = sgn((\mathbf{w}.\phi(\mathbf{x})) - \rho)$, où sgn désigne la fonction signe (*i.e.* $sgn(z)$ est égale à 1 si $z > 0$ et -1 sinon).

Formulation duale En introduisant le Lagrangien, on peut déduire la formulation duale de OC-SVM. Si les $\alpha_i,\ i \in [1,n]$ dénotent les multiplicateurs de Lagrange associés aux contraintes d'inégalité $(\phi(\mathbf{x}_i)\cdot\mathbf{w}) \geq \rho - \xi_i$, alors la formulation duale est :

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^T K \boldsymbol{\alpha} \\ \text{avec} & \sum\limits_{i=1}^{n}\alpha_i = 1 \\ \text{et} & 0 \leq \alpha_i \leq \frac{1}{\nu n} \quad i \in [1,n], \end{cases}$$

Cette formulation duale correspond à un problème quadratique qui peut être résolu en utilisant un solveur standard.

L'algorithme SVDD

Principe Cette méthode de classification a été proposée par [Tax and Duin (2004)]. On cherche cette fois non pas à trouver l'hyperplan séparateur qui maximise la marge mais la sphère de volume minimal. De même que pour OC-SVM, les données d'apprentissage sont projetées dans un espace de plus grande dimension où la sphère de centre $\mathbf{a}$ et de rayon R qui contient la majorité des exemples de la classe cible est recherchée.

Formulation dans le primal [Tax and Duin (2004)] propose aussi d'utiliser dans la fonction de perte un terme de régularisation L_2 et un coût de type Hinge pour l'erreur empirique.

Le problème de minimisation dans le primal est alors :

$$\begin{cases} \min\limits_{R,\mathbf{a},\boldsymbol{\xi}} & \underbrace{R^2}_{\text{erreur structurelle}} + C \underbrace{\sum\limits_{i=1}^{n}\xi_i}_{\text{erreur empirique}} \\ \text{avec} & ((\phi(\mathbf{x}_i) - \mathbf{a})\cdot(\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i, \qquad i \in [1,n] \\ \text{et} & \xi_i \geq 0, \qquad\qquad\qquad\qquad\qquad\quad i \in [1,n]. \end{cases}$$

Le terme $((\phi(\mathbf{x}_i)-\mathbf{a})\cdot(\phi(\mathbf{x}_i)-\mathbf{a}))$ correspond à la distance entre l'observation $\mathbf{x}_i$ et le centre de la sphère $\mathbf{a}$ dans l'espace projeté. Le paramètre C est le paramètre de régularisation qui permet d'équilibrer les deux types d'erreurs. La fonction de décision du SVDD est alors donnée par : $f_{SVDD}(\mathbf{x}) = sgn(((\phi(\mathbf{x}_i) - \mathbf{a})\cdot(\phi(\mathbf{x}_i) - \mathbf{a})) - R^2)$.

Formulation duale De même que pour OC-SVM, en introduisant les multiplicateurs de Lagrange α_i, on obtient la formulation duale suivante :

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^T K \boldsymbol{\alpha} - \boldsymbol{\alpha}^T diag(K) \\ \text{subject to} & \sum\limits_{i=1}^{n} \alpha_i = 1 \\ \text{and} & 0 \leq \alpha_i \leq C \qquad\qquad i \in [1, n], \end{cases}$$

où $diag(K)$ correspond à la diagonale de la matrice K définie par ses éléments $K_{ij} = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j)$. La formulation duale de l'algorithme SVDD est aussi un problème quadratique. Cette formulation est très similaire à celle de l'algorithme OC-SVM. Les cas d'équivalence entre ces deux formulations ont été discutés dans [Schölkopf *et al.* (2001)]. Une preuve de cette équivalence rédigée par le professeur Stéphane Canu est donnée en annexe D.

Optimisation des hyper-paramètres

Les deux algorithmes ont chacun un hyper-paramètre dont la valeur doit être optimisée. Les paramètres ν dans la formulation de OC-SVM et C dans la formulation de SVDD jouent un rôle similaire. En effet, ces deux paramètres permettent de trouver un compromis entre le terme d'attache aux données et le terme de régularisation. Utiliser une fonction de projection ϕ et donc un noyau pour définir le produit scalaire dans l'espace des caractéristiques revient à introduire un autre paramètre. En général, le choix du noyau est fixé, et il s'agit juste d'optimiser le paramètre associé à ce noyau. Par exemple, pour le noyau Gaussien, la largeur du noyau doit être optimisée.

Les métriques d'évaluation (*e.g.* sensibilité, spécificité et courbe ROC) classiquement utilisées dans les problèmes de classification ne peuvent pas être utilisées dans le contexte de détection de changement. En effet, lors de l'optimisation des paramètres, des exemples de classes autre que la classe cible ne sont pas à disposition pour permettre le calcul de ce genre de métriques.

[Tax and Duin (2004)] propose une méthodologie pour optimiser les paramètres du modèle SVDD. Cette méthodologie consiste faire varier les paramètres du modèle sur une grille de valeurs et d'estimer à chaque noeud de la grille l'erreur de classification mesurée par validation croisée (de type leave-one-out) sur les données d'apprentissage. Les paramètres optimaux sont choisis de sorte à avoir un taux de fausse alarme donné.

[Schölkopf *et al.* (2001)] propose d'utiliser une propriété du paramètre ν pour fixer un taux de faux positifs donné.

Chapitre 5 : Système CAD pour la détection de lésions épileptogènes

Dans ce chapitre nous proposons un système d'aide au diagnostic reposant sur une analyse multivariée de cartes de texture permettant de caractériser les lésions épileptogènes de type FCD et hétérotopies. Un modèle prédictif de type OC-SVM est appris à l'échelle du voxel en utilisant les images de témoins sains. L'algorithme OC-SVM a rarement été utilisé dans le contexte de classification de données de neuroimagerie [Mourão Miranda *et al.* (2011), Sato *et al.* (2012)]. L'intérêt principal de cette formulation est de s'affranchir de la difficulté relative à la construction de bases de données de patients annotées nécessaires pour apprendre un modèle de classification supervisée. Les performances du système CAD sont comparées à celles d'un modèle statistique optimisé pour cette tâche de détection. Deux types de simulations réalistes ont été réalisés et utilisés pour bien quantifier et évaluer les performances du système. Des données cliniques de 11 patients ont aussi servi à évaluer les performances du système.

Description des données

Base de données d'apprentissage : Deux bases de données de témoins ont été utilisées. La première base de données est composée des IRM T1 de 37 témoins sains âgés de 18 à 53 ans. La deuxième base de données de témoins est composée des images IRM T1 de 40 témoins sains âgés de 20 à 62 ans. Les deux bases de données ont été inspectées visuellement par des radiologues pour éliminer les images de sujets présentant des anomalies structurelles significatives. Ces deux bases de données ont été acquises sur le même scanner IRM mais diffèrent dans les paramètres de la séquence IRM T1.

Base de données de simulation : cinq IRM additionnelles de témoins sains ont été utilisées pour simuler deux types de lésions épileptogènes. Ces IRM ont été acquises selon le même protocole utilisé pour la première base de données de témoins sains.

Pré-traitements

Pour pouvoir faire une analyse voxel à voxel, il est indispensable d'avoir une correspondance entre les voxels des images des différents sujets. Pour cela, les images IRM T1 de tous les sujets de la base sont plongées dans un même espace de référence. Cette procédure se fait en deux étapes : une étape de segmentation et une étape de recalage (ou normalisation). L'espace Montreal Neurological Institute (espace MNI) a été choisi comme espace de référence et la segmentation et la normalisation ont été réalisées en utilisant l'algorithme "unified segmentation" [Ashburner and Friston (2005)] implémenté dans le package SPM8. Cet algorithme permet de faire la segmentation et la normalisation et la correction d'artefacts (non-homogénéité) en une seule étape.

Segmentation Le processus de segmentation est basé sur une analyse statistique de l'intensité de chaque voxel des images IRM multi-séquences. Une carte de probabilités à priori est utilisée : elle donne la probabilité à priori pour chaque voxel d'appartenir à l'une des classes que l'on cherche à segmenter. Cette carte de probabilités, mise à disposition par le Montreal Neurological Institute est largement utilisée, et permet une segmentation à priori du cerveau en 6 classes tissulaires, dont les trois plus importantes sont : la matière blanche (MB), la matière grise (MG) et le liquide céphalo-rachidien (LCR). Cette carte de probabilités a été construite à partir de 305 IRM-T1, à partir desquelles une segmentation semi-automatique a été effectuée.

Dans l'algorithme "unified segmentation" un modèle de mixture de gaussiennes basé sur l'algorithme d'Espérance-Maximisation est utilisé. Ce modèle suppose que la distribution de l'intensité de chaque classe de tissus k est approximée par une gaussienne $G(\mu_k, \Sigma_k)$ de moyenne μ_k et de matrice de covariance Σ_k. La probabilité à priori d'un voxel i d'appartenir à une classe k est disponible dans la carte de probabilités. Les probabilités à posteriori sont calculées en appliquant la loi de Bayes et correspondent aux probabilités d'appartenance à chaque classe tissulaire. L'algorithme alterne entre le calcul des paramètres de classe μ_k et Σ_k- étape de Maximisation - et la mise à jour des probabilités à posteriori - étape d'Espérance.

Normalisation spatiale Pour utiliser correctement les informations fournies par la carte de probabilités, il est nécessaire de placer les images de la carte de probabilités et du patient dans le même repère. Le choix a été fait de se placer dans l'espace MNI. Le recalage utilisé dans l'algorithme "unified segmentation" est un recalage non-rigide. Le principe de base de ce recalage est de trouver une transformation spatiale non linéaire qui permet de faire correspondre une image donnée à une image de référence tout en minimisant un certain critère (par exemple : minimisation de l'erreur quadratique, maximisation du coefficient de corrélation). En particulier, dans l'algorithme utilisé, la transformation spatiale correspond à une combinaison linéaire d'un millier de fonctions de base de la transformée en cosinus discrète 3D et l'objectif du recalage est de faire correspondre la matière grise de l'image considérée à la matière grise de l'image de référence.

Définition du volume d'intérêt Les images normalisées dans l'espace MNI sont ensuite masquées pour retirer les zones non concernées par la dysplasie. Ce masque a été créé à partir de l'atlas anatomique de maximum de probabilité de Hammersmith [Hammers *et al.* (2003)] en retirant le cervelet, le tronc cérébral et toutes les structures centrales (noyaux caudés, corps calleux, ...). L'application de ce masque aux images normalisées donne un volume d'intérêt contenant 1,5 million de voxels à classifier.

Extraction de caractéristiques

L'extraction de caractéristiques discriminantes pour la pathologie qu'on cherche à identifier est une étape importante de la chaîne de traitement du CAD. L'étude de l'état de l'art des systèmes CAD dédiés à l'épilepsie (cf. Chap. 2) a montré que trois caractéristiques qui correspondent à la description clinique des lésions de type FCD ont été utilisées. Ces caractéristiques sont :

- une carte de jonction : cette carte permet de caractériser la limite entre la matière grise et la matière blanche. Elle permettra donc de détecter des anomalies dues à une limite floue entre ces deux tissus
- une carte d'extension : cette carte permet de caractériser une avancée anormale de la matière grise dans la matière blanche. Elle aidera donc dans la détection des anomalies de gyration telles que des sillons trop profonds.
- une carte d'épaisseur : cette carte donne la mesure de l'épaisseur corticale à chaque voxel et permet donc de caractériser des changements dans l'épaisseur du manteau cortical.

Des détails sur le calcul de ces trois cartes sont donnés dans [Huppertz *et al.* (2005), Bernasconi *et al.* (2001)]. Les trois cartes ont été essentiellement calculées à partir de

la carte de probabilités de la matière grise. Pour conserver de l'information sur les deux autres tissus (matière blanche et liquide céphalo-rachidien), les cartes de probabilités des trois tissus ont aussi été considérées comme des caractéristiques discriminantes. Chaque voxel de l'image est ainsi caractérisé par un vecteur de six composantes correspondant aux valeurs que prend ce voxel dans chacune des six cartes paramétriques qui ont été extraites.

Classification

Estimation du modèle à l'échelle du voxel Pour avoir une cartographie précise des zones suspectes, une analyse voxel à voxel a été choisie. Chaque voxel k du volume d'intérêt est représenté par une matrice $M^k \in M_{n,p}(\mathbb{R})$ où $n = 37$ ou 40, le nombre de témoins de la base d'apprentissage et $p = 6$, le nombre de caractéristiques. Ainsi $M^k_{i,j}$ correspond à la valeur de la $j^{ème}$ caractéristique pour le $i^{ème}$ sujet de la base d'apprentissage pour le voxel k.
Chaque voxel k est ensuite associé à un classifieur OC-SVM indépendamment des autres voxels. Le volume d'intérêt étant composé de 1.5 millions de voxels, on se retrouve avec 1.5 millions de classifieurs OC-SVM.

Optimisation des paramètres Pour garantir après projection la séparabilité des données d'apprentissage de l'origine du l'espace des caractéristiques, nous avons utilisé un noyau Gaussien paramétré par son écart-type σ. Les hyper-paramètres de l'algorithme OC-SVM ont été optimisés par validation croisée de type *leave-one-out*. Cette technique est fortement conseillée lorsque l'étude repose sur peu d'exemples d'apprentissage. Le *leave-one-out* consiste à apprendre le modèle sur $n-1$ échantillons et de le valider sur l'échantillon restant. On répète ensuite cette opération en choisissant un autre échantillon de validation parmi les $n-1$ échantillons qui n'ont pas encore été utilisés pour la validation. L'opération est répétée n fois jusqu'à ce que tous les échantillons aient été utilisés exactement une fois pour la validation. Pour chaque combinaison de (C, σ), nous avons réalisé une validation croisée de type *leave-one-out*. Pour chaque boucle du *leave-one-out*, nous avons calculé le nombre de mauvaises classifications. La combinaison optimale (C, σ) est celle qui donne le plus petit taux de mauvaises classifications sur l'ensemble des boucles.
En théorie, cette procédure d'optimisation doit être menée pour chaque voxel séparément, mais en pratique, pour réduire les temps de calcul, nous avons sélectionné aléatoirement 4000 voxels dans le volume d'intérêt et nous avons retenu la combinaison (ν, σ) qui permet d'avoir le plus petit taux de mauvaises classifications pour l'ensemble de ces voxels. La sélection aléatoire de ces voxels, permet de garantir que les valeurs de (C, σ) retenues ne dépendent pas d'une région en particulier dans le cerveau. En appliquant cette procédure, les valeurs $\nu = 0.03$ et $\sigma = 4$ ont été retenues pour la première base de données et les valeurs $\nu = 0.05$ et $\sigma = 3$ pour la deuxième base de données.

Post-traitements

Lorsque les modèles OC-SVM sont appliqués à une image test normalisée avec les mêmes dimensions que les images ayant servi pour l'apprentissage, on obtient en sortie une carte de distances dans laquelle la valeur d'un voxel est donnée par sa distance à l'hyper-plan séparateur estimé à cet endroit. Le seuillage de cette carte de distances permet de faire ressortir les voxels qui s'écartent le plus de la description de la classe cible.

Nous proposons une méthode originale pour fixer la valeur de ce seuil qui permet de contrôler l'erreur de type I (le taux de faux positifs). Le principe de cette méthode est de modéliser la distribution des scores OC-SVM des voxels correspondant à du tissu sain pour

ensuite déduire la probabilité pour un voxel test d'être pathologique étant donné son score OC-SVM. Cette distribution normative a été construite à partir de la distribution des scores obtenus pour les témoins sains de la base d'apprentissage et suivant une procédure de *leave-one-out*. Les 37 ou 40 histogrammes de distribution sont ensuite moyennés pour obtenir une seule distribution qui est approximée en utilisant une méthode statistique non-paramétrique de type fenêtres de Parzen. Enfin, la valeur du seuil est choisie par rapport à cette distribution pour avoir une p-valeur fixée. Dans nos expériences, les seuils correspondants à une p-valeur de 0.001 ont été utilisés. Après seuillage de la carte des distances, des groupements de voxel (clusters) sont détectés, ordonnés puis labellisés en considérant la distance OC-SVM minimale des voxels faisant partie de chaque cluster.

Évaluation du système CAD

Comparaison avec l'analyse SPM Une analyse statistique voxel à voxel uni-variée a été utilisée dans le cadre du formalisme SPM afin de tester l'image d'un patient contre les images des témoins sains de la base de données. Cette analyse repose sur le modèle linéaire. Ce modèle cherche à partir d'un vecteur d'observations Y à trouver les p facteurs X_i qui permettent d'expliquer les observations. Cela se traduit par l'équation suivante : $Y = X\beta + e$, où X est la matrice des facteurs et e représente les composantes du bruit résiduel. L'hypothèse forte de ce modèle est que les composantes du bruit sont supposées être identiques, indépendantes et normalement distribuées. Le paramètre β_i quantifie l'effet du facteur X_i dans le modèle. Les paramètres β_i sont estimés par la méthode des moindres carrées de sorte à minimiser l'erreur quadratique e^2. On définit alors des contrastes pour quantifier l'effet d'un facteur donné sur la variable observée. Cela permet ensuite de faire des tests statistiques (e.g. test T) en considérant comme hypothèse nulle que le facteur n'a pas d'effet sur la variable observée.

Nous avons réalisé une analyse de variance (ANOVA) en utilisant quatre facteurs d'intérêt : la carte de jonction du patient, les cartes de jonction des témoins sains, la carte d'extension du patient et les cartes d'extension des témoins sains. Deux contrastes ont été définis [1,-1,0,0] et [0,0,1,-1] pour tester les différences significatives entre les cartes de jonction et d'extension du patient et celles des témoins sains. Nous avons aussi utilisé le contraste combiné (conjonction des deux contrastes) qui permet de tester une différence significative dans l'une des deux cartes. Les cartes de scores T ont été seuillées en utilisant une p-valeur de 0.001 pour faire ressortir des clusters de voxels qui restent significatifs étant donné ce seuil de détection. Pour permettre une comparaison plus juste entre les résultats de SPM et ceux de notre système CAD, seules les cartes de jonction et d'extension ont été utilisées pour apprendre le modèle OC-SVM au lieu des six caractéristiques de départ.

Évaluation des performances pour les données simulées Travailler sur des données simulées permet de s'affranchir de toutes les problématiques liées à la définition de la vérité terrain. Nous avons essayé, à partir de la description clinique des malformations corticales, de simuler des lésions très semblables à des lésions épileptogènes. En particulier, à partir des images IRM des témoins sains, nous avons simulé deux type de lésions : des anomalies de jonction associées avec des FCD et des hétérotopies. Nous avons simulé un total de 5 lésions de type anomalie de jonction localisée autour d'un sillon et 90 lésions de type hétérotopie. Pour les lésions de type hétérotopie, le contraste ainsi que la localisation des lésions ont été variés.

Pour ces données de simulations, la courbe ROC ainsi que la sensibilité à différents niveaux de faux positifs ont été utilisées pour évaluer les performances du système CAD

et de l'analyse SPM.

Évaluation des performances pour les données cliniques Pour chaque patient de la base de test, les cartes des clusters labellisés résultant du traitement par OC-SVM et par les trois analyses SPM ont été analysées par un expert. Seuls les clusters dont la taille en voxels excède les 82 voxels ont été retenus. Chacun de ces clusters a été qualifié par l'expert comme étant 1) un faux positif, 2) une lésion mais non responsable des crises ou 3) la lésion recherchée. Pour prendre cette décision, l'expert s'est basé sur le dossier médical de chaque patient.

Résultats

Résultats pour les données simulées :
- *Anomalie de jonction* : L'analyse des courbes ROC montre que les deux méthodes OC-SVM et SPM réussissent à identifier les lésions avec une performance équivalente en terme d'aire sous la courbe ROC. Cependant, le système OC-SVM permet une meilleure sensibilité dans la détection pour des taux de faux positifs bas.
- *Lésion type hétérotopie* : L'analyse des courbes ROC montre un avantage significatif en faveur du système OC-SVM par rapport aux analyses SPM et en particulier pour les lésions les plus contrastées. De même que pour les anomalies de jonction, des taux de faux positifs plus bas peuvent être considérés pour OC-SVM sans une grande perte de sensibilité dans la détection.

Résultats pour les données cliniques : 11 patients (avec un total de 13 anomalies) ont été utilisés pour comparer les performances du modèle OC-SVM et des analyses SPM.
- *Résultats pour les IRM positives* : le système OC-SVM a permis la détection de 3 lésions sur 3 avec en moyenne 1.7 faux positifs par patient. Pour les analyses SPM, les meilleures performances ont été obtenues avec le contraste de la jonction. Celui-ci a permis la détection de 3/3 lésions mais avec en moyenne 6.3 faux positifs par patient.
- *Résultats pour les IRM négatives* : le système OC-SVM a détecté 7 sur 10 lésions avec en moyenne 3.7 faux positifs par patient. Parmi les différents contrastes SPM, le contraste de conjonction a permis d'avoir les meilleures performances, en détectant 5 lésions sur 10 avec en moyenne 5.7 faux positifs par patient.
- *Performances globales* : pour l'ensemble des lésions, le système OC-SVM a détecté 10 lésions sur 13 avec en moyenne 3.2 faux positifs par patient. L'analyse SPM basée sur la conjonction de contrastes a détecté 7 lésions sur les 13 avec en moyenne 6.3 faux positifs par patient. Les analyses SPM basées sur les contrastes individuels (jonction et extension) ont détecté 6 sur 13 lésions avec en moyenne 7 faux positifs par patient pour le contraste de la jonction et 21 faux positifs par patient pour le contraste de l'extension.

Chapitre 6 : Méthode de détection de changement robuste au bruit

Motivation

Dans ce chapitre on s'intéresse à la robustesse des méthodes de détection de changement de type OC-SVM et SVDD. A l'aide d'un exemple clinique, on illustre le fait que ces méthodes ne sont pas très robustes à la présence de bruit d'étiquetage dans la base d'apprentissage. Ce type de bruit peut avoir différentes sources et est naturellement présent lorsque l'étiquetage est réalisé par des experts humains.

Coût Hinge et sensibilité au bruit d'étiquetage

Le problème de minimisation associé à l'algorithme SVDD est :

$$\begin{cases} \min\limits_{R,\mathbf{a},\boldsymbol{\xi}} & \underbrace{R^2}_{\text{erreur structurelle}} + C \underbrace{\sum_{i=1}^{n} \xi_i}_{\text{erreur empirique}} & \\ \text{avec} & ((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i, & i \in [1, n] \\ \text{et} & \xi_i \geq 0, & i \in [1, n]. \end{cases}$$

Le terme de l'erreur empirique correspond à un coût de type Hinge. La minimisation du coût Hinge revient à pénaliser plus sévèrement les larges valeurs d'erreur (ξ_i) par rapport aux petites valeurs d'erreur. Cela est dû au fait que le coût de mauvaise classification d'une observation $\mathbf{x}_i$ est proportionnel à $C\xi_i$. La présence de données mal étiquetées dans la base d'apprentissage est susceptible de générer de larges erreurs. Le paramètre C peut être utilisé pour équilibrer les deux termes d'erreur. Cependant, en pratique optimiser ce paramètre dans le cadre des problématiques de détection de changement est difficile à réaliser car des exemples étiquetés appartenant à des classes autres que la classe cible ne sont pas disponibles pour l'optimisation des paramètres. Par conséquent, il est primordial de prendre en compte la présence de ce type de bruit dans l'algorithme de détection pour ne pas obtenir de mauvaises performances de détection.

Etat de l'art

Deux adaptations de l'algorithme SVDD ont été proposées pour prendre en compte la présence de données mal étiquetées dans la base d'apprentissage.

[Lee *et al.* (2007)] et [Liu *et al.* (2013)] ont proposé de modifier le terme de coût Hinge en associant à chaque observation un terme de pondération w_i. La nouvelle fonction objective à minimiser est alors donnée par :

$$\min_{R,\mathbf{a},\boldsymbol{\xi}} R^2 + C \sum_{i=1}^{n} w_i \xi_i$$

Dans [Lee *et al.* (2007)] les termes de pondération correspondent à une mesure de densité locale. Dans la formulation proposée par [Liu *et al.* (2013)], les termes de pondération correspondent à un score de confiance calculé par rapport à la distance du centre de masse des observations d'apprentissage.

Formulation L_0-SVDD

Nous proposons une nouvelle formulation du problème SVDD basée sur le coût 0-1, qui vaut 1 en cas de mauvaise classification (erreur) et 0 sinon. La nouvelle formulation du problème dans le primal est alors donnée par :

$$\begin{cases} \min\limits_{\mathbf{a},R,\boldsymbol{\xi}} & R^2 + C\|\boldsymbol{\xi}\|_0 \\ \text{avec} & ((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i \\ \text{et} & \xi_i \geq 0 \qquad i \in [1, n]. \end{cases}$$

Dans cette formulation, l'erreur empirique correspond uniquement au nombre d'erreurs et n'est plus proportionnelle à la somme des erreurs.

Malheureusement, la pénalité de type l_0 est non différentiable et le problème de minimisation associé ne peut pas être résolu numériquement de façon efficace. Pour remédier à cela nous procédons en deux étapes. Nous proposons d'abord d'approximer le terme de coût l_0 par une fonction coût logarithmique. Le nouveau problème d'optimisation est ensuite résolu en utilisant une approche itérative qui résout à chaque itération un problème de minimisation avec une pénalité de type l_1. Cela résulte de l'écriture du terme de pénalité logarithmique sous forme de différence de deux fonctions convexes (DC programming). La deuxième fonction est ensuite linéarisée localement.

Pour notre formulation L_0-SVDD, résoudre le problème primal revient à résoudre itérativement le problème suivant :

$$\begin{cases} \min\limits_{\mathbf{a},R,\boldsymbol{\xi}} & R^2 + C\sum\limits_{i=1}^{n} w_i\xi_i \\ \text{avec} & ((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i \\ \text{et} & \xi_i \geq 0 \qquad i \in [1, n] \end{cases} \qquad \text{où} \qquad w_i = \frac{1}{\gamma + \xi_i^{\text{old}}}.$$

On remarque que la formulation obtenue ressemble à celles proposées par [Lee *et al.* (2007)] et [Liu *et al.* (2013)] car elle fait apparaître un terme de pondérations pour chaque observation de la base d'apprentissage. La différence étant que dans la formulation proposée, les termes de pondération sont calculés de façon adaptative, qui ne dépend d'aucune heuristique.

Expériences et résultats

Application à la détection de lésion épileptogènes Dans le système CAD proposé dans le chapitre 5, nous avons remplacé le classifieur OC-SVM par notre la formulation L_0-SVDD. Pour illustrer l'apport de cette nouvelle formulation, nous avons utilisé un exemple de lésion simulée de type hétérotopie. Les performances du système ont été évaluées en calculant l'indice DICE qui représente le recouvrement voxel à voxel entre la vérité terrain et la sortie du CAD. Les performances du système ont aussi été comparées à celles de la méthode proposée par [Liu *et al.* (2013)].

Le modèle L_0-SVDD a bien détecté la lésion simulée avec un indice de recouvrement (DICE) total de 60%, alors que la méthode proposée par [Liu *et al.* (2013)] n'a pas permis de détecter la lésion.

Application aux bases de données référencées UCI Quatre bases de données référencées du dépôt UCI ont été utilisées pour comparer la performance de L_0-SVDD à celle de la formulation initiale de l'algorithme SVDD. Nous avons en particulier évalué

la robustesse au bruit d'étiquetage. Les bases de données UCI correspondent à des bases de données de classification contenant plusieurs classes. La classe majoritaire a été considérée comme la classe cible. Pour simuler la présence de bruit d'étiquetage, les observations appartenant aux classes autres que la classe cible ont été ajoutées dans la base de données d'apprentissage après avoir changé les étiquettes associées. Différents niveaux de bruit ont été considérés. Les résultats de cette expérience ont démontré que l'utilisation de la formulation L_0-SVDD permet d'obtenir de meilleure performance de détection et est plus robuste à la présence de bruit d'étiquetage par rapport à la formulation initiale de l'algorithme SVDD.

Chapitre 7 : Méthode multi-modale pour la détection de changement

Objectif

Dans ce chapitre, nous nous intéressons à l'amélioration du système CAD en terme de performance, en particulier face à des cas difficiles. Nous faisons l'hypothèse qu'un gain significatif sur les performances de diagnostic peut être obtenu en complétant les caractéristiques initiales par d'autres paramètres dérivés du FLAIR et/ou du DTI. Une stratégie de fusion de données optimale qui maximise les performances de détection est proposée. Les résultats de détection pour trois patients sont ensuite validés en les confrontant à une carte d'indices d'épileptogénicité dérivée de l'examen SEEG examen de référence.

Fusion de données

La fusion de données, peut être vue comme une application particulière de l'apprentissage par ensemble de classifieurs [Polikar (2012)]. Elle correspond à faire un apprentissage à partir de mesures enregistrées à l'aide de divers capteurs. La combinaison de ces mesures peut fournir des informations complémentaires et augmenter les performances des systèmes de prise de décision automatisés [Atrey *et al.* (2010)]

Dans [Lahat *et al.* (2015)], les auteurs passent en revue les différents domaines d'application pratique dans lesquels les données multimodales (données provenant de différentes sources) sont disponibles et peuvent être efficacement exploitées pour mieux comprendre les observations. Ces domaines d'application comprennent les systèmes multi-sensoriels, les systèmes biomédicaux et sanitaires et quelques études environnementales.

Le niveau de fusion fait référence au niveau de l'étape consistant à fusionner les informations provenant de différentes sources. Par souci de simplicité, nous allons considérer seulement deux niveaux principaux de fusion : le niveau de fusion précoce et le niveau de fusion tardive. D'un côté, le niveau de fusion précoce correspond à la combinaison des caractéristiques provenant de différentes sources avant l'étape d'apprentissage. La façon la plus simple de le faire est de concaténer les différentes caractéristiques dans un vecteur de caractéristiques unique. Un modèle global est ensuite appris en utilisant ce vecteur de caractéristique. La décision finale est en général obtenue par seuillage direct du résultat de l'algorithme d'apprentissage ou après une étape de calibration qui est destinée à faciliter l'interprétation de la sortie du système. D'un autre côté, la fusion tardive vise à combiner des décisions locales obtenues à partir de sources de données individuelles. Dans ce cas, les caractéristiques des différentes modalités sont utilisées en parallèle pour apprendre des modèles à source unique et leurs sorties combinées pour former la décision finale du système.

La fusion tardive présente de nombreux avantages par rapport à la fusion précoce. En particulier, elle permet d'utiliser des algorithmes d'apprentissage qui sont optimisés pour chaque modalité, avant de combiner les sorties. [Atrey *et al.* (2010)] distingue entre trois familles de méthodes de combinaison : les méthodes basées sur des règles de combinaisons définies à l'avance, celles basées sur la classification et celles basée sur l'estimation.

Application à la détection des lésions épileptogènes

Dans cette section, nous étendons le système de CAD proposé en introduisant deux approches de fusion de données multi-modalités provenant de trois séquences IRM différentes, à savoir T1, FLAIR et DTI. La base de données d'apprentissage utilisée est composée de

38 sujets témoins sains âgés de 20 à 62 ans. La méthode proposée a été testée sur trois patients souffrants d'épilepsie réfractaire (une lésion par patient). Afin de trouver une stratégie de fusion optimale qui exploite efficacement les informations contenues dans les trois séquences IRM, nous comparons deux approches de fusion avec différents niveaux de fusion. La stratégie de fusion précoce est préformée en extrayant cinq caractéristiques des trois modalités (T1, T2 FLAIR et DTI). Pour chaque voxel, un seul modèle OC- SVM (modèle global) est appris. Ceci donne une carte avec un seul groupement. Les groupements sont classés en fonction de la valeur moyenne des distances OC- SVM sur chaque groupement. Pour la stratégie de fusion tardive, nous définissons trois modèles locaux, chacun associé à une modalité (i.e une séquence). Selon le modèle local considéré (T1, FLAIR T2 ou DTI), on entraîne un classifieur OC- SVM pour chaque voxel. Trois cartes de groupements différents sont ainsi trouvés. Pour produire une seule carte de groupement étiqueté, nous utilisons une règle de vote à la majorité. Afin d'obtenir des scores calibrés, l'approche de fusion a été couplée avec des tests statistiques uni-variés dont les scores-z ont été combinés à l'aide de la méthode de Stouffer pour produire un seul score interprétable en termes de probabilité.

Performances et validation contre les résultats SEEG

Pour chaque modèle, les groupements détectés ont été comparés aux anomalies détectées visuellement et aux résultats des autres examens (par exemple FDG-PET, MEG). Les groupements ont été revus par un expert pour désigner vrais positifs et les faux positifs. Les enregistrements SEEG de chaque patient ont, aussi, été utilisés pour identifier les régions du cerveau générant des crises.

Nos résultats montrent que les approches de fusion des trois modèle donnent de meilleures performances de détection, en termes de spécificité et de sensibilité, par rapport aux modèles individuels basés uniquement sur une seule séquence. Nous avons noté aussi que la fusion des scores (stratégie tardive) a donné de meilleurs résultats que la fusion des caractéristiques (stratégie précoce). En particulier, nous avons observé une très bonne corrélation spatiale entre les zones détectées et les zones les plus activées dans les enregistrements SEEG.

Chapitre 8 : Conversion des scores SVDD en probabilités

Motivation

Dans ce chapitre on s'intéresse à la conversion des scores calculés par l'algorithme SVDD en probabilités. Avoir des scores qui correspondent à des probabilités présente plusieurs avantages. En effet, cela améliore l'interprétation de la sortie du système de détection, facilite la combinaison de plusieurs détecteurs et permet l'intégration du détecteur dans un processus plus global.

Dans le contexte de détection de changement, les approches statistiques paramétriques ou non-paramétriques estiment en général la distribution de probabilité à partir d'exemples d'apprentissage. Ainsi, elles permettent d'avoir directement en sortie des scores calibrés en probabilité. Cependant, un grand nombre d'exemples d'apprentissage est nécessaire pour obtenir une bonne estimation de la distribution. En pratique, dans divers applications, seul un petit nombre d'exemples est disponible pour l'estimation et les modèles estimés par approches statistiques ont de faibles capacités de généralisation. Les algorithmes de détection tels que SVDD et OC-SVM permettent d'estimer un bon modèle prédictif même lorsqu'un nombre faible d'exemples est disponible pour l'apprentissage. Cependant, les scores SVDD et OC-SVM ne peuvent pas être interprétés en termes de probabilités à posteriori [Chandola *et al.* (2009), A.F. Pimentel *et al.* (2014)].

État de l'art

Différentes approches ont été proposées dans la littérature pour transformer les scores en sortie d'un algorithme de détection de changement [Gao and Tan (2006), Nguyen *et al.* (2010), Kriegel *et al.* (2011)]. Cependant, dans toutes ces approches, des hypothèses sur la distribution des scores sont formulées et utilisées pour obtenir des scores unifiés.

[Vert and Vert (2006)] ont démontré une propriété théorique intéressante pour les algorithmes OC-SVM et SVDD. D'après cette propriété, la frontière de décision estimée par ce type d'algorithmes correspond asymptotiquement à une courbe de niveau (MV-set) de densité au moins égale à $1 - \nu$, où ν est l'hyper-paramètre de l'algorithme défini dans le chapitre 4.

Généralisation de l'algorithme SVDD

L'algorithme SVDD permet d'estimer une courbe de niveau (MV-set) de densité paramétrée par $\nu = \frac{1}{Cn}$, où n représente le nombre d'exemples d'apprentissage. Nous proposons d'étendre l'algorithme SVDD pour l'estimation de plusieurs courbes de niveaux de densité croissante. Pour cela, deux généralisations de l'algorithme SVDD sont envisagées.

Approche naïve : iSVDD Dans cette première généralisation, on estime q modèles SVDD indépendamment les uns des autres. Chaque model SVDD est associé chacun à une valeur de ν différente. Si on note $\mathbf{a}_j$ et R_j le centre et le rayon du $j^{ème}$ modèle SVDD. Le score iSVDD attribué par ce modèle à une observation $\mathbf{x}$ est donné par :

$$g_j(\mathbf{x}) = (\mathbf{x} - \mathbf{a}_j)^T(\mathbf{x} - \mathbf{a}_j) - R_j^2.$$

Modèles SVDD concentriques : cSVDD Nous proposons d'étendre l'algorithme SVDD pour l'estimation de q courbes de niveaux avec une structure hiérarchique. Pour cela nous estimons q modèles SVDD. La structure hiérarchique est imposée en forçant

les q modèle à être concentriques (*i.e.* avoir le même centre $\mathbf{a}$). Les q modèles SVDD concentriques sont estimés en résolvant le problème d'optimisation global suivant :

$$\begin{cases} \min\limits_{R_j,\mathbf{a},\boldsymbol{\xi_j}} & \sum\limits_{j=1}^{q} R_j^2 + \sum\limits_{j=1}^{q} \dfrac{1}{\nu_j n} \sum\limits_{i=1}^{n} \xi_{ji} \\ \text{s.t} & (\mathbf{x}_i - \mathbf{a})^\top (\mathbf{x}_i - \mathbf{a}) \leq R_j^2 + \xi_{ji}, \qquad i = 1, \ldots, n \ , \ j = 1, \ldots, q \\ \text{and} & \xi_{ji} \geq 0, \qquad\qquad\qquad\qquad\quad i = 1, \ldots, n \ , \ j = 1, \ldots, q. \end{cases}$$

Le score cSVDD attribué par le $j^{ème}$ modèle SVDD concentrique à une observation $\mathbf{x}$ est donné par : $f_j(\mathbf{x}) = (\mathbf{x} - \mathbf{a})^T(\mathbf{x} - \mathbf{a}) - R_j^2$.

Nous démontrons que la formulation duale de ce problème correspond à un problème quadratique qui peut être résolu de façon efficace en utilisant des solveurs standards.

Calibration des scores iSVDD et cSVDD

Nous avons considéré deux fonctions de calibration pour convertir les scores de iSVDD et cSVDD en probabilité. La première fonction correspond à la fonction sigmoïde classiquement utilisée dans le contexte de la classification binaire. Dans le cas d'une classification binaire supervisée, les étiquettes des exemples d'apprentissage des deux classes sont utilisées pour estimer les paramètres de la sigmoïde. Dans le contexte de détection de changement, les exemples d'apprentissage appartiennent tous à la même classe. Par conséquent, les paramètres de la sigmoïde ne peuvent pas être estimés en utilisant uniquement les exemples d'apprentissage. Nous proposons ici de tirer profit de la propriété démontrée par [Vert and Vert (2006)] pour déduire une estimation de la probabilité à posteriori associée à chaque courbe de niveau et donc à chaque modèle SVDD. Les vecteurs supports se situant sur la frontière de décision (la courbe de niveau) peuvent être utilisés pour apprendre les paramètres de la sigmoïde.

La deuxième fonction de calibration correspond à la loi d'extremum généralisée. Dans ce cas la probabilité à posteriori d'être une anomalie est donnée par la fonction de répartition de cette loi. Les paramètres de cette loi sont aussi estimés en considérant les probabilités associées aux différentes courbes de niveau estimées par SVDD.

Expériences et résultats

Nous avons comparé les performances des méthodes proposées à celles obtenues en utilisant directement l'approche statistique d'estimation par noyau (robust kernel density estimator rKDE).

7 jeux de données du dépôt UCI ont été utilisés pour évaluer les performances des différentes méthodes. Nos résultats montrent que la généralisation de l'algorithme SVDD et la conversion des scores en probabilités, a permis d'améliorer les performances de détection par rapport à la méthode statistique rKDE.

Nous avons aussi évalué la robustesse de ces méthodes par rapport à la présence de bruit dans la base de données d'apprentissage. Nos résultats montrent que la méthode iSVDD est la moins robuste à la présence de ce type de bruit.

Conclusion générale et perspectives

L'objectif de ce travail de thèse était de concevoir un système de diagnostic assisté par ordinateur (CAD) capable d'extraire les informations discriminatoires principales à partir de données de neuroimagerie utilisé lors du bilan pré-chirurgical des patients atteints d'épilepsie réfractaire.

Contributions

Une étape importante dans la conception des système de CAD est la modélisation et l'analyse du problème. Les données utilisées dans ce projet consistaient de données de neuroimagerie (principalement des images MR) de patients d'épilepsie réfractaire non étiquetés. Ainsi que des bases de données de sujets sains pour le contrôle. Le système de CAD a pour objectif de produire une carte de cluster étiquetés montrant les zones qui présentent des comportements qu'on pourrait associer à des zones épileptogènes. Pour une meilleure interprétation la sortie obtenue à partir du système de CAD, et afin de faciliter le traitement, les étiquettes des groupements doivent fournir une représentation bien calibrée des scores de suspicion et correspondent parfaitement à la probabilité que le groupement présente une anomalie liée à l'épilepsie.

Compte tenu de ces spécifications, nous avons identifié plusieurs défis dont la plupart découlent de la nature des données de neuroimagerie. En particulier, nous étions amenés à gérer le bruit (par exemple, les artefacts), la grande dimensionnalité des données (environ 1,5 million de voxels dans une séquence MR), la nature multimodale des images (par exemple IRM/PET et/ou de multiples séquences IRM) et la petite taille de l'échantillon d'apprentissage. La gestion du déséquilibre des classes a également été une tâche difficile. Ce déséquilibre est inhérent à la tâche de détection que nous considérons dans cette thèse, car le nombre de données provenant de sujets témoins en bonne santé est supérieur à celui des données de patients annotés, qui s'avèrent très coûteuses. Ce déséquilibre est encore accentué par la petite taille des lésions d'épilepsie réfractaire et leur hétérogénéité. Un dernier défi était de fournir des sorties significatives du système de CAD qui peuvent être combinés avec d'autres tests de diagnostic afin d'aider les experts à prendre leurs décisions finales.

Tout au long de ce travail, nous avons essayé de fonder toutes nos contributions sur notre analyse du problème afin de proposer des méthodes pouvant faire face aux défis identifiés. Notre première contribution était de modéliser le problème comme un problème de détection de changements. Au lieu de construire un classificateur qui permet la discrimination entre les observations pathologiques et les observations saines, nous avons proposé la construction d'une description d'un cerveau sain typique et puis la comparer aux données des patients qui étaient à notre disposition. À cette fin, nous avons proposé un premier système de CAD, basé sur une extension du classifieur séparateur à vaste marges (SVM) adapté au cas d'une seule classe -one-class- (OC-SVM). Le classifieur OC-SVM a été entraîné à l'échelle du voxel à l'aide des caractéristiques extraites à partir de images IRM de témoins sains. La principale motivation pour cette formulation était d'éviter la dépendance à l'égard des données des patients étiquetés et faire face au déséquilibre de classes. Les sorties du système de CAD ont été transformées en scores calibrés par le biais d'une distribution de scores normative estimée en utilisant une procédure *leave-one-out* sur les distributions de scores des témoins sains. Ceci a également permis de contrôler l'erreur de type I à travers le choix d'une valeur de seuil qui correspond à une p-valeur définie par l'utilisateur. Le système de CAD a été évaluée à la fois sur des données de simulation réalistes et des données cliniques de patients. Nous avons comparé ces performances à celle

d'une version optimisée de l'analyse statistique uni-variée statistical parametric mapping (SPM) et avons montré que le système proposé donne de meilleurs résultats. Les comparaisons avec l'état de l'art des méthodes récentes, basées sur des schémas de classification à deux étapes et sur des caractéristiques plus complexes, ont aussi été favorables au système proposé.

Notre deuxième contribution est une reformulation de l'algorithme support vector data description SVDD qui permet de prendre en compte des observations incertaines. Cette reformulation consiste à remplacer le coût Hinge par un coût $(0-1)$ (une pénalité ℓ_0) qui permet de réduire l'effet des observations bruitées sur l'estimation de la description de la classe cible. Une procédure itérative a été adoptée pour résoudre le problème d'optimisation après une relaxation de la pénalité ℓ_0 grâce à une approximation logarithmique. L'algorithme L_0-SVDD proposé a été évaluée sur des données de simulation réaliste basées sur des données cliniques et sur des bases de données référencées parvenant du dépôt UCI. Nos résultats ont montré que, par rapport à la formulation initiale, l'utilisation de la pénalité ℓ_0 réduit avec succès l'effet des observations mal étiquetées.

Notre troisième contribution était d'étudier une stratégie de fusion optimale pour combiner des séquences de données d'IRM multiples. Nous avons proposé deux stratégies de fusion. Une première stratégie, dite précoce, consiste en un seul classifieur global OC-SVM construit à partir des caractéristiques extraites de trois séquences IRM (T1-w, T2 FLAIR et DTI). Puis, une deuxième stratégie, dite tardive, où trois modèles de classifieurs OC-SVM étaient formés séparément, chacun à l'aide des caractéristiques extraites à partir d'une des trois séquences IRM. Les sorties des classifieurs ont ensuite été combinés en utilisant un vote à la majorité. Afin d'obtenir des scores calibrés, l'approche de fusion a été couplée avec des tests statistiques uni-variés reposant sur la méthode de Stouffer. Les stratégies de fusion proposées ont été évaluées sur des données cliniques et validées grâce aux références SEEG (électroencéphalographie intracrânienne). Nous avons observé une bonne corrélation entre les zones cérébrales suspectes identifiées en utilisant la deuxième approche de fusion et les électrodes SEEG les plus activées.

Notre dernière contribution était de proposer un cadre général pour convertir les scores SVDD en des estimations de probabilités. Une stratégie en deux étapes a été proposée. Tout d'abord, nous avons généralisé l'algorithme SVDD pour l'estimation des courbes de niveau de probabilité dotés d'une structure hiérarchique (MV-sets) avec des masses de probabilité spécifiques. Pour cela, nous avons proposé d'estimer des modèles SVDD multiples et de contrôler les masses de probabilité des MV-sets associés *grâce* à la ν-propriété de l'algorithme de SVDD. La structure hiérarchique a été imposée en forçant les hyper-sphères des différents modèles de SVDD à avoir le même centre. Les résultats de la méthode de SVDD généralisée ont, ensuite, été converties en probabilités a posteriori en considérant deux fonctions de calibration; une fonction sigmo'ide ou une loi d'extremum généralisée. Le cadre proposé a été testé sur des données synthétiques et des bases de données ddu dépôt UCI.

Perspectives

Le travail présenté dans les deux premières parties de ce manuscrit est essentiellement dédié à la conception d'un système de CAD pour la détection des lésions d'épilepsie réfractaire. Le système proposé dans le Chapitre 5 est basé uniquement sur des caractéristiques extraites d'images IRM T1. Dans le Chapitre 7, nous avons proposé une extension de ce système qui utilise les caractéristiques extraites de séquences IRM multimodales qui comprenaient des séquences T1, T2 FLAIR et d'IRM de diffusion (DTI). Ceci était une

première étape vers un système CAD réellement multimodal pour la détection des lésions d'épilepsie réfractaire. Une perspective intéressante de ce travail serait de tester le système proposé en utilisant à la fois les différentes séquences IRM et la TEP. La plupart des systèmes CAD proposés dans la littérature (on réfère au Chapitre 2) utilisent exclusivement l'IRM. Seules quelques auteurs ont essayé de développer des méthodes automatisées pour la détection des lésions épileptogènes sur les images TEP. L'analyse conjointe des deux modalités a été limitée à l'inspection visuelle des images TEP après recalage sur l'image IRM T1. Cependant, une constatation intéressante dans ces études est le fait que les zones d'hypo-métabolisme TEP potentiellement associées à l'épilepsie réfractaire sont souvent localisées dans des zones qui correspondent à la matière grise dans l'IRM T. Cette remarque a beaucoup aidé à l'amélioration des résultats de détection visuelle des lésions épileptogènes, surtout dans les cas d'IRM négatives. L'incorporation de ce type d'à priori dans le cadre de l'apprentissage peut aussi aider à l'amélioration des performances des systèmes de détection automatisés. Le cadre proposé dans le chapitre 7 pourrait être facilement adapté pour tenir compte d'un tel à priori. Une façon de faire cela consisterait à modifier la règle de combinaison des scores. Par exemple, le vote par majorité peut être remplacé par un système de vote pondéré qui donnerait plus de poids aux détecteurs qui fournissent des résultats cohérents avec l'à priori de localisation. Bien entendu, ceci peut également être incorporé dans une étape antérieure du système CAD. On pourrait, par exemple, modifier la matrice du noyau de similarité utilisée par chaque classifieur afin de prendre en compte cet à priori.

Dans le Chapitre 6 et le Chapitre 8, nous avons proposé deux extensions intéressantes de l'algorithme d'apprentissage SVDD. Ces deux extensions ont été motivées par notre application clinique. La formulation L_0-SVDD introduite dans le Chapitre 6 visait à réduire l'effet de la présence d' observations mal étiquetées sur la frontière de décision du SVDD. A travers un exemple de motivation, nous avons montré, au Chapitre 6, que ce type de bruit est présent dans l'ensemble des données que nous avons utilisées dans notre étude. Nous avons illustré les avantages de l'utilisation de la nouvelle formulation sur les données de simulation réaliste. Toutefois, Nous n'avons pas appliqué cette méthode sur nos données cliniques. En effet, contrairement aux bases de données UCI, nos données cliniques ne contenaient pas d'observations pathologiques étiquetées pour aider à la sélection du modèle. Comme signalé dans le Chapitre 6, nos tentatives d'optimisation des hyper-paramètres de l'algorithme L_0-SVDD en utilisant l'estimation *leave-one-out* de l'erreur de la classe cible n'étaient guère encourageantes. Une première possibilité est d'envisager de recueillir plus de patients et de demander aux neurologues experts de fournir des étiquettes pour ces nouveaux cas. Une autre possibilité, moins fastidieuse, pourrait être d'utiliser des simulations. Les simulations artificielles (par exemple, en supposant un a priori d'uniformité sur la distribution des valeurs aberrantes *i.e.* n'appartenant pas à la classe cible) peuvent être utilisées pour fournir des exemples de classes ayant de telles distributions. Ensuite, des mesures de performance de sensibilité standard peuvent être utilisées pour optimiser les hyper-paramètres de l'algorithme L_0-SVDD. Une autre alternative serait d'utiliser les simulations réalistes décrites au Chapitre 5 pour valider le modèle et optimiser les hyper-paramètres.

La méthodologie proposée au Chapitre 8 pour convertir les scores SVDD en estimations de probabilité a été évaluée, uniquement, sur des données du dépôt UCI. Une perspective intéressante de ce travail serait d'appliquer cette méthode à nos données cliniques. Le principal défi empêchant l'emploi de cette méthode dans notre application clinique est la taille de la base d'apprentissage. Dans toutes les bases de données de témoins sains que nous avions à notre disposition, le nombre d'observations ne dépassait jamais 40. Ce

nombre n'est pas suffisant pour permettre une estimation précise des courbes de niveau de probabilité avec une structure hiérarchique. En ce qui concerne les données UCI, nous avons utilisé 200 observations pour l'estimation des MV-sets. Dans le cas où les données sont de grande dimension (par exemple dans la base de donnée SPECTF Heart), les performances n'étaient pas satisfaisantes, ce qui suggère le besoin d'utiliser plus d'observations pour l'apprentissage. Pour faire face à ce défi, deux orientations futures peuvent être étudiées. La première direction pourrait consister à essayer de tirer profit de toutes les données de contrôles (témoins sains) disponibles. Grâce à nos collaborateurs, nous avons déjà eu accès à trois bases de données de contrôle sains totalisant, ainsi, 112 images T1 de témoins sains. Cependant, les images disponibles dans ces trois bases de données ne sont pas nécessairement acquises suivant le même protocole d'acquisition et donc ne partagent pas, nécessairement, les mêmes caractéristiques. Les méthodes d'adaptation de domaine permettent d'effectuer l'apprentissage à partir de différentes sources et peuvent être étudiées dans ce contexte en utilisant toutes les bases de données disponibles. La seconde direction pourrait consister à introduire des modèles d'apprentissage au niveau des régions et non pas à l'échelle du voxel. Cela permettrait à la fois d'accroître la taille de la base d'apprentissage et de prendre en compte des informations de voisinage. L'extraction aléatoire de patchs réguliers pour former les régions et utiliser un algorithme OC-SVM pourrait donner lieu à l'estimation d'une frontière de décision qui ne caractérise pas la distribution de la classe normale cible. Cette approche ne pourrait pas permettre de détecter les lésions épileptogènes qui sont, généralement, situées au niveau des frontières (ou bords) des différentes structures du cerveau. L'augmentation de la taille de la base d'apprentissage n'a pas pour but d'ajouter de nouvelles informations qui ne serraient pas déjà présentes dans la base de données initiale, mais plutôt de donner plus de robustesse statistique à l'estimateur. Une possibilité est de regrouper les voxels proches dans au sens d'une certaine métrique de similarité. La sortie de cette étape de regroupement serait des régions homogènes en termes de caractéristiques qui peuvent être utilisées pour estimer les MV-sets. Différentes approches de regroupement peuvent être utilisées. Chacune des méthodes existantes dans la littérature a ses propres hyper-paramètres qui nécessiteraient un réglage particulier, et qui serait encore une fois difficile à réaliser dans ce cadre complètement non-supervisé. L'une des possibilités ici est d'utiliser la métrique introduite pour évaluer les estimations de probabilité dans Chapitre 8 afin de trouver les meilleurs hyper-paramètres pour l'approche de regroupement.

Contents

General introduction

Epilepsy is a neurological disease which affects the nervous system. It is characterized by unpredictable seizures that result from disturbance in the electrical activity of the brain. Most of the time, the origin of the seizures is unknown. Approximately, 65 million people around the world suffer from epilepsy. In 70% of epilepsy cases, antiepileptic drugs allow controlling the seizures [Nagae *et al.* (2016)]. In the remaining cases, seizures are referred to as intractable. For patients with intractable epilepsy, surgical removal of the epileptogenic zone offers the possibility of a cure. The outcome of the surgery relies critically on the clinicians' ability to accurately identify the epileptogenic zone. The pre-surgical evaluation first includes the analysis of electroencephalography (EEG) recordings and of neuroimaging data to obtain an approximate localization of the epileptogenic zone. Then, more invasive intracranial EEG is performed to confirm the lesion location [Duncan *et al.* (2016)]. Thanks to the recent advances in neuroimaging modalities such as magnetic resonance imaging (MRI) and positron emission tomography (PET), neuroimaging data analysis tends to play an increasing role in the pre-surgical evaluation protocol and help greatly with restricting the extent of the area to be explored with the invasive intracranial EEG.

Pattern recognition methods have been successfully applied to neuroimaging data to help clinicians screen abnormalities and improve diagnosis of different pathologies [Norman *et al.* (2006), Klöppel *et al.* (2008), Gray *et al.* (2013), Orrù *et al.* (2012)]. In particular, various computer aided diagnosis (CAD) systems based on the analysis of neuroimaging data with a special focus on MRI data, have been proposed in the past few years to assist in identifying epileptogenic lesions [Antel *et al.* (2003), Duchesne *et al.* (2006), Keihaninejad *et al.* (2012), Hong *et al.* (2014), Ahmed *et al.* (2015)].

In this PhD project, we designed a CAD system based on multivariate and multimodality data analysis. The proposed framework allows extracting discriminative features from neuroimaging data (different imaging modalities and/or sequences) that is usually used as part of the pre-surgical evaluation of patients with intractable epilepsy. For each individual test patient, the output of the CAD system corresponds to a labelled cluster map highlighting suspicious brain areas exhibiting abnormalities associated with intractable epilepsy. The proposed CAD system was evaluated using clinical patient data coming from Lyon's neurological hospital via our collaboration with Pr. F. Mauguière, Dr. J. Jung and Dr. A. Hammers.

Working with neuroimaging data to design such a CAD system presents various challenges. These include handling noisy, high dimensional and multi-modal data, the lack of annotated patient data and the imbalanced nature of the detection task. One of the main contributions of this work is taking into account the specificities of neuroimaging data and proposing efficient ways to deal with the identified challenges. Our first contribution was to formulate the problem of epileptogenic lesion detection as an outlier detection problem to avoid the dependence on labelled patient data and the class imbalance inherent to this detection task. The proposed CAD system builds upon the one-class support vector machines (OC-SVM) classifier. The OC-SVM classifier was trained in a voxel-wise basis to allow handling the high dimensional nature of neuroimaging data and to provide an accurate localization of the epileptogenic zone. The proposed framework was further extended to deal with the remaining challenges. To handle the presence of noise in the training data, we propose a reformulation of the support vector data description (SVDD) algorithm by considering an l_0 cost instead of the original Hinge loss. The new L_0-SVDD formulation is then solved using an iterative procedure providing data specific weighting terms. To deal with the multi-modal nature of the neuroimaging data, an optimal fusion strategy was investigated among two strategies: an early fusion approach consisting in building a single predictive model and a late fusion approach consisting in learning multiple base OC-SVM classifiers each associated with a MRI sequence. Finally, to help with score interpretation we propose to convert the outputs of the SVDD algorithm into well calibrated probabilities. For this purpose, generalizations of the SVDD algorithm and two calibration schemes are proposed.

The manuscript is divided into three parts. In Part I, we first give in Chap. 1 a description of intractable epilepsy and the pre-surgical evaluation protocol. We then give in Chap. 2 a general overview of a CAD system structuring steps and discuss design choices. In light of this description, we review the state-of-the-art CAD systems that were proposed in the context of intractable epilepsy. We conclude Part I by proposing a thorough analysis in Chap. 3 of the diagnostic task at hand given the available inputs and the desired outputs. In Part II, we first describe, in Chap. 4, OC-SVM and SVDD, two one-class learning algorithms that are at the core of the proposed CAD system. We then give in Chap. 5 a detailed description of a first CAD system and evaluate its performance using both realistic simulations and clinical data. In Part III, we propose three extensions of this first CAD system to deal with the remaining challenges. We first propose in Chap. 6 to handle the presence of label noise by extending the SVDD methodology to the case of learning with uncertain training observations. We then investigate in Chap. 7 an optimal fusion strategy for combining multiple MRI sequences. Finally, in Chap. 8, we propose to convert the output of the CAD system into well calibrated probabilities. We conclude this manuscript by overall conclusions and perspectives of the present work.

I Medical and scientific context

Chapter 1

Intractable epilepsy

1.1 Disease description

Epilepsy is one of the most common neurological disorders affecting up to one percent of the population worldwide [Nagae *et al.* (2016)]. Treatment of epilepsy often imposes an exposure to various antiepileptic drugs and requires long-term commitment and compliance from the patient. Despite significant advances over the past 15 years, antiepileptic drugs do not control seizures in a third of epilepsy patients. In this case, seizures are referred to as intractable and in the case of focal epilepsies the patients should be referred for surgical consultation. Temporal lobe epilepsy secondary to mesiotemporal sclerosis and extratemporal lobe neocortical epilepsy secondary to dysplasias are the two most common drug-resistant epilepsies amenable to surgery.

1.1.1 Temporal lobe epilepsy

Temporal lobe epilepsy (TLE) is the most common form of focal epilepsy in adults and children. The underlying pathology is most commonly mesial temporal sclerosis (MTS). There are two main surgical approaches in TLE, selective amygdalohippocampectomy when epileptogenicity seems restricted to mesial temporal structures, and anterior temporal lobectomy, particularly if a neocortical origin is suspected.

1.1.2 Malformations of cortical development

Malformations of cortical development refer to neuronal migrational abnormalities which occur during cortical development. They are often associated with intractable focal

epilepsy. In particular, focal cortical dysplasias (FCDs) [Taylor *et al.* (1971)] are highly epileptogenic lesions that are often associated with intractable epilepsy and are present in up to 25% of patients with focal epilepsy. FCDs are mostly cortical abnormalities; histological subtypes I-III are distinguished [Blümcke *et al.* (2011)]. If the neuronal migration from the subventricular zone to the cortex is disturbed during brain development, other malformations occur. For example, focal subcortical heterotopia are characterized by the presence of neurons (*i.e.* grey matter (GM)) located deep in the white matter (WM); there are also band-shaped heterotopia, and subependymal heterotopia next to the ventricles. The epileptogenic zone is usually centred on the dysplastic cortex. Its surgical removal often leads to freedom from seizures [Fauser *et al.* (2004),Krsek *et al.* (2009),Lerner *et al.* (2009)].

The detection of malformative lesions is complex and challenging as they are very heterogeneous in terms of type, size and location (see examples in Fig. 1.1).

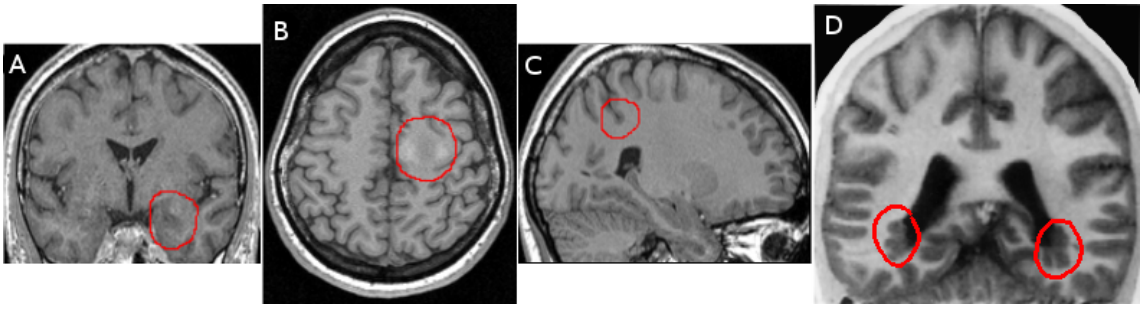


Figure 1.1: Example of three different epileptogenic lesions (highlighted in red) as seen on T1-weighted MRI. A- coronal slice showing a hippocampus anomaly, B- axial slice showing signal and texture change, C- sagittal slice showing an abnormally deep sulcus harbouring an FCD, and D- coronal slice showing bilateral periventricular nodular heterotopia.

1.2 Diagnosis

For patients with intractable epilepsy, surgical removal of the epileptogenic zone (EZ), without causing permanent neurological deficits, offers the possibility of a cure. The outcome of the surgery relies critically on the clinicians' ability to delineate the EZ during the pre-surgical workup of patients [Krsek *et al.* (2009),Lerner *et al.* (2009),Téllez-Zenteno *et al.* (2010)].

The pre-surgical evaluation of intractable epilepsy is divided into two main phases. Phase I consists in the joint analysis of electroencephalography (EEG), video-EEG recordings and neuroimaging data. Phase II consists of recording of intracranial EEG, often via stereotactically guided intracerebral EEG electrode (SEEG) implantation. Fig. 1.2 gives the common pathways for intractable epilepsy pre-surgical evaluation [Duncan *et al.* (2016)].

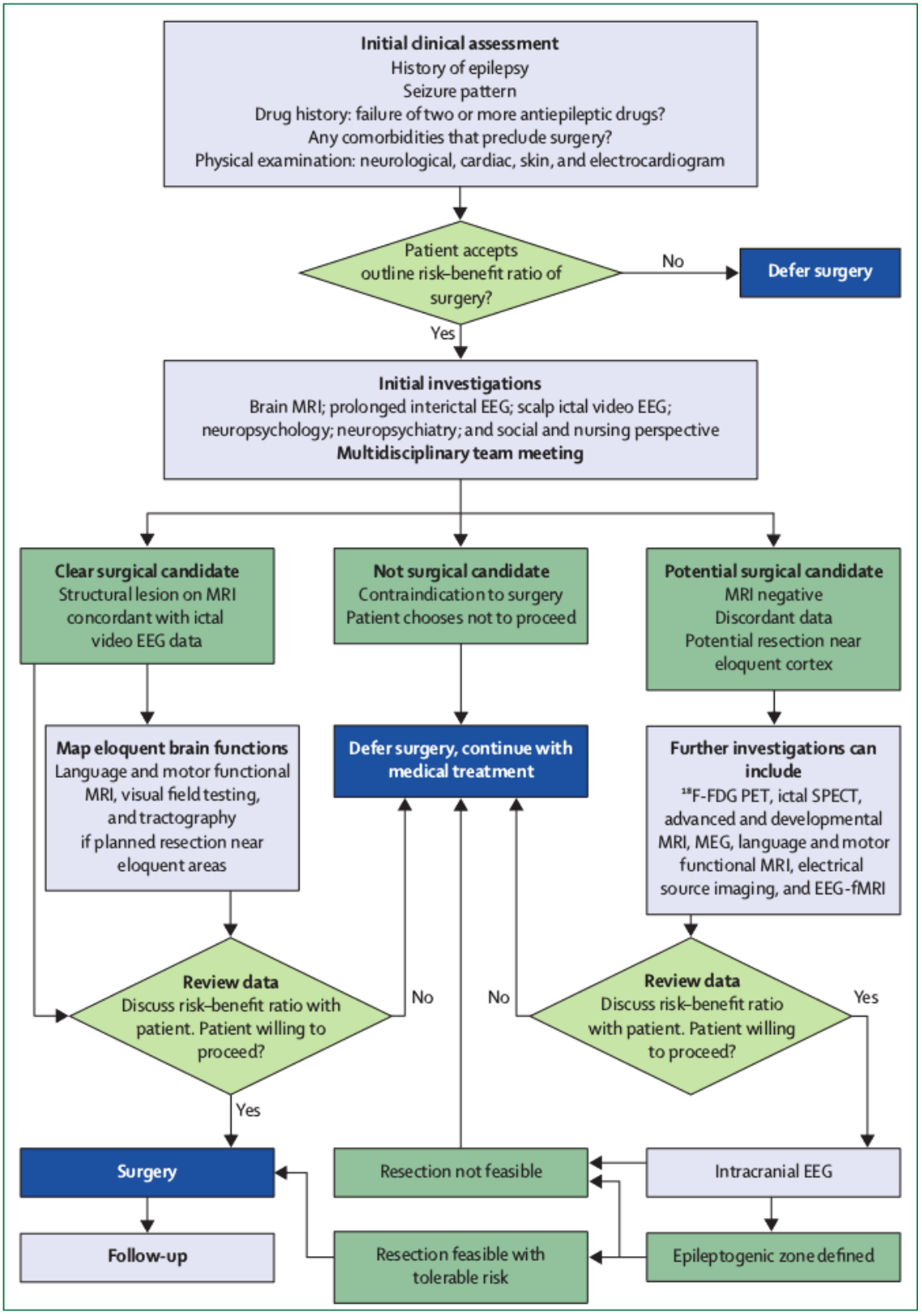


Figure 1.2: The common pathways for intractable epilepsy pre-surgical evaluation [Duncan et al. (2016)].

1.2.1 EEG and video EEG monitoring

Electroencephalography (EEG) is an electrophysiological monitoring method to record electrical activity of the brain. It is typically non-invasive, with the electrodes placed along the scalp. The brain's electrical charge is maintained by billions of neurons. The electric potential generated by an individual neuron is far too small to be picked up by EEG. EEG measurements therefore always reflect the summation of the synchronous activity of thousands or millions of neurons that have similar spatial orientation, this explains the low spatial resolution of EEG (5-9 cm), and the ability to only measure neural activity that occurs in the upper layers of the brain (the cortex). EEG has however a very good temporal resolution (milliseconds).

Continuous EEG monitoring and video monitoring has become a necessary step for all patients undergoing a pre-surgical evaluation [Kelly and Chung (2011), Duncan *et al.* (2016)]. Gathering ictal (during seizures) and interictal (between seizures) EEG data facilitates the localization of seizures. Interictal spikes and focal slowing reflect irritability and dysfunction in the cortex between seizures, while ictal EEG records the seizure onset and its evolution over time and space. Based on video-EEG recordings, epileptologists can distinguish between epileptic seizures and or non-epileptic events, characterize the abnormal electrical brain activity, classify epileptic seizures as focal or generalized in onset and allow an approximate localization of the EZ. This approximate localization is then confronted with the findings obtained using complementary modalities (*e.g.* MRI) that offer higher spatial resolution.

1.2.2 Neuroimaging

What for? Neuroimaging has become a powerful tool for *in vivo* investigation of the brain structure and function. Neuroimaging techniques have been successfully applied to healthy control subjects and patients with neurological diseases to identify possible imaging biomarkers which could be used for early diagnosis, treatment planning and monitoring of disease progression.

In particular, neuroimaging plays an increasingly decisive role in the pre-surgical evaluation of intractable epilepsy. It contributes to outlining the EZ and defining regional eloquent cortex, reducing in some cases the need for or the extent of invasive intracranial EEG (see most right and left paths in Fig. 1.2). It is important to note however that the epileptic network, as defined by EEG, may not exactly correspond to the abnormalities identified using neuroimaging data.

Which type of neuroimaging in epilepsy? The most commonly used imaging methods in the pre-surgical evaluation include magnetic resonance imaging (MRI), positron emission tomography (PET) and magnetoencephalography (MEG). Each modality has its strengths and weaknesses with respect to spatial or temporal resolution and invasiveness. MRI is by far the main neuroimaging technique for the identification of an epileptogenic

lesion. For non lesional epilepsies *i.e.* MRI negative epilepsies (MRI-), further investigation using the other imaging techniques is required to infer the localization of the EZ (see the rightmost path in Fig. 1.2).

Magnetic resonance imaging

Structural MRI allows obtaining high quality anatomical images of the body. It is based on the principle of nuclear magnetic resonance (NMR) and primarily images the NMR signal from hydrogen nuclei which are present in fat and water, highly abundant in the body (70%). During MRI scanning, the subject is placed in a strong static magnetic field. The NMR signal from hydrogen nuclei in the subject's body aligns with the applied magnetic field. This alignment is perturbed by the application of a radio frequency (RF) electromagnetic pulse resulting in the emission of a RF signal. Magnetic field gradients are applied in 3D to encode the emitted RF signals spatially. A more detailed description of this imaging modality and a description of the most commonly used sequences (T1-w, T2-w and FLAIR) are given in the appendix A.

Different tissues can be distinguished by studying the characteristics of the emitted RF signals and also by considering specific MRI sequences. For instance, the time constant T1 characterises the capacity of the imaged tissue to recover the energy transmitted by the RF pulse. This time constant is longer for tissues where the nuclei are free to move such as the water in the CSF and shorter for tissues where the movement is constrained like it is the case in white matter tracts. Consequently, in T1-weighted images (T1-w), we obtain a tissue contrast between the WM, the GM and the CSF. The same analysis can be made considering the T2 time constant.

In the context of intractable epilepsy, MR imaging protocols have been optimized to maximise the potential to identify epilepsy related abnormalities. The imaging hardware has also improved in terms of field strength, coils and gradients. This results in images with an improved signal to noise ratio and a higher spatial resolution. The optimized protocol established by the International League Against Epilepsy (http://www.ilae.org) includes T1-w 3D volume, T2-weighted (T2-w) and FLAIR, acquired with the minimum slice thickness possible. These sequences allow highlighting specific features associated with different types of intractable epilepsy. For instance, 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)].

Fig. 1.3 shows example slices of T1-w and T2-w MRI featuring an FCD lesion. For this FCD lesion, the T1-w images presents a blurred WM/GM interface at the lesion location, whereas the T2-w image presents a hyperintense signal.

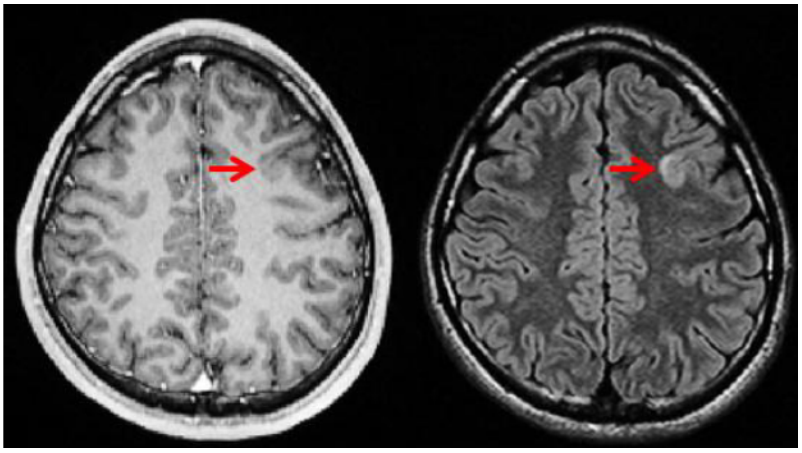


Figure 1.3: Sample T1-w (left) and T2-w (right) axial MRI images taken from a 21-year old male epilepsy patient. The FCD (red arrows) presents as loss of gray-white contrast on T1-w imaging as well as a hyperintensity on T2-w imaging. Illustration from [Kini et al. (2016)].

Diffusion tensor imaging (DTI) has also been investigated in the context of intractable epilepsy detection [Lee *et al.* (2004), Thivard *et al.* (2006), Chen *et al.* (2008), De Carvalho Fonseca *et al.* (2012)]. This MRI sequence measures the diffusion of water molecules to create an anisotropy map. The diffusion of water molecules along the fibers is often highly anisotropic (*i.e.* limited in one direction) whereas the diffusion within the CSF is isotropic. Therefore, using the index of fractional anisotropy allows finding the orientation of the WM tracts. In temporal lobe epilepsy, fractional anisotropy is consistently decreased and for FCD lesions, abnormalities in diffusion indices are present in the subcortical WM adjacent to the lesion [De Carvalho Fonseca *et al.* (2012), Bernasconi and Bernasconi (2015)]. Evidence also suggests that the appearance of DTI tracts can predict postoperative outcome, with displaced tracts recovering more favourably than those infiltrated by the target lesion [Bagadia *et al.* (2011)].

Recent retrospective studies based on surgical epilepsy patients indicate that up to 33% with typical FCD type II lesions and 87% with FCD type I (*i.e.* intracortical) lesions have unremarkable routine MRI [Bernasconi and Bernasconi (2015)]. Similarly, subtle heterotopia may only become apparent after MRI post-processing [Huppertz *et al.* (2009)].

Positron emission tomography

Positron emission tomography is a nuclear medicine technique that is used to observe physiological processes in the body such as metabolism. The basic procedure for a PET scan involves injecting patients with a tracer, labelled with a positron-emitting radionuclide, and then scanning them. A positron emitted inside the body can travel only a short distance through tissue, losing kinetic energy by Coulomb scattering from atomic electrons, until it is almost at rest. When this low energy positron interacts with an atomic electron, the particles can annihilate to produce two gamma ray photons that are detectable outside the body. To conserve energy and momentum, the photons must be

emitted in opposite directions and each with an energy of 511 keV. Since the elements of the PET detector form closed rings around the patient, the two photons are detected simultaneously in opposite detector elements. This process, known as coincidence detection, allows spatial localisation of the tracer in the body and the production of an image showing its distribution.

In the context of intractable epilepsy, the most widely available and clinically used PET tracer is [^{18}F]fluorodeoxyglucose (^{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 EZ, but are generally highly correlated with seizure onset zones and/or areas of seizure spread [Juhász *et al.* (2000), Rathore *et al.* (2014)]. The FDG-PET technique has been used in the pre-surgical planning in intractable epilepsy long before the recent advances in MRI and is still an important investigation method in MRI- focal epilepsies or when MRI and EEG findings are non concordant.

One of the most important contributions of FDG-PET, analysed in conjunction with coregistered MRI, is the detection of FCD with a good surgical prognosis [Salamon *et al.* (2008), Chassoux *et al.* (2010), Hammers (2015)]. Fig. 1.4 illustrates the added value and complementarity of combined FDG-PET and MR imaging. In this example, the FDG-PET alone does not show a clear-cut asymmetry. The MRI alone only shows unusual gyration. The superposition of both images following posthoc coregistration reveals a focal area of hypometabolism, restricted to a single sulcal bottom.

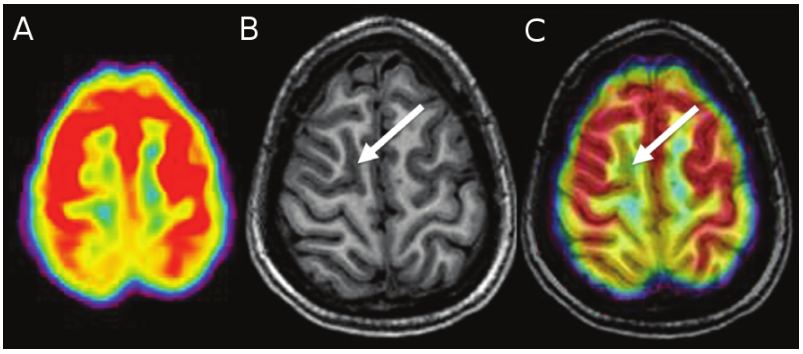


Figure 1.4: Example of the conjunction analysis of FDG-PET and MRI: A- Axial slice of FDG-PET, B- corresponding axial slice of T1-w MRI and C- Superposition of coregistered axial slices of FDG-PET and T1-w MRI. Illustration from [Hammers (2015)].

1.2.3 Intracranial EEG

If the lesion is not seen on imaging or if the imaging and the scalp EEG findings are non-concordant, more invasive monitoring by temporarily placing depth electrodes or other types of intracranial EEG is considered. These procedures, while more invasive, are needed to define the EZ in 20–30% of candidates for epilepsy surgery. They are currently the gold standard for definitively localizing seizure onset [David *et al.* (2011)].

Intracranial EEG allows improving the low spatial resolution obtained with EEG to 1 cm or less. In particular, depth electrodes (thin wires 0.8 mm diameter containing 5

to 18 contacts points ∼2 mm long) can be stereotactically implanted within the desired brain areas to record the electrical activity during epileptic seizures, thus contributing to defining with accuracy the boundaries of the EZ. They allow imaging deep structures such as the hippocampus and the amygdala that cannot be explored using surface or cortical measurements. However, an inherent limitation of depth electrode EEG is the poor spatial sampling as only a limited number of electrodes can be used (*e.g.* 5 to 12 electrodes). At present, standard clinical practice for electrode implantation is a complex task and caries risks. Preoperative planning of electrode trajectories using multimodal imaging can minimise implantation risk by ensuring that the electrodes avoid critical structures such as arteries or veins. Non-invasive neuroimaging methods can also help in maximizing the likelihood of successful SEEG by restricting the extent of the area to be explored with the intracranial electrodes.

1.3 Detection sensitivity and Prognosis

The methods investigated during the pre-surgical evaluation of intractable epilepsy patients in order to localize the EZ have varying detection sensitivity. The performance depends both on the acquisition (*e.g.* the use of an optimized epilepsy protocol as established by the International League Against Epilepsy) and scan interpretation. For instance, in using MRI to detect focal lesions, this sensitivity was 39% when non-optimized imaging was used and reported by non-expert neuroradiologists. Sensitivity in detection was 50% when reported by experts and 91% when optimised acquisition was used and reported by experts [Von Oertzen *et al.* (2002), Duncan *et al.* (2016)].

Sensitivity in detection depends also on the type of epileptogenic abnormalities. MRI is negative in 20–30% of TLE cases and in 34% of FCDs [Nagae *et al.* (2016)]. For MRI-cases, FDG-PET was able to detect the EZ in 85% of TLE cases, and 80% of FCDs (reviewed by [Hammers (2015)]). This performance was further improved after coregistration of FDG-PET with structural MRI (8% gain in performance). Overall, seizure freedom after surgery depends greatly on the presence of a lesion on histopathology or MRI. For lesional epilepsies, the prognosis for seizure freedom is of 69% for TLE and 66% for extratemporal epilepsies. For non-lesional epilepsies, the prognosis is 45% for TLE and 34% for extratemporal epilepsies [Téllez-Zenteno *et al.* (2010)].

Chapter 2

CAD systems for intractable epilepsy

Despite recent improvements in imaging technology and the identification of imaging features associated with intractable epilepsy, the ability to detect the EZ has significant room for improvement. In Sec. 1.3 of the previous chapter, we discussed sensitivity in detection for the various imaging techniques. Overall, surgery success is very much correlated with the detection of a lesion on neuroimaging data and on MRI in particular. Seizure freedom after surgery is almost 2 times higher in cases that exhibit an identifiable lesion on histopathology or MRI. Furthermore, the visual analysis of neuroimaging data is a challenging task. It requires not only having optimized imaging but also skilled epileptologists who are able to identify biomarkers from all available neuroimaging data in order to delineate the epileptogenic zone.

Image processing techniques, especially pattern recognition methods, have been successfully applied to neuroimaging data to help clinicians screen abnormalities and improve diagnosis [Norman *et al.* (2006), Klöppel *et al.* (2008), Gray *et al.* (2013), Orrù *et al.* (2012)]. In particular, computer aided detection systems (CAD) are aimed at identifying suspicious areas in the image and therefore alerting clinicians to these regions during the interpretation of neuroimaging data. The CAD system diagnosis is in general faster or more accurate and reproducible.

In this chapter, we first give a general overview of a CAD system structuring steps, we then discuss choices in CAD design. In light of this general description, we finally give a summary of state-of-the-art CAD systems for intractable epilepsy.

2.1 Pipeline of image-based CAD systems

Image-based computer aided detection systems have been successfully applied in many medical applications including breast cancer detection, lung cancer detection and Alzheimer's disease diagnosis. Despite a wide spectrum of application, these image-based CAD systems share a common architecture. They are typically composed of four main steps:

1. *Image pre-processing*: aimed at improving image quality. It consists of image denoising, artefact reduction and image standardization (or scaling) that allows comparing images obtained under different conditions. Examples of common image standardization steps are intensity scaling and spatial registration.

2. *Object definition and classification level*: defining an object consists essentially in choosing the classification level. For an image-based classification system, the object can either be a voxel of the image, a region of interest (ROI) in the image, a patch from the image, or the whole image. This often allows reducing the size of the data to be analysed and in some cases results in simpler region-specific models. The computational cost should however be balanced with spatial sensitivity and the ability to find features that allow discriminating between the different categories.

3. *Feature extraction:* consists in deriving other measurements from the raw data (initial measurements). The derived measurements are ideally more informative and non-redundant compared with the initial set of measurements. Depending on the classification level, these features are grouped in a feature vector to form a synthetic description of an object. This is aimed at facilitating the subsequent analysis of the data and in some cases results in a deeper understanding of the underlying scientific question.

4. *Classification:* uses the extracted features to infer a predictive model that assigns a label to a given input observation. The model is learned using a training data set. Depending on the nature of this data set, classification can be supervised or unsupervised. The learned model is then used to infer the label of new unseen test observations.

Designing a CAD system requires a thorough analysis of the detection or the diagnostic task at hand. One has to consider, the type of input data (labelled vs unlabelled observations, class imbalance, continuous vs categorical features), *a priori* knowledge about the distribution of the different categories, the characteristics of the target abnormalities and also the type of the desired outputs (scores, labels, probabilities). In the remainder of this section, we discuss important aspects of steps 2-4.

2.1.1 Object definition and classification level

The choice of the classification level depends essentially on the detection task and the type of available inputs. One can distinguish between three main classification levels.

Patient-based classification:

For this classification level, the analysis is made at the patient level and the goal is to discriminate between different groups of subjects *e.g.* patients and healthy controls. This classification level does not allow a precise localisation of target abnormalities and the output is binary. It can still however be used to perform a rough localisation.

ROI-based classification:

For this level of classification the detection model is inferred using features computed from specific regions of interest. The main advantages of using ROI-based classification is 1) to reduce the dimensionality of the input data and 2) to obtain a local representation of the features of interest which then allows a better localisation of target abnormalities. Multiple ways of extracting the ROIs from the input images exist.

Voxel-based classification:

The analysis is made at the voxel level (normal versus pathological) and usually results in a cluster map indicating the most suspicious cluster of voxels. The standard way of performing this type of analysis is to perform a binary classification where the model is learned using features corresponding to normal and pathological voxels in patients' scans.

2.1.2 Feature extraction and selection

A priori knowledge of specific characteristics of observations belonging to different categories (for instance image biomarkers) can help engineering features that are discriminant for the considered classification task. If no such information is available, then a very large amount of features can be extracted from the initial input data. These features are referred to as hand-crafted features. Examples of such features include: textural features [Haralick *et al.* (1973)] based on the computation of grey-level co-occurrence matrices, filters (edge and shape detectors), and robust image descriptors such as SIFT [Lowe (1999)] and HOG [Dalal and Triggs (2005)] that have been introduced for the detection of salient points in the context of object recognition.

Having a very large number of features requires however a large input training data set from which the classification model can be inferred. Often, only a small number of training observations is available, and the learned model is very likely to overfit the training data and suffer from poor generalization performance (*i.e.* the inability to make accurate predictions on new unseen test observations). In the machine learning literature this problem is known as the curse-of-dimensionality or small-n-large-p problem. To deal with this issue a subsequent step of feature selection has to be performed. Feature selection can be thought of as a way of balancing the lack of *a priori* knowledge by retaining only the features that improve classification performance. Many feature selection approaches have been proposed in the context of neuroimaging data. A complete review of these

methods can be found in [Mwangi *et al.* (2014)]. It should be noted however that no study recommends any technique as the best in all neuroimaging machine learning tasks. Therefore, designing and selecting the appropriate features is an important step of the processing pipeline that must be considered carefully.

The main disadvantage of hand-crafted features is that their modelling capacities are limited by the fixed transformations (filters) that stay the same regardless of the different sources of the data. To avoid obtaining poor performances, methods for learning the representations from the data have recently been proposed. The features that are learned using these method are often referred to as data driven features as opposed to hand-crafted features. Examples of such approaches include dictionary learning methods [Mairal *et al.* (2009)] and deep learning approaches. Dictionary learning methods are aimed at learning representative elements from the input data such that each data point can be represented as a weighted sum of the representative elements. Deep learning approaches are based on the idea that an object can be described using a hierarchy of concepts, with higher level concepts building upon lower level ones to offer a higher degree of abstraction [Bengio *et al.* (2013)]. In general, the first layers of a deep learning architecture correspond to low level features that are very similar to hand-crafted features (*e.g.* simple filters). High layers however correspond to more complex concept that are sometimes not easy to interpret. The lack of interpretability and the need for huge datasets are one of the main disadvantages of these type of architecture. In the past few years, much effort have been made to develop several approaches for understanding and visualizing Convolutional Networks [Girshick *et al.* (2014),Zeiler and Fergus (2014),Mahendran and Vedaldi (2015)].

2.1.3 Classification algorithms

The overall goal of classification is to assign a label to a given observation. More formally, let us consider n observations each represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ and K classes C_k with labels $y_i \in \{1, \ldots, K\}$. Each observation is thus represented as one point in the feature space $\mathcal{X}$. A classification algorithm is a mapping function f such that each point $\mathbf{x}_i$ is assigned to a given class C_k:

$$\begin{array}{rccc} f: & \mathcal{X} = \mathbb{R}^p & \longrightarrow & \{1, \ldots, K\} \\ & \mathbf{x}_i & \longmapsto & y_i \end{array} \tag{2.1}$$

The mapping function f is learned by using a training dataset $\mathcal{X}^{tr}$ and optimizing a given performance criterion. In supervised classification, the decision function is learned using a training dataset composed of the n observations $\mathbf{x}_i$ and the corresponding labels y_i ($\mathcal{X}^{tr} = \{(\mathbf{x}_i, y_i | i = 1, \ldots, n)\}$). In unsupervised classification, no labels are presented to the learning algorithm and only the n observations are ($\mathcal{X}^{tr} = \{\mathbf{x}_i | i = 1, \ldots, n\}$), the goal being in this case to discover unknown, but useful, classes of objects. In both cases, after training, the model can be used for mapping new unseen observations.

Most often, the type of function f is chosen beforehand and just a few parameters ($\mathbf{w}$)

of the function have to be determined. To find the optimal parameter $\mathbf{w}^*$ for the function f using the considered training set, an error function $\mathcal{E}$ has to be defined. Multiple choices are possible for the error function including the mean squared error, the 0-1 loss, the hinge loss, etc. The optimal parameter $\mathbf{w}^*$ can be obtained by solving the following minimization problem:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{E}_{\text{true}}(f, \mathbf{w}, \mathcal{X}),$$

where $\mathcal{E}_{\text{true}}$ is defined as:

$$\mathcal{E}_{\text{true}}(f, \mathbf{w}, \mathcal{X}) = \int \mathcal{E}(f(\mathbf{x}; \mathbf{w}), y) p(\mathbf{x}, y) d\mathbf{x} dy,$$

where, $p(\mathbf{x}, y)$ is the true data distribution. In almost all classification problems, $p(\mathbf{x}, y)$ is unknown. In practice it is approximated by the empirical error $\mathcal{E}_{\text{emp}}$ on the training set $\mathcal{X}^{tr}$ defined as:

$$\mathcal{E}_{\text{emp}}(f, \mathbf{w}, \mathcal{X}^{tr}) = \frac{1}{n} \sum_i \varepsilon(f(\mathbf{x}_i; \mathbf{w}), y_i))$$

For this to work, several precautions have to be taken to guarantee good generalization properties. Generalization properties depend greatly on:

- the training dataset: the observations used for estimating the model should in general be a representative sample from the true distribution. Otherwise, minimizing the empirical error $\mathcal{E}_{\text{emp}}$ would result in overfitting the training data and would not guarantee a minimal true error $\mathcal{E}_{\text{true}}$ when tested using an independent test set. It should also be equally representative of the different classes C_k. Ideally, if an infinite number of observations is available for each class C_k, the optimal training set would be the smallest subset of observations that captures most of the information about each class with low redundancy.

- the features: as discussed above, an infinite number of features can be computed from the initial data. These features must however be descriptive of the instances in the different classes and contain enough information to allow accurate discrimination between the classes. Their number should not be too large, because of the curse of dimensionality.

- the structure of the learned function: different families of mapping functions can be considered. For instance, linear discriminant functions, kernel-based functions, logic-based functions (*e.g.* decision trees and random forests) and probabilistic model (Bayes classifier) are the most frequently used models.

- hyper-parameter optimization procedure: most models have some hyper-parameters that have to be chosen before learning the model. The selection of these hyper-parameters and their optimization impacts the generalization performance heavily. A given performance measure has to be chosen for optimizing these hyper-parameters.

This performance measure should avoid selecting models that perfectly fit the training data (*i.e.* over-fit), and only select models that are reliable after deployment in the real use.

In practice, *a priori* knowledge can be very helpful when choosing an adequate model f. When no prior knowledge is available, a relatively complex model f is considered and an extra error term $\mathcal{E}_{\text{struct}}$ is added to the empirical error. This error term tries to capture the complexity of the model $f(\mathbf{x};\mathbf{w})$ and is often referred to as a regularization term. The new cost (or error) that has to be minimized is now given by:

$$\mathcal{E}(f,\mathbf{w},\mathcal{X}^{tr}) = \mathcal{E}_{\text{emp}}(f,\mathbf{w},\mathcal{X}^{tr}) + \lambda\mathcal{E}_{\text{struct}}(f,\mathbf{w}),$$

where λ is a user-defined regularization parameter that sets the relative influence of the structural error with respect to the empirical error. The structural error term does not depend on the training observations. Setting the regularization parameter λ to a very high value would therefore result in very simple models that completely ignore the training observations. In many cases, the structural error is introduced to impose the smoothness of the function f.

A- Supervised classification methods

Much effort has been expended to solve supervised classification tasks. Numerous methods have been proposed which differ in the function $f(\mathbf{x};\mathbf{w})$, the definition of the empirical error $\mathcal{E}_{\text{emp}}$, and the optimization routine for finding the optimal parameter $\mathbf{w}^*$. A comprehensive review of different supervised classification algorithms can be found in [Kotsiantis *et al.* (2007)].

In a nutshell, most classification methods can be grouped in five main categories:

- *Logic-based algorithms*: these algorithms are based on the construction of a set of rules or tests based on feature values. Decision trees [Murthy (1998)] and rule-based classifiers are included in this category. For decision trees, the goal is to find the feature that best divides the training set at each node whereas for rule-based classifiers, the goal is to find the best collection of (if ... then) rules that best divides the training data. Both methods are easy to comprehend and handle numerical and categorical features. However, in practice finding the optimal trees or rule set is an NP-complete problem which requires introducing heuristics and the learned models do not generalise well (overfitting).

- *Perceptron-based algorithms*: a perceptron [Rosenblatt (1962)] is an algorithm for binary classification based on a linear model f combining a set of weights $\mathbf{w}$ with the feature vector $\mathbf{x}$ and an adjustable threshold. The perceptron is usually trained by running the algorithm repeatedly through the training observations until it finds a prediction vector that is correct on all of the training observations. Processing the

observations one at a time makes it an online learning algorithm. As the underlying model is linear, perceptrons can only handle linearly separable classes. Artificial neural networks (ANN) or Multilayered perceptrons (MLP) [Rumelhart *et al.* (1985)] have been proposed to handle non-linearly separable classes. They are brain-inspired learning algorithms that consist of a large number of "neurons" joined together using a connectivity pattern and distributed in three layers: the input layer, the hidden layer and the output layer. Each "neuron" has an activation value that represents a feature. The training data is used to learn the weights of the connections between the different connected "neurons" by comparing the output of the network to the desired label for each training observation. In practice, it is hard to choose the size of the network as underestimating the number of "neurons" can lead to poor generalization properties while overestimating it can lead to overfitting. Depending on the size of the network, a large number of training observations can be required to obtain a good estimate of the optimal weight parameter. These methods require considerable processing and storage resources, however with the recent advances in hardware, many researchers are working on their development.

- *Probabilistic models*: these methods have an explicit underlying probability model. Statistical inference is used to find the best class for a given observation. Unlike other algorithms, which simply output a class label, probabilistic algorithms output a probability of the observation being a member of each of the possible classes. The naive Bayes classifier is the most well-known representative of this category of methods. This classifier assigns a label to a given observation by applying the Bayes decision rule according to which an observation, represented by its feature vector $\mathbf{x}$, is assigned to the class C_k that has the largest posterior probability $p(C_k|\mathbf{x})$. This rule while being theoretically optimal, requires knowing the true posterior probability of all classes C_k and all feature vectors $\mathbf{x}$.

- *Instance-based learning methods*: also referred to as *memory-based learning*, these methods rely on a simple comparison of any new observation to all observations in the training set which have been stored in memory. The k-nearest neighbour (kNN) algorithm [Cover and Hart (1967)] is an example of such methods. In kNN, the label for a test observation is given by the most frequent label among the labels of its k nearest neighbours according to a given distance metric. Various distance metrics and voting strategies can be adopted. These methods while being very simple, have a large storage requirement and are very sensitive to the presence of label noise in the training dataset.

- *Support vector machines (SVM)* [Vapnik (1998)]: are binary classification algorithms that try to find an optimal hyperplane to separate observations from two classes with maximum margin. To deal with classes that are not linearly separable , a soft-margin version of the original algorithm was proposed. Slack variables are introduced to

allow misclassifying some of the training observations at the price of a cost that has to be minimized. For complex datasets that involve highly non-separable classes, the observations can be mapped onto a higher-dimensional space, referred to as the feature space, where a linear separating hyperplane is sought. Deriving the model parameters only involves computing dot products between projected observations in the feature space. Kernels are a special class of function that allows computing dot products in the feature space without knowing the mapping function. This is known as the kernel trick, and one only needs to select an appropriate kernel function to generalize the SVM methodology to obtain non-linear separations in the input space. One advantage of the SVM algorithm over other classification methods is that it necessarily reaches a global minimum and avoids ending in a local minimum.

An empirical comparison of various supervised learning algorithms can be found in [Caruana and Niculescu-Mizil (2006)]. The performance achieved by the different algorithms varies across datasets and depends greatly on the considered performance metrics. Overall, learning methods such as random forests (an ensemble classifier consisting of a collection of decision trees) [Breiman (2001)] and SVMs especially after calibration achieve the best performance.

B- Unsupervised classification methods

Also called clustering algorithms, these methods allow partitioning the data into a certain number of clusters. A cluster is a subset of observations that are similar in terms of a given similarity measure and dissimilar to observations in other clusters. These algorithms operate on unlabelled training observations and allow uncovering hidden structure from the observations. Different clustering methods have been proposed over the years [Xu and Wunsch (2005)]. The three main families of methods are:

- *Hierarchical clustering*: These algorithms organize the data into a hierarchical structure according to a proximity matrix computed using a similarity distance. The resulting hierarchy is usually represented using a dendrogram. The ultimate clustering results can be obtained by cutting the dendrogram at different levels. There are mainly two types of hierarchical clustering algorithms: agglomerative methods and divisive ones. Agglomerative methods start with n clusters consisting each of one observation. Similar clusters are then merged together successively until there is just one final cluster including all observations. Divisive methods proceed in an opposite order, starting with the entire dataset in one cluster and successively dividing it until all clusters are singleton clusters. In practice, divisive methods are very expensive in computation and consequently not commonly used.

- *Squared error-based methods*: These methods assign a set of training observations into K fixed clusters (classes). The optimal partition can in theory be found by enumerating all possibilities, however, due to the computational cost associated with

such an approach, heuristics have been proposed to find an approximate solution of the partition problem. The K-means algorithm [MacQueen *et al.* (1967)] is the best known squared error-based algorithm. This algorithm starts with a random partition and then repeats two main steps until there are no changes to the partition. The first step consists in computing the centroid or prototype of each cluster. The second step consists in assigning observations to the nearest cluster represented by its prototype. This algorithm can be implemented very easily and parallel versions of the algorithm allow speeding the algorithm. However, it requires fixing the number of clusters K beforehand and is very sensitive to the presence of outliers and noise in the training data.

- *Probability density function (pdf) estimation-based methods*: in this case, the training observations are assumed to be generated according to a mixture of probability distributions associated each with a given cluster (or class). Multivariate Gaussian densities are the most commonly used mixtures. In general, maximum likelihood estimates of the mixture parameters are then derived from the training observations using the expectation-maximization (EM) algorithm [McLachlan and Krishnan (2007), Duda *et al.* (2012)]. The major disadvantage for these approaches is that they require fixing beforehand the number of classes K, a probabilistic model for each class, and that the final partition is very sensitive to the initialisation of the EM algorithm.

C- One-class classification methods

The problem in one-class classification is to make a description of a target set of observations and to detect which new observations resemble this training set. Observations that do not fit into the description are called outliers. The difference with conventional classification is that in one-class classification, only examples of one class are available for training the model ($\mathcal{X}^{tr} = \{\mathbf{x}_i | y_i = 1,\ i = 1 \dots n\}$). Depending on the application domain, this problem is also referred to as: outlier detection, anomaly detection, novelty detection, or concept learning. The most straightforward application of one-class classification methods is the detection of unusual observations in a dataset given the description learned from the training observations. Standard classification approaches are not well-suited for this kind of tasks as they only provide reliable class estimates for test observations that resemble the training set. For instance, outlier detection approaches can be used prior to classification to identify and reject outliers in the training dataset. Another case where this type of approaches can be useful is when the class imbalance is very high. If no strategy is considered to deal with class imbalance, a standard classification approach will be biased towards the majority class. To avoid this bias, one can consider the majority class as the target class and construct a descriptive model of this class. Observations from the severely undersampled class are then viewed as outliers with regards to the target class. Like for classification, several models $f(\mathbf{x}; \mathbf{w})$ have been proposed for one-class prob-

lems. Most reviews on the subject classify these algorithms according to five major categories [Chandola *et al.* (2009), A.F. Pimentel *et al.* (2014)]:

- *Probabilistic or statistical methods*: these methods are based on the estimation of the generative probability density function of the training data. A threshold on this density is then set to define the boundaries of the target class. Parametric methods assume that observations from the target class are generated from an underlying parametric distribution. Because of the Central Limit Theorem, the most commonly used form of distribution is the Gaussian [Bishop (1995)]. For more flexibility, the Gaussian distribution has been extended to a mixture of Gaussians: a linear combination of Gaussian distributions. Distribution parameters are derived using the training data and maximum likelihood methods (*e.g.* EM algorithm). Kernel or Parzen density estimators [Parzen (1962)] are a non-parametric generalization of the Gaussian distribution. The estimate of the probability density function is given as a linear combination of Gaussian kernels centered on the individual training observations. The smoothness of the distribution is controlled by choosing an appropriate kernel width, the only parameter of this method. Probabilistic methods require assuming a good probability model and in general having a large number of training observations to overcome the curse of dimensionality. One advantage however is that once a threshold is set, the output of these approaches is a minimum volume for the given probability density model. This advantage will be discussed further in Chap. 8.

- *Distance-based methods*: these include both clustering and instance-based methods (see supervised and unsupervised classification approaches). In the context of outlier detection, instance-based approaches make the assumption that target (or normal) observations occur in dense regions while outliers occur far from their closest neighbours. k-NN [Altman (1992)] is the most commonly used nearest neighbour approach. Several distance measures and strategies for reducing the search domain when dealing with large datasets have been proposed in the context of outlier detection. For instance, [Breunig *et al.* (2000)] proposed computing a measure called local outlier factor (LOC) for each observation. This factor is based on the ratios of the local density of the area around the observation and the local densities of its neighbours. The number of neighbours has however to be specified by the user. Clustering approaches used in the context of outlier detection make the assumption that normal observations lie close to their closest cluster centroid, while outliers are far away from their closest cluster centroid. K-means and its variants can therefore be used to detect outliers by considering the distance of an observation to its nearest cluster centroid as a measure of normality. Besides the disadvantages of these approaches discussed earlier, depending on the type of outliers in the dataset, the assumptions made in the context of outlier detection may not hold. For instance, outliers can potentially have a cluster structure in which case assuming that outliers are isolated observations will lead to false detection results.

- *Reconstruction-based methods*: these methods use prior knowledge about the training observations and make assumptions about the generating process. When test observations are tested using the chosen model, the reconstruction error, the distance between the test observation and the output of the model, can be used to assess the normality of the test observation. Perceptron-based algorithms such as MLP can be used in the context of outlier detection. Subspace-based algorithms, also called spectral analysis algorithms, are another type of reconstruction-based methods. Such algorithms assume that the training observations can be embedded into a lower dimensional subspace in which normal observations and outliers appear significantly different. Principal component analysis (PCA) and kernel PCA [Schölkopf *et al.* (1997), Hoffmann (2007)], that extends the standard PCA to non-linear data distributions (see the kernel trick discussed in the SVM description), are examples of spectral methods.

- *Domain or classification-based methods*: these methods avoid solving the more general problem of estimating the whole density distribution and only focus on deriving the boundary of the target class [Vapnik (1998)]. One-class SVM (OC-SVM) [Schölkopf *et al.* (2001)], the one-class version of SVM, support vector data description (SVDD) [Tax and Duin (2004)] and their variants are examples of such approaches. OC-SVM finds a separating hyperplane that separates the target training observations from the origin of the feature space with maximum margin. Test observations are classified as normal or outliers depending on which side of the hyperplane they fall on. SVDD builds a closed boundary around the target observations by finding the enclosing hypersphere with minimum volume. These methods offer the possibility of deriving non-linear boundaries that have a sparse representation. The main disadvantage of these methods is that outputted scores cannot be interpreted as a probability. Both approaches will be discussed in more detail in Chap. 4.

- *Information theoretic methods*: these methods compute the information content of the training dataset using measures such as entropy, relative entropy, etc. They make the assumption that anomalies introduce irregularities in the information content of the dataset. In general, metrics are computed using the whole training dataset and then the subset of observations whose elimination introduces the largest difference in the metric are identified to be the outliers. For example, the entropy measures the degree of disorder in a given dataset. Outliers can be identified by successively detecting observations with the highest entropy. These methods were mostly used in the context of sequential data analysis. Their performance depends heavily on the choice of the information theoretic measure and its ability to detect the effect of outliers with high sensitivity.

2.2 Performance evaluation

The development of CAD systems goes hand in hand with their evaluation. A variety of metrics can be used to evaluate the performance of a given CAD system. Unifying these metrics allows for a better estimation of the CAD system performance, and facilitates comparison with other CAD systems designed for the same task. In [Petrick *et al.* (2013)], a set of "best-practice" recommendations for assessing CAD systems has been proposed by the "computer aided detection in diagnostic imaging committee".

A number of factors can affect the estimation of the performance of a CAD system. The main factors are:

1. the selection of training and test observations for system training, parameter optimization and performance assessment. In practice, only a random and limited number of observations is available for model training with often an unequal number of observations for each class. This raises the issues of representativeness of the training data; data splitting strategies; and class imbalance. A review of strategies that deal with class imbalance can be found in [Sotiris *et al.* (2006)].

2. ground truth definition (*i.e.* defining the pathological cases, their location and extent) and the labelling rule definition (*i.e.* false positive and true positive definition).

3. performance assessment metrics (*e.g.* accuracy, specificity, sensitivity, ROC curve, etc.)

2.2.1 Data splitting strategies and cross validation

For data splitting strategies, two sets must be derived from all available observations: 1) a model selection set which is in turn divided into a training set used for learning and estimating the optimal parameter $\mathbf{w}^*$ of the model and a validation set used to evaluate the model, usually for model parameter selection; and 2) a model evaluation set or testing set composed of examples used exclusively to assess the predictive performance of the model. This distinction between these two separate sets is essential for not obtaining overly optimistic estimates of the performance achievable by the learned model. To reduce the bias associated with choosing a given partition to select the model, this procedure is usually repeated several times and the performance averaged. The most commonly used strategy is cross-validation. In k-fold cross-validation, the model selection dataset is split into k disjoint folds of the same size, where k is a parameter of the method. In each k turn one fold is used for validation and the remaining k-1 folds for model training. The resulting classification performance is the average of all turns. Leave-one-out (LOO) cross-validation is the special case of k-fold cross-validation where only one observation of the model selection dataset is left out at each turn and used for model validation. This method makes the best use of the data and does not involve any random sub-sampling. Performance estimators obtained using LOO are nearly unbiased but usually have a large

variance [Efron and Tibshirani (1997), Varma and Simon (2006)]. Nested cross-validation has also been investigated to further reduce partition bias. In this case, an outer cross-validation loop is used to split the data into a model selection set and a model evaluation set. An inner cross-validation loop is used to further partition the model selection set into a training set and a validation set [Varma and Simon (2006)].

Bootstrap [Efron and Tibshirani (1994)] is another data splitting or resampling strategy that can be used to partition the data into a model selection set and an evaluation set. It mainly consists in generating distinct data sets, of size equal to that of the original dataset, by repeatedly sampling observations from the original dataset with replacement. One of the main differences between cross validation and bootstrapping is that unlike in the cross validation folds, in a a given bootstrap dataset some observations may appear more than once and some not at all. Using bootstrap allows computing standard errors of an estimate and confidence intervals. In the classification context, the bootstrap datasets can be used to learn the model and the original data to test its performance. This however often results in underestimating the true prediction error as there is a significant overlap between the original dataset and the bootstrap datasets. Different methods have been proposed to take into account this bias and provide a better estimate of the prediction error [Fukunaga and Hayes (1989)].

2.2.2 Ground truth definition and labelling rules

Ground truth for a CAD system refers to the "true" label for each observation. It is often constructed using the results of a gold-standard test or exam. A ground truth definition can be associated with each object definition level discussed in Sec. 2.1.1. For instance, at the patient classification level, the ground truth associated with a given test observation would be either patient case or healthy control subject. A cluster-level ground truth would also include the location and the extent of the target abnormality. Voxel-level ground truth definition is even more challenging, especially when high resolution imaging techniques are considered. In some cases, a gold standard reference exam is lacking. A possibility for establishing the ground truth is by having experts review all available diagnostic information and form a consensus ground truth. For the cluster-level analysis, the performance also depends on a set of labelling rules used to decide which cluster corresponds to the targeted abnormalities. A cluster is considered as a true positive detection if it satisfies the set of labelling rules, and a false positive detection otherwise. Many labelling rules have been used by researchers to evaluate their CAD system performance. Using simulation data can potentially offer a better evaluation of performance provided the simulations are representative of the true abnormalities. In this case, the ground truth can easily be defined and precise labelling rules can be used for performance evaluation and method comparison.

2.2.3 Performance measures (sensitivity, specificity and ROC curves)

The performance of a CAD system can be summarized in a confusion matrix (see Table 2.1). In this example, class C_1 is referred to as the positive class and class C_2 as the negative class. The confusion matrix depends on the definition of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) observations.

		Predicted class	
		C_1	C_2
True class	C_1	True Positive	False Negative
	C_2	False Positive	True Negative

Table 2.1: Confusion matrix for a binary (C_1 versus C_2) classification problem.

The number of observations that were correctly labelled by the classifier is reported in the diagonal of the matrix. TP corresponds to observations belonging to the first class (C_1) that were correctly labelled. TN correspond to observations from the second class (C_2) that were correctly labelled. FP correspond to observations belonging to class C_1 and that were labelled C_2, and FN correspond to observations that belong to class (C_1) and were incorrectly labelled C_2.

Different metrics can be computed from the confusion matrix:

- *Accuracy*: This is the proportion of observations that were correctly labelled by the classifier.

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

  This metric does not always provide a good estimate of the CAD performance, especially if the class distribution of the testing dataset is imbalanced. For instance, the classifier that always predicts the majority class will always have a high accuracy.

- *Sensitivity*: also called true positive rate, measures the proportion of positives (*i.e.* observations from class C_1) that are correctly identified as positives.

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

- *Specificity*: also called true negative rate, measures the proportion of negatives (*i.e.*

observations from class C_2) that were correctly labelled.

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

Sensitivity and specificity measures are less sensitive to class imbalance than the accuracy measure. An optimal CAD system should have a sensitivity and specificity of 100%. In practice however, there is a trade-off between sensitivity and specificity. The best operating point of the CAD system, the point that offers the best compromise, can be found by constructing a receiver operating characteristic (ROC) curve.

Receiver operating characteristic (ROC) curve: this is typically used to asses the performance of CAD systems. It operates on the scores outputted by the CAD system prior to thresholding and label assignment. To construct the ROC curve, the threshold is varied to cover the entire range of possible scores, and sensitivity is plotted as a function of the false positive rate (FPR) expressed as: $(1 - \text{Spe})$.

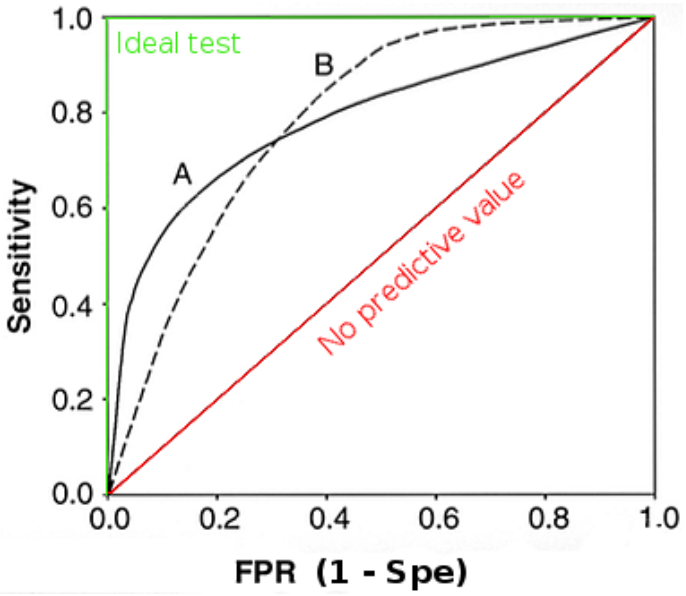


Figure 2.1: ROC curve examples for: an ideal system (100% sensitivity and specificity) in green, a system with chance performance and therefore no predictive value in red, and two systems (A and B) with an identical AUC value but different best operating points.

Figure 2.1 gives examples of ROC curves corresponding to four detection systems. At a given threshold, the couple (Sen, 1-Spe) is called an operating point of the system. Different methods have been proposed to fit empirical ROCs to smoothout undesirable artefacts such as non-convexity and to make it easier to derive other measures from the ROC curve. The area under the ROC curve (AUC), partial AUC [Jiang *et al.* (1996)], and specific pairs of operating points can be used to summarize the performance of a CAD system. Replacing the ROC curve by a single value alone (*e.g.* the AUC value) sometimes introduces bias when comparing different systems [Hand (2009)]. Systems A and B in Fig. 2.1 have an equal AUC value but the two systems do not have an identical

performance. In the high sensitivity range, system B is better than system A, whereas in the low false positive rate range (high specificity range) system A is better than system B. In general, for clinical applications with highly imbalanced datasets or cost-sensitive systems, the AUC value does not give a reliable estimate of the system performance. It should therefore be accompanied by other measures of performance.

2.3 State-of-the-art CAD methods for intractable epilepsy

Given the various aspects of CAD system design discussed in the previous sections, we classified the state-of-the-art studies that deal with the detection of intractable epilepsy related lesions in Tab. 2.2, 2.3 and 2.4.

2.3.1 Object definition and classification level

Patient-based classification level has been used to infer the lateralization of the epileptogenic focus in the case of temporal lobe epilepsy [Duchesne *et al.* (2006), Focke *et al.* (2012), Keihaninejad *et al.* (2012)]. The binary output in this case was: lesion or EZ located in the left versus right medial temporal lobe.

Regarding FCD detection, most studies used voxel-based methods to perform a comparison between a patient and a cohort of normal subjects [Bernasconi *et al.* (2001), Huppertz *et al.* (2005), Colliot *et al.* (2006), Thesen *et al.* (2011)]. These methods generally require registering patient and control scans to guarantee voxel correspondence between the different scans. Voxel based morphometry (VBM) [Ashburner and Friston (2000)], the most commonly used method, consists of registering structural images of the brain into a reference space and making spatial comparison within the statistical parametric mapping (SPM: fil.ion.ucl.ac.uk/spm; Wellcome Trust Centre for Neuroimaging) framework which performs an analysis of variance at each and every voxel of the images. As a mass univariate approach, this method tends to produce many false positive detections and requires correcting for multiple comparisons. In [Besson *et al.* (2008), Hong *et al.* (2014), Ahmed *et al.* (2015)], patch and vertex-based classification levels were investigated. The authors in [Hong *et al.* (2014), Ahmed *et al.* (2015)] argue that voxel-based methods do not optimally characterize morphology as they neglect anatomical relationships across the folded cortex. As an alternative, they proposed vertex-based methods that rely on surface-based morphometric features. This however results in restricting the target epileptogenic abnormalities (and therefore reducing model complexity) as they are most suited for detecting abnormalities that are located at the surface of the cortex. In the context of intractable epilepsy detection,

2.3.2 Feature extraction

In the context of intractable epilepsy detection, nearly all of the time knowledge-guided neuroradiological markers have been considered. As part of the pre-surgical evaluation of

intractable epilepsy patients, neuroradiologists examine neuroimaging data and look for certain distinct markers in order to identify the EZ. A description of these imaging markers has been given in Chapter 1. Several groups have tried to translate these findings into computable features with a special focus however on MRI findings (see Tab. 2.2, 2.3 and 2.4).

- Signal intensity changes associated with epileptogenic lesions are assessed by using grey-level values from T1-w, T2-w or FLAIR images after intensity standardization [Duchesne *et al.* (2006), Focke *et al.* (2012), Riney *et al.* (2012), Cantor-Rivera *et al.* (2015)].
- Increased cortical thickness associated with malformations of cortical development and in particular with FCD is assessed by measuring the radial distance between WM and GM surfaces [Antel *et al.* (2003), Srivastava *et al.* (2005), Thesen *et al.* (2011)].
- GM/WM junction blur is modelled via the computation of a gradient map after convolution with a Gaussian kernel [Antel *et al.* (2003), Thesen *et al.* (2011), Ahmed *et al.* (2015)].
- Sulcal and gyral abnormalities are assessed using surface-based measures such as gyrification index, curvature, and sulcal depth [Thesen *et al.* (2011), Hong *et al.* (2014), Ahmed *et al.* (2015)].

To better highlight theses changes, most features are transformed into a z-score map computed with respect to a nominal distribution constructed using features from healthy control subjects. Textural maps derived from grey-level co-occurrence matrices including angular second momentum, difference entropy and contrast, have also been investigated as features for detecting FCD lesions [Antel *et al.* (2003)] .

2.3.3 Classification

Not many machine learning classification approaches have been used to build CAD systems for the detection of epilepsy related abnormalities. [Focke *et al.* (2012), Keihaninejad *et al.* (2012)] used SVM for the lateralization of TLE lesions. [Antel *et al.* (2003)] used two cascaded Bayesian classifiers for the detection of FCD lesions.

In the neuroimaging community, statistical regression-based methods have been more extensively investigated than classification models. Regression-based methods attempt to explicitly model the relationship between inputs or independent variables and the outputs, typically in the form of parametric equations in which the parameters are estimated from the data. These methods often provide explicit estimates of measures of association between individual inputs and the outcome, adjusted for other inputs, with standard error estimates provided from the modelling paradigm used. The most common class of regression methods in the literature comes from the class of generalized linear models [McCullagh

and Nelder (1989)], which includes linear regression and logistic regression. In regression, the output is either a continuous (linear regression) or categorical variable (logistic regression) as opposed to the categorical labels obtained in classification. The general linear model is a generalization of linear regression where more than one independent variable and dependent variable are considered. Once a model is fitted using the training data, post-hoc inferences on the dependent variables of interest are made with a standard mass univariate statistical test resulting in statistical scores of T or F values for each independent variable. Dedicated software implementing various regression models has been developed for the analysis of neuroimaging data. Statistical parametric mapping (SPM) software is the most commonly used in the neuroimaging community. Linear discriminant analysis (LDA) [Fisher (1936), McLachlan (2004)] is a statistical approach that resembles logistic regression as it also explains a categorical variable by the values of continuous independent variables. It tries to find the best projection of labelled training observations that minimizes the intra-class variance while maximizing the inter-class variance. It assumes that the independent variables are normally distributed.

The general linear model (GLM) is the most used regression method. It is often used, in an unsupervised or one class setting, to test a single patient against a cohort of healthy control subjects. In the neuroimaging community this model is often referred to as voxel-based morphometry (VBM) [Srivastava *et al.* (2005), Bruggemann *et al.* (2007), Chassoux *et al.* (2010), Thesen *et al.* (2011)]. In two recent studies LDA and logistic regression were also used to detect FCD lesions [Hong *et al.* (2014), Ahmed *et al.* (2015)].

2.3.4 Performance evaluation

Leave-one-patient-out cross validation was used in most classification studies [Antel *et al.* (2003), Duchesne *et al.* (2006), Concha *et al.* (2012), Focke *et al.* (2012), Hong *et al.* (2014), Ahmed *et al.* (2015), Cantor-Rivera *et al.* (2015)]. For TLE detection studies, accuracy was used to evaluate system performance (see Tab. 2.2). For FCD detection, sensitivity and specificity were used for performance evaluation (see Tab. 2.3 and 2.4). In some cases, specificity was measured by considering a leave-one-patient-out on the controls only.

In the context of intractable epilepsy, the gold standard reference exam for defining the epileptogenic zone is intracranial EEG. Still, different ground truth definitions have been considered. In most cases, outputted cluster maps are visually reviewed by an expert to check for co-localization with the known or inferred location of the abnormality. In other cases, true positive clusters are defined as clusters overlapping with the surgical resection zone for patients with good surgical outcome [Hong *et al.* (2014), Ahmed *et al.* (2015)]. As there is no way of checking the truth of a cluster detected outside of the resection zone, false positive detections are often evaluated by considering the proportion of detections in healthy control subject scans. In practice, the resection area contains both lesional and nonlesional tissue. Defining the ground truth based on the resection area alone would

therefore introduce bias in performance evaluation. In the two studies by [Hong *et al.* (2014), Ahmed *et al.* (2015)], a mask reduction step consisting of the joint analysis of the resection area and texture maps or cortical measures was used to define the ground truth. Improved performance was obtained after the mask reduction step.

2.3.5 Summary of CAD systems for TLE

The studies in Tab. 2.2 are mainly focused on the diagnosis of temporal lobe epilepsy (TLE). In this diagnostic task, a precise localization of the epileptogenic focus is not required. It is therefore easier to obtain labelled training datasets for training supervised classification algorithms. Overall, the accuracy in the lateralization of the epileptogenic foci depends greatly on whether hippocampal sclerosis (HS) is present. In case of HS, 100% accuracy can be reached [Duchesne *et al.* (2006), Keihaninejad *et al.* (2012)]. Distinguishing TLE patients from normal controls has also been investigated [Focke *et al.* (2012), Keihaninejad *et al.* (2012), Cantor-Rivera *et al.* (2015)]. In this diagnostic task as well, the presence of HS allowed for a better distinction between TLE patients and controls reaching an overall accuracy of 89-96% versus 86% without HS.

2.3.6 Summary of CAD systems for FCD

CAD systems for the detection of malformations of the cortical development mainly focused on detecting FCDs. Compared with the diagnostic tasks addressed in the context of TLE, the detection of FCDs is a more challenging task as it requires finding the exact location of the lesion. Most studies in Tab. 2.3 and 2.4 used the GLM model to test on a voxelwise basis one single patient against a cohort of normal controls. The main limitation of these approaches is that they only perform mass univariate tests that often take into account a single feature (effect). Finding the location of the EZ requires in this case examining the results of multiple statistical maps, one for each considered feature. Only one study [Bruggemann *et al.* (2007)], tried to take into account more than one effect of interest and proposed the use of a conjunction model to test both the effect of the GM and WM maps.

Only a few studies used supervised classification algorithms [Antel *et al.* (2003), Hong *et al.* (2014), Ahmed *et al.* (2015)]. This is mainly due to the difficulty of labelling pathological observations. The scarcity of pathological observations is partly due to the difficulty of defining a gold standard on these examples. In most cases, the labelling is performed by considering all neuroimaging findings for lesional epilepsies and the resection zone for non-lesional ones. From a learning point of view, the subjectivity associated with labelling observations introduces label noise that must be taken into account either by using robust learning methods or by carefully choosing the data splitting strategy. Additionally, labelled observations may not fully represent the physiopathological heterogeneity of epileptogenic abnormalities as well as the image pattern variability that is likely to be observed for any lesion type depending on its location in the cortex or white matter. To

deal with this model complexity, the tree studies proposed to focus only on a certain type of epileptogenic lesions, namely FCD subtype II and to proceed in two steps for detecting the FCD related abnormalities.

It is also worth noting that the most commonly used imaging modality is MRI, with a special focus on T1-w sequences. In two studies [Focke *et al.* (2008), Riney *et al.* (2012)], using the FLAIR sequence allowed improving the performance in detecting FCDs.

Table 2.2: State-of-the-art methods for TLE detection.
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.* (2006)]	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, linR 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%

Table 2.3: State-of-the-art methods for FCD detection – part I

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: 14/16 (WM GM), 10/16 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

Table 2.4: State-of-the-art methods for FCD detection – part II

Study	Data	Imaging	Ground Truth	Object definition	Features	Classifier	Evaluation	Results
[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
[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
[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
[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

Chapter 3

Problem analysis

In the previous chapters, we first discussed the medical context of this work and gave a description of the target pathologies. In Chap. 2, we then discussed different aspects of CAD system design and gave an overview of the previously proposed CAD systems in the context of epilepsy lesion detection. Going beyond simply proposing another CAD system for intractable epilepsy, in this chapter we propose a thorough analysis of the diagnostic task at hand given the available input data and the desired output. This problem analysis can be assimilated to drafting the specifications for this research work. Figure 3.1 gives a schematic representation of the objectives and challenges of the present work. In light of this analysis, in the remainder of this chapter, we give a brief description of the different contributions that we propose to meet the specifications for this project.

3.1 Objectives & Challenges

The objective of this project is to design a CAD system that can extract discriminative features from multi-modal data consisting of the different imaging modalities and/or sequences that are usually used during the pre-surgical evaluation of patients with intractable epilepsy. For each individual test patient, the desired output of the CAD system should be a labelled cluster map highlighting suspicious brain areas exhibiting abnormalities associated with intractable epilepsy (see Fig. 3.1). For a better interpretation of the CAD system output, cluster labels should represent well calibrated suspicion scores and ideally correspond to the probability of the cluster being an epilepsy related abnormality.

Another important aspect that must be taken into account in the problem analysis is the data available for building and evaluating the CAD system.

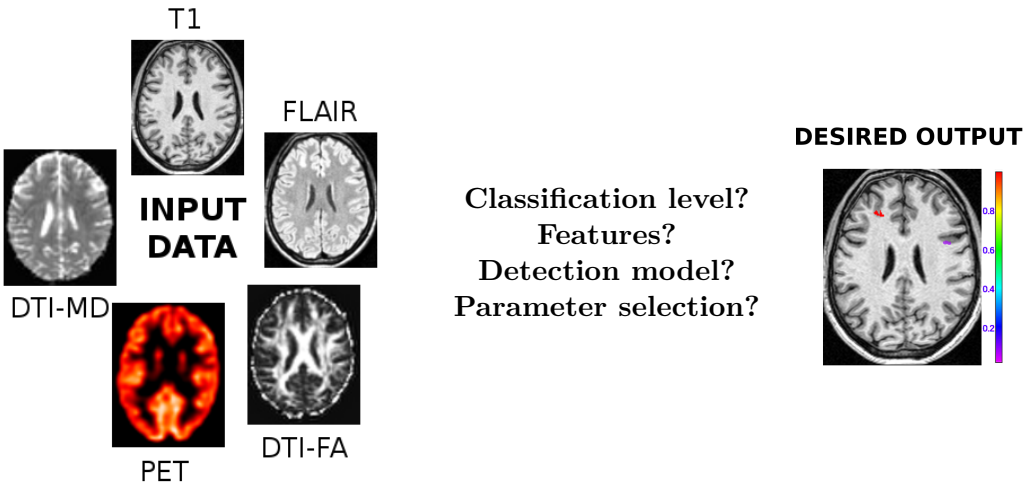


Figure 3.1: Problem analysis: inputs, design choices and outputs.

- **Patient data**: in this project, we have access to unlabelled patient data coming mainly from Lyon's neurological hospital either via our collaboration with Dr. A. Hammers (a first group of five patients, patient database PDB1) or as part of an ongoing research program PHRC (programme hospitalier de recherche clinique) initiated by Pr. F. Maugière and Dr. J. Jung (a second group of patients, PDB2). This research program is aimed at evaluating the diagnostic value of using multimodal neuroimaging data in the pre-surgical evaluation of intractable epilepsy. The associated clinical database will expectedly comprise 100 patients with intractable epilepsy; however for this project only twelve were available. All patients will systematically undergo an imaging protocol that includes both FDG-PET and MRI (T1-w, FLAIR and DTI sequences).
- **Healthy control data**: through our collaboration with the CERMEP, the imaging platform that was used to obtain patient scans, we have access to three healthy normal control databases. For the first database (normal database NDB1), the imaging protocol matches the first group of patients in PDB1 and consists only of T1-w MRI images of 37 healthy subjects. As part of the PHRC research program, two healthy control databases were acquired. The first database (NDB2) consists of MRI (T1-w, FLAIR and DTI) images of 40 healthy control subjects. The second database (NDB3) consists of 35 healthy control subjects who underwent both MRI (T1-w, FLAIR and DTI) and FDG-PET exams. Imaging protocols used in both NDB2 and NDB3 match the ones used for the second group of patients in PDB2.

An overview of the characteristics of the different databases that were used in this project is given in Tab. 3.1. The imaging protocols used in each database will be detailed later in the experiment sections in the next chapters.

Given these specifications, various challenges can be identified:

1. Class imbalance: this results from the higher availability of healthy control data

Database	Nb subjects	1.5T MRI			PET	SEEG
		T1-w	FLAIR	DTI		
PDB1	5	✓	x	x	x	x
NDB1	37	✓	x	x	x	NA
PDB2	12	✓	✓	✓	✓	✓
NDB2	40	✓	✓	✓	x	NA
NDB3	35	✓	✓	✓	✓	NA

Table 3.1: Characteristics of patient and control databases used in this project.

compared with annotated patient data. Annotated patient data is very expensive to obtain as it requires a manual labelling of the pathological observations by an expert based on imaging data or histology. Class imbalance is further accentuated by the small size of intractable epilepsy abnormalities in comparison with the whole brain volume. In particular, for focal lesions, lesion size in voxels does not exceed 1% of the entire brain volume (*e.g.* 1.5 million millimetric cubic voxels).

2. Label noise: this arises mainly from subjectivity in manually delineating the epileptogenic zone either after examination of neuroimaging data by considering the resection zone. In general, the resection zone can also include normal tissue. Labelling the whole resection zone as pathological would therefore result in assigning false labels to the normal tissue within the resection zone.

3. High dimensional data: the lack of *a priori* knowledge about the potential localization of the target epileptogenic abnormalities makes it very hard to restrict the analysis to given areas of the brain. Due to the relatively high resolution of neuroimaging data used in epilepsy protocols, investigating the whole brain would require handling high dimensional data and can be very costly computationally.

4. Multi-modal data: as part of the pre-surgical work-up, intractable epilepsy patients undergo several imaging exams. MRI (T1-w, FLAIR) and PET are the most commonly used ones. The definition of the EZ is based on the examination of all available neuroimaging data. Each modality or sequence has its own characteristics (dimension, spatial resolution, contrast) that help capturing given features of epilepsy related abnormalities and provide complementary information that can help localize the EZ. Taking into account all modalities requires in general using multi-variate approaches.

5. Feature noise: this can result from artefacts present in the raw images or from errors in the pre-processing of the images and feature extraction steps, for instance, during image registration and segmentation. The challenge here is to find a way to reduce the impact of noisy training observations on the learned model.

Some of the methods presented in the-state-of-the-art overview presented in Chap. 2 addressed some of these challenges. In [Antel *et al.* (2003), Hong *et al.* (2014), Ahmed *et al.* (2015)] supervised approaches were used. Class imbalance was addressed by using all available pathological observations and proposing a sampling strategy to select a subset of observations from the non-pathological class [Antel *et al.* (2003)] or randomly selecting an equal-sized sample of normal observations [Hong *et al.* (2014), Ahmed *et al.* (2015)]. In the other studies, class imbalance was avoided by using parametric statistical models and the GLM that only require data from one class to infer the model parameters. In the supervised classification settings [Hong *et al.* (2014), Ahmed *et al.* (2015)], label noise was dealt with by restricting the definition of the ground truth. Pathological observations were extracted by defining a mask that restricts the resection zone to areas that also show an abnormal texture feature or abnormal cortical thickness measure according to a given threshold, while non-pathological observations were extracted only from healthy normal control subjects. In [Antel *et al.* (2003), Hong *et al.* (2014), Ahmed *et al.* (2015)], the high dimensional nature of neuroimaging data was addressed by first extracting the cortical surface and adopting a vertex-based classification scheme. Sampling strategies adopted for reducing the class imbalance also helped with reducing the overall number of observations (*e.g.* 10 000 to 50 000 vertices instead of 1.5 million voxels).

It should be noted however that the last two challenges, the multi-modal nature of the data and feature noise, were not specifically addressed by the state-of-the-art methods and that no approach that combines all imaging sequences or modalities has been investigated.

3.2 Our contributions

Our first contribution is to formulate the problem of epileptogenic lesion detection as an outlier detection problem. This formulation was mainly motivated by the lack of labelled patient data generally and also in the datasets provided by our expert collaborators. This also alleviates the need to address the first challenge (*i.e.* class imbalance), discussed above. Training a model using an outlier detection (or one-class classification) algorithm only requires having observations from one class and avoids obtaining a model biased toward the majority class. In our case, observations from healthy normal control subjects scans were used to train the model. Healthy control subjects scans were all previously reviewed and controls presenting significant structural abnormalities had been excluded from the healthy control databases. From all possible outlier detection algorithms, we chose one-class support vector machines (OC-SVM) and support vector data description (SVDD) as classifiers. These classifiers were trained on a voxel-wise basis to allow handling the high dimensional nature of neuroimaging data, to guarantee an accurate and precise localization of the epileptogenic zone, and to avoid estimating complex models due to the anatomic complexity of the brain. The features used to train the model were extracted from MRI T1-w images only. The CAD system was validated and evaluated using both realistic simulations and patient data from the two datasets PDB1 and PDB2 (see Tab. 3.1). To

reduce the computational cost associated with training a model per voxel, the CAD system was implemented using a parallel distributed computing architecture and was deployed on a cluster. This first contribution is explained in detail in Chap. 4 and Chap. 5.

In this first CAD system, challenges 2 and 4 were not specifically addressed. The proposed framework was therefore further extended. To handle the presence of noise in the training data (label noise), we propose a reformulation of the support vector data description (SVDD) algorithm where the L_1 penalty term in the primal problem was substituted by an L_0 penalty term. We demonstrate that the resulting L_0-SVDD problem can be solved using an iterative procedure providing data specific weighting terms. The details related to this contribution are given in Chap. 6.

In Chap. 7, an optimal fusion strategy for combining multiple base OC-SVM/SVDD classifiers is investigated to deal with the multi-modal nature of neuroimaging data. Two data fusion levels were considered. The early fusion approach consists in building a single global model learned using features extracted from all imaging modalities; the late fusion approach consists in building local models associated each with a single imaging modality and then combining their outputs. In our experiments, we tried fusing information extracted from three MRI sequences, namely T1-w, FLAIR and DTI. The results of the CAD system were validated against SEEG findings and suggest that the best strategy for this detection task is the late fusion approach.

Finally, in Chap. 8, we propose to transform the outputs of the CAD system into well calibrated probabilities to help with score interpretation, threshold selection and score combination. A two step strategy is proposed. We first generalize the SVDD method by reformulating the associated problem to estimate nested probability level-sets and then use a calibration function to convert the outputted scores into well-calibrated probability estimates. Two calibration functions are proposed: the sigmoid function classically used in binary classification problems, and the generalized extreme value distribution, much more suited for long-tailed probability distributions that can be encountered in the context of outlier detection. Hyper-parameter selection is performed by optimizing the quality of the predicted level set using cross-validation. Optimizing the quality of the probability estimates instead of the detection performance (*e.g.* accuracy, AUC) has the great advantage of not requiring examples from the second class to select model parameters.

II A building block CAD system

Chapter 4

Outlier detection

The problem of outlier detection consists in learning a description of a target class of observations and detect which new observations fit into the learned description and which do not. The main difference with standard classification is that only example observations of one class are available and used to estimate the model. One-class classification approaches have been successfully applied in many application domains, with a special focus on applications where labelling outliers is prohibitively expensive. The most popular application domains are: fraud and intrusion detection, medical anomaly detection (*e.g.* detection of anomalous records) and image processing (*e.g.* pattern recognition tasks).

In Chap. 2, Sec. 2.1.3, we gave an overview of the five main categories of one-class classification algorithms previously proposed in the literature. In this chapter, we will focus on two domain based algorithms: the one-class support vector machines (OC-SVM) algorithm proposed by [Schölkopf *et al.* (2001)] and its variant the support vector data description (SVDD) algorithm proposed by [Tax and Duin (2004)]. The choice of OC-SVM and SVDD algorithms was not arbitrary. Both algorithms allow constructing flexible descriptions of the target data in a multivariate way. Unlike probabilistic models, OC-SVM and SVDD do not require making any assumptions on the distributions of the target observations and the potential outliers and can give good generalisable descriptions with relatively few training observations. When compared with distance-based methods (*e.g.* k-NN and LOC), these algorithms are less sensitive to the presence of noise in the training observations and have a lower memory requirement as only a fraction of the training observations are used to describe the target class domain. Additionally, OC-SVM and SVDD testing phases are fast compared with the other approaches since they only require comparing the test observation with the pre-computed model. A last advantage of these

approaches is that they allow an explicit control of the fraction of training observations that can be excluded from the target class domain definition, thus allowing a control of the type I error (*i.e.* the false positive rate).

4.1 OC-SVM: primal and dual formulations

Principle The one-class SVM methodology proposed by [Schölkopf *et al.* (2001)] is a special case of the SVM algorithm [Vapnik (1998)] for assigning labels $y_i \in \{-1, 1\}$ corresponding to two distinct classes of objects, based on n training samples $\{\mathbf{x}_i, y_i\}_{i=1,\dots,n}$, $\mathbf{x}_i \in \mathbb{R}^p$ from the positive (normal) class only. The learning samples are first mapped into a higher dimensional space (called the feature space) via a feature map ϕ associated with a kernel K such as: $K(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$, where $(- \cdot -)$ denotes the inner product. In this feature space, a linear separation between the training observations and the origin of the feature space is sought. The kernel is chosen as to maximize the separability of the training data from the origin of the feature space. Learning sample separability from the origin is guaranteed when using kernels for which properties in equation 4.1 hold.

$$K(\mathbf{x}_i, \mathbf{x}_i) = 1 \text{ and } K(\mathbf{x}_i, \mathbf{x}_j) > 0. \tag{4.1}$$

In equation 4.1, the first condition implies that all examples lie on the surface of the unit sphere in the feature space. The second condition implies that all mapped examples lie within some orthant (all inner products are positive). An example of such a kernel is the radial basis function (RBF or Gaussian) kernel where: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$. It should be noted however that these are sufficient but not necessary conditions for the separability in the feature space. This essentially amounts to solving a binary SVM classification problem where, the origin is treated as the only member of the negative class and therefore the objective is to find the hyperplane of equation $(\mathbf{w} \cdot \phi(\mathbf{x})) - \rho = 0$ that separates the learning examples (normal examples) from the origin with maximum margin.

Primal formulation Analogous to the support vector classifier [Vapnik (1998)], [Schölkopf *et al.* (2001)] propose to use an l_2 regularization term (structural error) and a hinge loss (see Fig. 4.1) for the empirical error. The total error which has to be minimized is then:

$$\begin{aligned} \mathcal{E}_{\text{OC-SVM}}(\mathbf{w}, \rho, \mathcal{X}^{tr}) &= \mathcal{E}_{\text{struct}} + \lambda\, \mathcal{E}_{\text{emp}} \\ &= \tfrac{1}{2}\|\mathbf{w}\|^2 - \rho + \tfrac{1}{\nu n} \sum_{i=1}^{n} max(0, -((\phi(\mathbf{x}_i) \cdot \mathbf{w}) - \rho)), \end{aligned}$$

where $\|.\|$ corresponds to the norm in the feature space (and should be noted $(\mathbf{w} \cdot \mathbf{w})$).

Unlike the 0-1 loss, the hinge loss function in Fig. 4.1 is convex. However it is not differentiable at 0. Slack variables ξ_i associated with each training observation $\mathbf{x}_i$ are introduced to make the minimization problem differentiable and therefore easier to optimize.

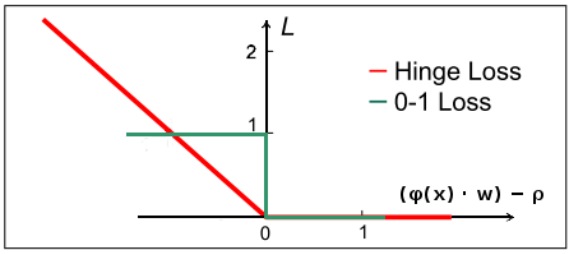


Figure 4.1: Comparison between the hinge loss function used in OC-SVM and the 0-1 loss.

The slack variables are used to replace the $(\phi(\mathbf{x}_i) \cdot \mathbf{w}) - \rho$ and are constrained such as $\xi_i \geq max(0, -((\phi(\mathbf{x}_i) \cdot \mathbf{w}) - \rho))$.

This results in the following primal problem for OC-SVM:

$$\begin{cases} \min\limits_{\mathbf{w},\rho,\boldsymbol{\xi}} & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum\limits_{i=1}^{n}\xi_i \\ \text{subject to} & (\phi(\mathbf{x}_i) \cdot \mathbf{w}) \geq \rho - \xi_i, \qquad i \in [1, n] \\ \text{and} & \xi_i \geq 0, \qquad\qquad\qquad i \in [1, n], \end{cases} \tag{4.2}$$

$\frac{1}{\nu n}$ is the associated regularization parameter that controls the trade-off between model complexity (regularization term) and the number of errors (empirical error). The decision function (or the OC-SVM boundary) is given by the function: $f(\mathbf{x}) = sgn((\mathbf{w}.\phi(\mathbf{x})) - \rho)$, where sgn is the sign function (*i.e.* $sgn(z)$ equals to 1 if $z > 0$ and -1 otherwise). This function will be positive for most training observations $\mathbf{x}_i$, and the normal class domain corresponds to regions where f takes positive values.

Dual formulation The primal problem in Eq. 4.2 is often solved by considering its dual formulation and using Karush-Kuhn-Tucker (KKT) optimality conditions. In case of convex and differentiable objective function and constraints, KKT conditions are necessary and sufficient for finding the optimal solution to the problem.

To derive the dual formulation of problem 4.2, we introduce a Lagrange multiplier for each constraint. Let $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ be the positive Lagrange multipliers associated with the inequality constraints $(\phi(\mathbf{x}_i) \cdot \mathbf{w}) \geq \rho - \xi_i$ and $(\beta_1, \beta_2, \ldots, \beta_n)$ be the positive Lagrange multipliers associated with constraints $\xi_i \geq 0$.
The Lagrangian of problem 4.2 is :

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

The optimal parameters can be found by minimizing the Lagrangian with respect to the primal variables $\mathbf{w}$, ρ and $\boldsymbol{\alpha}$ and maximizing it with respect to the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Setting to 0 the partial derivatives of the Lagrangian with respect to the primal vari-

ables gives:

$$
\begin{aligned}
&\bullet \quad \nabla_{\mathbf{w}} L = 0 \quad \Rightarrow \quad \mathbf{w} - \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) = 0 \qquad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) \quad (*) \\
&\bullet \quad \tfrac{\partial L}{\partial \rho} = 0 \quad \Rightarrow \quad -1 + \sum_{i=1}^{n} \alpha_i = 0 \qquad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i = 1 \\
&\bullet \quad \nabla_{\boldsymbol{\xi}} L = 0 \quad \Rightarrow \quad \tfrac{1}{\nu n} \boldsymbol{e} - \boldsymbol{\alpha} - \boldsymbol{\beta} = 0, \qquad \Rightarrow \quad \boldsymbol{\alpha} = \tfrac{1}{\nu n} \boldsymbol{e} - \boldsymbol{\beta}
\end{aligned}
$$

$$\text{where } \boldsymbol{e} = [1, ..., 1]^T \in \mathbb{R}^n.$$

Using these three conditions to substitute the primal variable $\mathbf{w}$ in the Lagrangian gives (after simplification) the following dual problem:

$$
\begin{cases}
\min\limits_{\boldsymbol{\alpha}} & \frac{1}{2} \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \\
\text{subject to} & \sum\limits_{i=1}^{n} \alpha_i = 1 \\
\text{and} & 0 \leq \alpha_i \leq \frac{1}{\nu n} \quad i \in [1, n],
\end{cases}
\tag{4.3}
$$

where K is the matrix of inner products whose elements are $K_{ij} = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$. The inequality constraints on α_i are obtained by considering the relation between α_i and β_i given by setting to 0 the derivative of the Lagrangian with respect to ξ_i and the positivity constraint on all Lagrange multipliers (α_i and β_i).

The resulting dual problem in Eq. 4.3 is a standard quadratic programming (QP) problem with box constraints. This dual formulation is much easier to solve than the primal. In the primal, we must minimize over $\mathbf{w}$, ρ and $\boldsymbol{\xi}$, with two inequality constraints per training observation $\mathbf{x}_i$. In the dual, we must minimize over $\boldsymbol{\alpha}$ with one box constraint per observation and one simple equality constraint. Off-the-shelf optimization algorithms can be used for solving the QP problem in Eq. 4.3. However, due to the simplicity of the constraints, optimized algorithms can be used to efficiently solve this problem with a better time complexity.

After solving the dual problem, the optimal $\boldsymbol{\alpha}^*$ can be used to compute the optimal primal variables. The expression for the primal variable $\mathbf{w}^*$ is directly given by condition $(*)$. Condition $(*)$ also gives the representer theorem stating that the optimal classifier function that minimizes the total error can be written as: $f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* (\phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i))$. The two other primal variables, namely the bias ρ^* and the slacks ξ^*, can be found by using the complementary slackness KKT condition or by deriving the bi-dual problem of problem 4.2 and identifying the primal variables (see Appendix C for a more detailed explanation of both methods).

Three categories of training observations can be distinguished based on the Lagrange multipliers $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

- normal observations: correspond to observations for which the first inequality con-

straint of the primal problem is strictly satisfied $(\phi(\mathbf{x}_i)\cdot\mathbf{w}) > \rho$ and $\xi_i = 0$. In terms of the dual variables, this corresponds to having: $\alpha_i = 0$ and therefore $\beta_i = \frac{1}{\nu n}$.

- essential support vectors (Ess_SVs): correspond to observations lying on the decision boundary *i.e.* $(\phi(\mathbf{x}_i)\cdot\mathbf{w}) = \rho$ and $\xi_i = 0$. This corresponds to having: $0 < \alpha_i < \frac{1}{\nu n}$ and $0 < \beta_i < \frac{1}{\nu n}$.
- non-essential support vectors ($errors$): correspond to observations that were allowed to be outside of the normal description domain. For these observations, the slack variables are strictly positives $\xi_i > 0$ and α_i are equal to the upper bound $\alpha_i = \frac{1}{\nu n}$.

It is worth noting that only the essential and non essential support vectors have a non-zero α_i Lagrange multiplier and are therefore sufficient for defining the decision boundary (see condition $(*)$).

The $\nu-$property In [Schölkopf *et al.* (2001)], the authors showed that the parameter ν that balances the structural and the empirical error actually corresponds to an upper bound on the fraction of permitted errors and a lower bound on the fraction of support vectors.

$$\frac{\#errors}{n} \leq \nu \leq \frac{\#errors + \#Ess\_SVs}{n}.$$

4.2 SVDD: primal and dual formulations

Principle The SVDD algorithm proposed by [Tax and Duin (2004)] allows defining model $f(\mathbf{x};\mathbf{w})$ which gives a closed boundary around a target (normal) dataset. Like for the OC-SVM algorithm, the training observations are first mapped into a higher dimensional space via a feature map ϕ also associated with a kernel K. In the feature space, the hypersphere with a minimum volume that contains most of the training data is sought. This hypersphere is also sometimes called the minimum enclosing ball. The hypersphere is characterized by its radius R and center $\mathbf{a}$ (see Fig. 4.3).

Primal formulation [Tax and Duin (2004)] propose also to use an l_2 regularization term and a hinge loss for the empirical error. Like for the OC-SVM case, to obtain a differentiable objective function, slack variables ξ_i are introduced.

The resulting primal minimization problem is:

$$\begin{cases} \min\limits_{R,\mathbf{a},\boldsymbol{\xi}} & \underbrace{R^2}_{\text{structural error}} + C \underbrace{\sum_{i=1}^{n}\xi_i}_{\text{empirical error}} & \\ \text{subject to} & ((\phi(\mathbf{x}_i)-\mathbf{a})\cdot(\phi(\mathbf{x}_i)-\mathbf{a})) \leq R^2+\xi_i, & i\in[1,n] \\ \text{and} & \xi_i \geq 0, & i\in[1,n]. \end{cases} \tag{4.4}$$

The term $((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a}))$ corresponds to the distance between observation $\mathbf{x}_i$ after projection and the hypersphere center $\mathbf{a}$. For simplicity, we will note this term as: $\|\phi(\mathbf{x}_i) - a\|^2$. The decision function (SVDD boundary) is given by: $f(\mathbf{x}) = sgn(\|\phi(\mathbf{x}_i) - \mathbf{a}\|^2 - R^2)$.

Dual formulation The constraints in problem 4.4 can be incorporated into the objective function by introducing the positive Lagrange multipliers $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ associated with the first n inequality constraints and $(\beta_1, \beta_2, \ldots, \beta_n)$ associated with the inequality constraints on the slack variables ξ_i.

The Lagrangian is then given by:

$$L(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(R^2 + \xi_i - \|\phi(\mathbf{x}_i) - \mathbf{a}\|^2) - \sum_{i=1}^{n}\beta_i\xi_i$$

Satisfying the KKT stationarity conditions at the optimum gives:

- $\nabla_{\mathbf{a}} L = 0 \quad \Rightarrow \quad -2\sum_{i=1}^{n}\alpha_i(\phi(\mathbf{x}_i) - \mathbf{a}) = 0 \quad \Rightarrow \quad \mathbf{a} = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x}_i) \quad (**)$
- $\frac{\partial L}{\partial R} = 0 \quad \Rightarrow \quad 2R(1 - \sum_{i=1}^{n}\alpha_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\alpha_i = 1$
- $\nabla_{\boldsymbol{\xi}} L = 0 \quad \Rightarrow \quad C\boldsymbol{e} - \boldsymbol{\alpha} - \boldsymbol{\beta} = 0, \quad \Rightarrow \quad \boldsymbol{\alpha} = C\boldsymbol{e} - \boldsymbol{\beta}$

where $\boldsymbol{e} = [1, ..., 1]^T \in \mathbb{R}^n$.

Using these conditions, the Lagrangian can be simplified to obtain the dual formulation of SVDD:

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^T K \boldsymbol{\alpha} - \boldsymbol{\alpha}^T diag(K) \\ \text{subject to} & \sum\limits_{i=1}^{n}\alpha_i = 1 \\ \text{and} & 0 \leq \alpha_i \leq C \qquad\qquad i \in [1, n], \end{cases} \tag{4.5}$$

where $diag(K)$ corresponds to the diagonal of the matrix K of elements

$$K_{ij} = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j).$$

Like for OC-SVM, the dual SVDD problem in 4.5 is a quadratic programming problem for which standard solvers exist. After solving this dual problem, the optimal $\boldsymbol{\alpha}^*$ can be used to compute the optimal primal variables $\mathbf{a}^*$, R^* and $\boldsymbol{\xi}^*$. The relation between $\boldsymbol{\alpha}$ and $\mathbf{a}$ is given by condition $(**)$. The radius R^* can be deduced either by considering the complimentary slackness KKT condition or by deriving the bi-dual formulation of problem 4.5 (see Appendix C).

The training observations can be classified into three types depending on the Lagrange

multipliers values. When an observation $\phi(\mathbf{x}_i)$ lies inside the hypersphere its Lagrange multiplier will be equal to zero ($\alpha_i = 0$). If the observation lies on the hypersphere boundary (*i.e.* when the distance between the observation and the hypersphere center **a** equals the radius R) then its Lagrange multiplier satisfies $0 < \alpha_i < C$. When the observation is rejected by the data description (outlier), its Lagrange multiplier reaches its upper bound ($\alpha_i = C$). Fig. 4.2 illustrates these properties using a toy 2D distribution of data points.

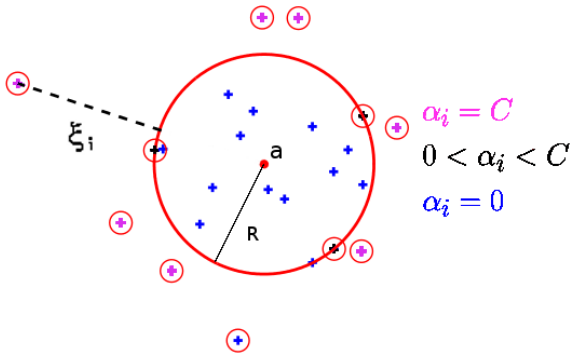


Figure 4.2: Classification of the training observations according to their Lagrange multiplier value. Data points in blue correspond to observations that lie inside the hypersphere. Data points with the red circle correspond to support vectors. Essential support vectors (in black) lie exactly on the boundary while errors (in magenta) are outside.

4.3 Comparison OC-SVM / SVDD

Figure 4.3 illustrates the decision boundaries of both OC-SVM and SVDD after projection of the training data observations into the feature space using a mapping function ϕ. In this case the kernel function associated with ϕ satisfies the properties in Eq. 4.1.

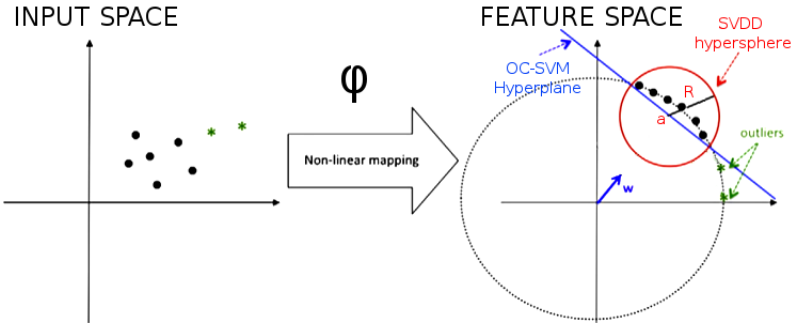


Figure 4.3: Principle of OC-SVM and SVDD.

Despite similar formulations, especially when considering the dual formulations in 4.3 and 4.5 that only differ by a linear term in the objective function, OC-SVM and SVDD do not always give the same descriptions. The SVDD algorithm looks for a closed boundary around the data (a hypersphere) while the OC-SVM tries to find a separating hyperplane

which does not necessarily give a closed boundary around the data. However, if the data is preprocessed to have a unit norm or when a kernel function that implicitly rescales the data to have a constant norm is used, the two algorithms give identical solutions. The Gaussian kernel is an example of such kernel functions.
More formally, let us consider a kernel function K such that $K(\mathbf{x}_i, \mathbf{x}_j)$ only depends on $\mathbf{x}_i - \mathbf{x}_j$. This implies that $K(\mathbf{x}_i, \mathbf{x}_i) = c$, where c is a constant. The linear term in the objective function of the dual SVDD problem in 4.5 is therefore also a constant and we do not need to minimize it. Indeed, we have: $\boldsymbol{\alpha}^T diag(K) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) = c \sum_{i=1}^{n} \alpha_i = c$ by using the equality constraint on α_i. Hence, the two dual problems become identical for hyper-parameters C and ν that verify $C = \frac{1}{\nu n}$. We can also show that the two decision functions are also identical in this case (see Appendix D).

4.4 Hyper-parameter optimization

Both approaches have one user defined hyper-parameter: ν for OC-SVM and C for SVDD. C and ν play an important role in balancing the two types of errors in the objective function. For instance, setting the C parameter of the SVDD algorithm to an infinite value (*i.e.* giving an infinite weight to the structural error term in the objective function) will result in forcing all the training observations to be inside the decision boundary and therefore allowing zero errors. When decreasing the value of C, respectively increasing the value of ν, the Lagrange multipliers α_i are more constrained, and due to the equality constraint on the α_is, more and more observations become support vector and the error on the target class increases (more training observations are rejected).

If a kernel function is used, other hyper-parameters have to be tuned. The Gaussian kernel ($K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$) is by far the most commonly used kernel. OC-SVM and SVDD both perform best when this kernel is used. The Gaussian kernel has one hyper-parameter, σ the width of the Gaussian. This parameter influences highly the decision boundary. For small values of σ, $K(\mathbf{x}_i, \mathbf{x}_j) \approx 0, \forall i \neq j$ and therefore all observations become support vectors with an equal weight $\alpha_i = \frac{1}{n}$. Increasing the value of σ, will result in fewer support vectors and therefore a more rigid hypersphere boundary.

In Chap. 2, we discussed the importance of tuning the model hyper-parameter. Cross-validation and associated data splitting strategies must be carefully performed to avoid overfitting the training observations and obtaining too complex models with poor generalization properties. In a standard classification setting, the hyper-parameters are searched over a grid to find the optimal values that minimize a given cross-validation classification error (*e.g.* AUC, TPR, FPR). In outlier detection, observations from only one class are available for training and selecting the model. In this setting, it is not possible to optimize standard classification performance measures and the risks of over-fitting the training observations is even harder.

[Tax and Duin (2004)] propose to compute an estimate of the target error by con-

sidering leave-one-out cross-validation. They showed that this error is bounded by the fraction of observations that become support vectors (essential and errors). Varying the model hyper-parameters in a given range then allows picking the hyper-parameters that minimize the target error estimate or that allow obtaining a specific target error. Of course for this to work, the training observations must be a representative sample from the true target distribution. It should be noted however that minimizing this criterion does not guarantee good detection performance on the outlier class. [Schölkopf *et al.* (2001)] propose to take advantage of the ν-property of the OC-SVM for tuning the hyper-parameters. This value can be set for obtaining a given upper bound on the fraction of errors.

Knowing the influence of the different hyper-parameters on the decision boundary can give the user some insights into how to choose these parameters. Still, hyper-parameter tuning in outlier detection settings remains an open problem. Recently [Xiao *et al.* (2014)] proposed two methods for selecting the Gaussian kernel width for OC-OSVM. The main idea is to find a measure of the compactness of the decision boundary and its tightness using information on the distance between observations and their neighbours (farthest and nearest) and/or by considering the decision values for midpoint observations for pairs of training samples. Fig. 4.4 illustrates the principle of the method proposed by [Xiao *et al.* (2014)] for measuring boundary tightness.

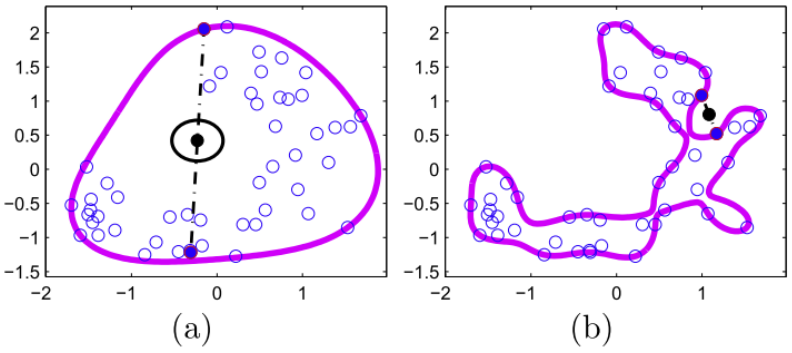


Figure 4.4: Example of a (a) loose OC-SVM boundary and (b) tight OC-SVM decision boundary. In both cases a Gaussian kernel was used [Xiao et al. (2014)].

Chapter 5

Application to epileptogenic lesion detection

In this chapter, we propose a flexible classification method combining textural maps [Huppertz *et al.* (2005)] relevant for FCD and heterotopia associated abnormalities and a one-class support vector machine (OC-SVM) [Schölkopf *et al.* (2001)] that is trained with 'negative' (normal) examples from a control database only. The model then allows the detection of outliers (i.e. abnormalities) in the test group on a voxelwise basis. The OC-SVM method has been successfully applied in many application domains, but rarely in the field of pattern recognition using neuroimaging data [Mourão Miranda *et al.* (2011), Sato *et al.* (2012)]. The main contribution of this work is a voxel-level machine learning system that performs outlier detection in neuroimaging data based on multivariate features adapted to capture the specificity of malformative epileptogenic lesions. We hypothesize that the OC-SVM framework overcomes the main limitations of the mass univariate statistical analysis by allowing 1) to better control outliers within the learning step, 2) to easily incorporate multiple feature maps in a multivariate analysis without any assumption on the feature statistical distribution and 3) opening the way for the integration of spatial *a priori* information that enables learning from the voxel and its neighbourhood.

We evaluate the CAD system on synthetic data and on a patient group, including a quantitative performance analysis against manually annotated epileptogenic lesions. Finally, we compare the OC-SVM scheme against the mass univariate SPM analysis optimised for this application. Our results indicate that the OC-SVM scheme outperforms the SPM analysis and competes favourably with more sophisticated systems based on a

two-step image processing.

Fig. 5.1 schematizes the different steps of the proposed CAD system. The steps are further described in the remainder of this chapter.

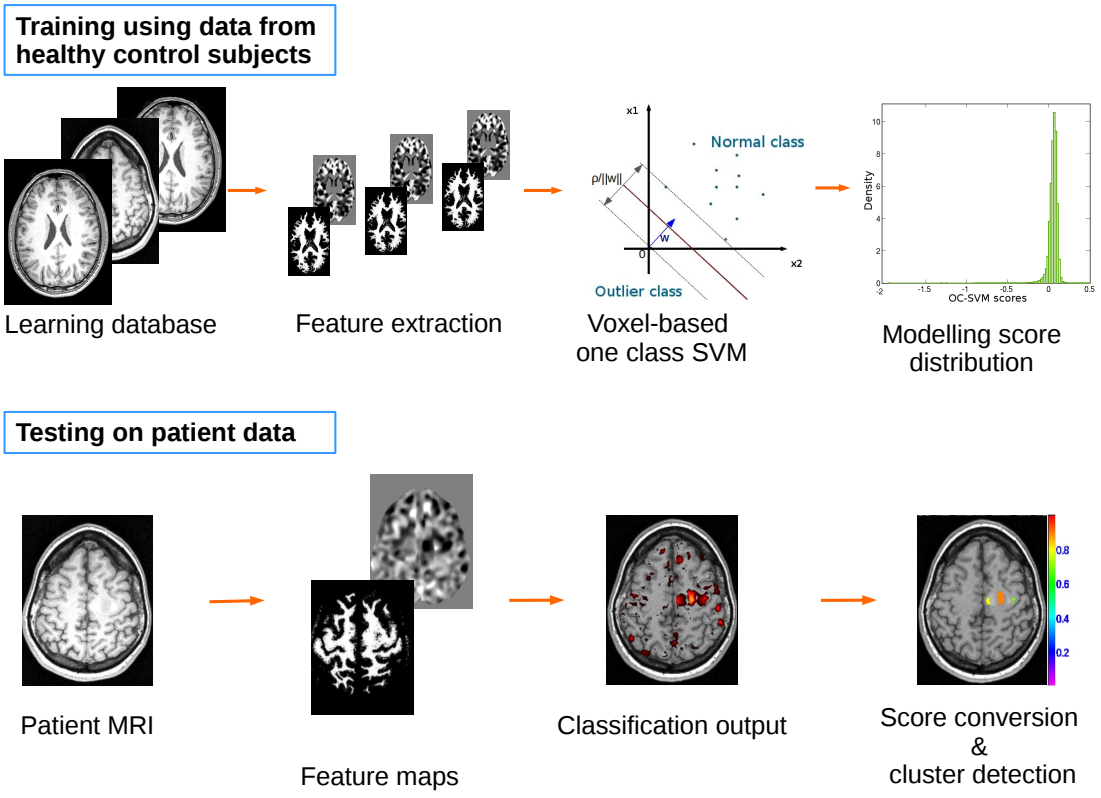


Figure 5.1: **Scheme of the CAD system illustrating the learning (top) and the testing phase (bottom).**

5.1 Data description

5.1.1 Study group

The study was approved by our institutional review board (Comité de protection des Personnes Sud-Est IV) with approval number: 140277A-12 and 2012-A00516-37 and written informed consent was obtained for all participants.

Clinical patient group: The clinical test group was composed of 11 patients (all patients from PDB1 and 6 patients from PDB2) admitted to Lyon's Neurological Hospital for medically intractable epilepsy. Five of these patients were operated upon between 2009 and 2013 and found to have histologically proven FCD. The presurgical workup included standard MR neuroimaging as well as SEEG and VEEG. Some patients also had a PET and/or a MEG as part of their clinical workup when clinically indicated. Post-processing of these anonymised scan data acquired for clinical purposes did not require individual

consent from the individuals who had been scanned. The second series included six patients who were admitted between 2014 and 2015 following the same inclusion criteria as those of the first group of five patients. These patients underwent a similar presurgical evaluation except that the MR imaging protocol was slightly different as explained in the next section. They also underwent systematic PET and MEG as part of a research protocol. All patients had an initial routine radiological assessment consisting in a blind visual inspection of standard T1-weighted 1.5 T MRI by two expert neurologists. Results of this visual inspection are reported in the first three columns of Table 5.2. Three FCD lesions were visually detected on three patients thus diagnosed as MRI-positive (MRI+). The same experts also reported the two hippocampal atrophies (HA) of patient #4 and patient #6. The two FCD lesions of the right amygdala for patient #4 and of the right temporal lobe for patient #6 were not visible on the T1-weighted MRI. These areas were suspicious on VEEG and SEEG respectively and lesions confirmed on histology. These two patients were thus classified as MRI-negative (MRI-). The MRI- group also included the six remaining patients.

Normal subject database: Two learning databases were used in this study to match the characteristics of the two series of clinical data detailed above. The first database referred to as NDB1 consisted of 37 T1-weighted MRI exams acquired on healthy control subjects aged 18-53 years. The second one referred to as NDB2 consisted of 40 T1-weighted MRI exams acquired on healthy control subjects aged 20-62 years. Both databases were visually inspected to exclude subjects with significant structural abnormalities.

Simulated patient group: Five additional MRIs of healthy control subjects (simulation subjects) acquired following the same protocol as for NDB1 were used to simulate two kinds of epileptogenic lesions.

5.1.2 MRI acquisition

Subjects from NDB1, the simulated patient group and the first series of 5 patients all had a 3D anatomical T1-weighted brain MRI sequence (TR/TE 9.7/4 ms; 176 sagittal slices of 256×256 millimetric cubic voxels) on a 1.5 T Sonata scanner (Siemens Healthcare, Erlangen, Germany).
Subjects from NDB2 and the second series of 6 clinical cases had a 3D anatomical T1-weighted brain MRI sequence on the same scanner but with a slightly different protocol (TR/TE 2400/3.55; 160 sagittal slices of 192×192 1.2 mm cubic voxels).

5.2 Pre-processing

5.2.1 Spatial normalization

Following the work in [Huppertz *et al.* (2005)], the preprocessing was performed based on the reference methods implemented in SPM8. The spatial normalization was performed using the unified segmentation algorithm (UniSeg) [Ashburner and Friston (2005)]. This

method combines all steps including the segmentation of the different tissue types, namely grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), correction for magnetic field inhomogeneities and spatial normalization under the same objective function that is iteratively solved. The 3D brain MRI of each subject was normalized to the standard brain template of the Montreal Neurological Institute (MNI) [Mazziotta *et al.* (2001)] as contained in SPM8 using the default parameters for normalization and a voxel size of $1 \times 1 \times 1$ mm as in [Ashburner and Friston (2005)].

A recent study by [Klein *et al.* (2009)] compared several registration algorithms of healthy control subjects brain images. In this comparison, the DARTEL method that was introduced by Ashburner in 2007 [Ashburner (2007)] as an alternative to the unified segmentation algorithm was found to be more accurate. DARTEL, a diffeomorphic image registration algorithm, involves performing first the correction for intensity non-uniformity and computing the three tissue class probability maps. The subject images are then used to create an initial template by averaging all subjects' images. An iterative process is then performed as follows: 1) the template (source) is deformed to each of the individual images (target) by simultaneously minimizing the sum of squares difference between the source and target images as well as the linear elastic energy of the deformations used to warp the target image, 2) the template is updated by applying the inverse of the deformations to the subject images followed by averaging, 3) the whole process is repeated. In this study, the processing of the MRIs was performed as described in the DARTEL tutorial (http://www.fil.ion.ucl.ac.uk/~john/misc/VBMclass10.pdf). The intensity correction and computation of the tissue class probability maps were carried out by the UniSeg method.

The cerebellum and brain stem were excluded from the spatially normalized images to restrict the analysis to brain regions susceptible to harbour FCDs. The masking image in the reference MNI space was derived from the Hammersmith maximum probability atlas in Fig 5.2 [Hammers *et al.* (2003)]. The resulting volume of interest contained 1.5 million voxels ($\approx$ 1.5 liters). Fig. 5.2 shows an example slice of the maximum probability map and the masking image.

5.2.2 Feature extraction

Previous work [Duchesne *et al.* (2006), Antel *et al.* (2003), Thesen *et al.* (2011), Huppertz *et al.* (2005)], referenced in Tables 2.3 and 2.4 , showed the discriminative power of different types of features for the detection of epileptogenic lesions on T1-weighted MRIs. This includes grey-level intensities of the T1-weighted images, volume of specific brain structures such as hippocampus, tissue class probabilities, image maps highlighting FCD characteristics, as well as second order texture features based on grey level co-occurrence matrices. In this study, we hypothesized that the CAD system would be selectively specific to FCD and heterotopia as typical epileptogenic lesions by considering only the features that model the clinical description of this type of lesions and that were previously shown to be discriminant (see for instance [Huppertz *et al.* (2005)] and [Wagner *et al.* (2011)]).

(a) (b)

Figure 5.2: Example slice of (a) the maximum probability atlas and (b) the resulting volume of interest.

Three parametric maps were thus computed for all subjects from the probabilistic tissue maps to capture suspicious patterns characterizing 1) the extension of the GM into the WM and referred to as extension map, 2) the junction between the grey and white matters and referred to as junction map, and 3) the thickness of the GM referred to as the thickness map.

The extension map was obtained from the segmented GM image derived from the segmentation by smoothing with a Gaussian kernel of width 6 mm. To compute the junction map, the T1-weighted intensity corrected MR image was transformed into a binary image by selecting voxels whose grey value ranges between low threshold $T_{low} = mean_{GM} + \frac{1}{2}SD_{GM}$ and high threshold $T_{high} = mean_{WM} - \frac{1}{2}SD_{WM}$ where *mean* and *SD* values correspond to the mean and standard deviation of the grey values in the respective tissue class, with $mean_{WM} > mean_{GM}$. A smoothing with a 6 mm width Gaussian kernel was then applied to the binary image. The thickness map that measures the distance between grey and white matter surfaces, was obtained from the segmented GM image by performing run-length coding in each point-to-point direction [Bernasconi *et al.* (2001)]. Details concerning the computation of these maps are given in [Huppertz *et al.* (2005)] and [Bernasconi *et al.* (2001)].

For each feature type and for each healthy control database (NDB1 and NDB2), a mean parametric map (referred to as "mean template") and a standard deviation map (referred to as "SD template") were created by averaging and computing the standard deviation over the parametric maps computed from the 37 control subjects of NDB1 and from the 40 control subjects of NDB2, respectively. Fig. 5.3 gives examples slices of mean and SD templates of the junction map obtained with both pre-processing methods (UniSeg and DARTEL).

The final individual parametric map, also called Z-score map, was obtained by subtracting the mean parametric map (mean template) from the parametric value of the

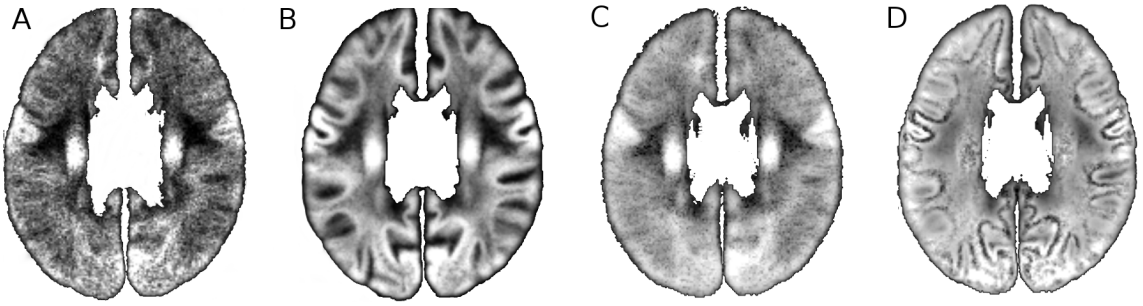


Figure 5.3: Example axial slices of the junction map templates for NDB1: (A) Uniseg mean template (B) DARTEL mean template (C) Uniseg SD template and (D) DARTEL SD template.

individual map and dividing by the standard deviation template. As a result, high signal indicates typical MRI features of epileptogenic lesions. Individual tissue probability maps (GM, WM, CSF) were also considered as discriminant features. Each voxel k from the volume of interest was thus described by a six component feature vector $\mathbf{V}_k$ comprising the three tissue probability values as well as the extension, the junction and the thickness value at this voxel location. Fig. 5.4 gives example slices of the three parametric maps as well as the corresponding individual tissue probability maps.

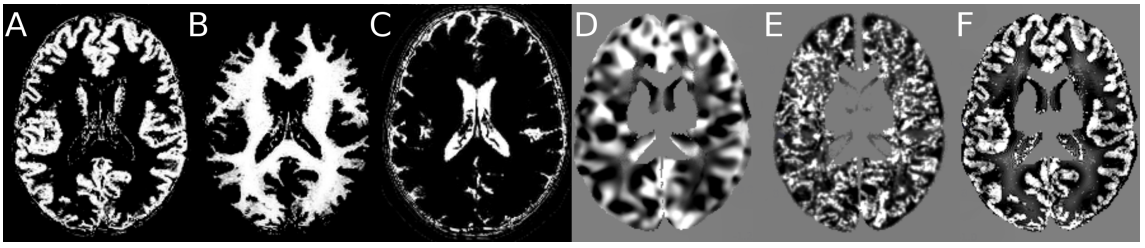


Figure 5.4: **Example slice of:** *(A) GM probability map (B) WM probability map (C) CSF probability map (D) extension map (E) junction map and (F) thickness map.*

5.3 Classification

As discussed in Chap. 4, the OC-SVM approach [Schölkopf *et al.* (2001)] allows learning from negative examples only while keeping all the desirable properties of the support vector machines (SVM). This includes combining different maps by specifying the features to be included in the learning step (multivariate approach), using a kernel function to find non-linear decision boundaries, and yielding a sparse representation of this decision boundary.

Each voxel k from the MRI scan was separately associated with a OC-SVM classifier. The classifier was trained using the matrix $\mathbf{M}^k \in M_{n,p}(\mathbb{R})$ with $n = 37$ for NDB1 and $n = 40$ for NDB2 and $p = 6$ where each row of M^k is an instance of the feature vector $\mathbf{V}_k$. A RBF kernel was used and the values of ν and the kernel parameter σ were derived as described below. For a voxel k, the output of the OC-SVM is the signed distance from the optimal hyperplane found during the learning phase at this voxel location. As we only learn from normal examples, most training examples will have a positive signed distance

to the optimal separating hyperplane and only a fraction of the examples (controlled by ν) will have a negative signed distance. The 1.5 million OC-SVM predictive models were computed. These models were applied to test images to produce a OC-SVM distance map, with the same dimensions as the normalized input image, where each voxel value is the local OC-SVM distance to the local hyperplane found during the learning step. The Matlab toolbox developed by [Canu *et al.* (2005)a] was used to solve the optimization problem (Eq. 4.3) at each voxel location.

5.4 Post-processing

Thresholding the OC-SVM distance map allows identifying clusters of voxels that will be regarded as pathological. As outlined above, the more negative the score, the more suspicious (pathological) the voxel. Selection of the threshold thus controls the trade-off between the sensitivity and the specificity of the CAD system.

We developed a novel method to adjust the threshold value so as to control the type I error (false positive detection rate). The idea is to model the distribution of the OC-SVM scores for normal voxels and then infer the probability for any given test voxel to be abnormal considering its score value relative to the normative distribution. This normative score distribution was computed from the score distribution of the control subjects based on a leave-one-out procedure as follows. $n-1$ control subjects first served to learn the map of OC-SVM models (one model for each of the 1.5 million voxels) that was then applied to compute the OC-SVM distance map of the remaining control subject. For each control subject left out, the histogram of the OC-SVM distance was approximated by a non-parametric distribution using a kernel density estimator [Bowman and Azzalini (1997)]. All n histograms were then pooled to obtain the normative score distribution. We assumed that the OC-SVM score distribution of any given test patient can be represented by this normative score distribution considering that it is not influenced by the small fraction of outlier examples ($< 1\%$) that are likely to correspond to typically sized lesions. This is equivalent to considering the OC-SVM scores for abnormal examples as outliers of the distribution of normal example scores. The type I error can then be controlled by using a threshold value that corresponds to a given p-value on the normative distribution. The good overlap between the score distribution of a control subject and that of a test patient in Fig. 5.5 illustrates the validity of this hypothesis. We set a p-value of 0.001 which is equivalent to a signed distance of -0.95 for NDB1 and -1.07 for NDB2 given the two normative score distributions estimated in this study. The clustering process then consists in scanning the thresholded map in a lexicographical voxel order and aggregating all non-null voxels that are linked for the 26-connectivity rule. Descriptive statistics per cluster (size, minimum, maximum and mean of the voxel scores) are then computed. Each cluster is assigned a value that corresponds to the minimum OC-SVM score value of its constituting voxels, i.e. the smallest probability value of belonging to the normal class. Finally, a labelled cluster map is generated in which cluster order is given by the minimum

(i.e. most pathological) OC-SVM score value.

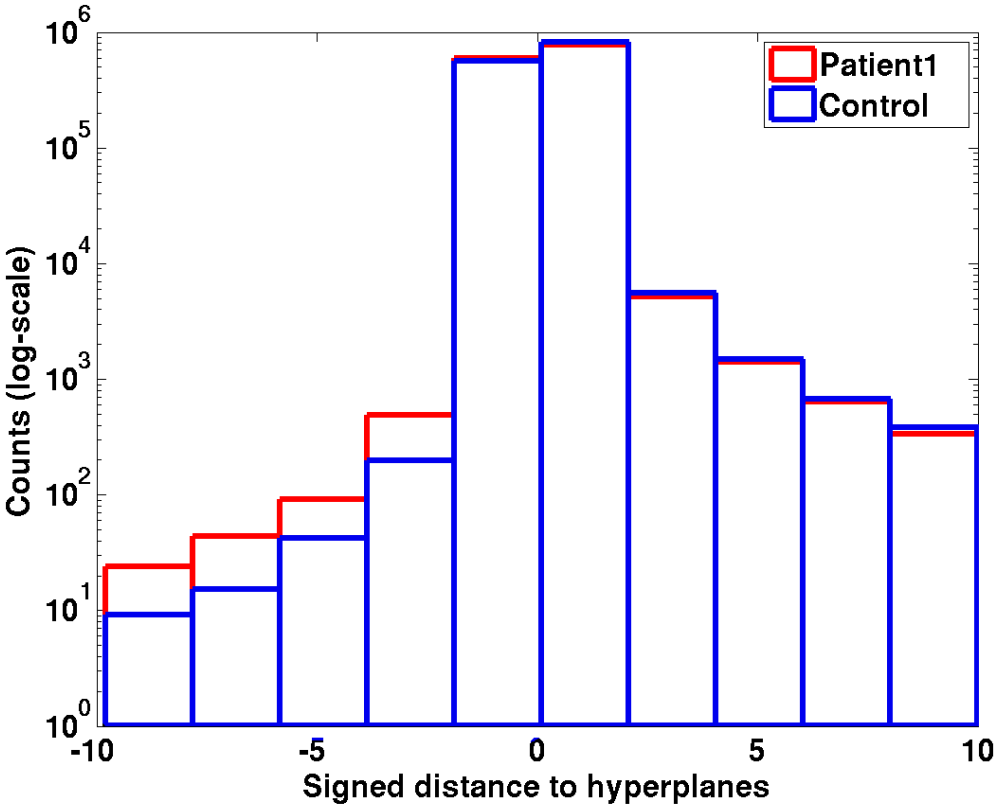


Figure 5.5: **Example of OC-SVM score histogram (on a log-scale) obtained for a control subject from NDB1 (blue) overlaid with that of patient #1 (red).** *There are differences in a small number of voxels, all with a negative signed distance to the hyperplanes indicating non-normal tissue.*

5.5 Evaluation of the CAD system

5.5.1 Comparison against SPM

The mass univariate single subject analysis [Ashburner and Friston (2000)] was performed within the framework of the statistical parametric mapping software (SPM: fil.ion.ucl.ac.uk/spm; Wellcome Trust Centre for Neuroimaging). A general linear model (GLM) is first fitted to each voxel based on user-predefined factors of interest (e.g. groups, frequency of seizure, etc) and/or confounding variables (e.g. sex, age) [Ashburner and Friston (2000)]. Then, post-hoc inferences on the effect of interest are made with a standard mass univariate statistical test resulting in statistical maps of T or F values for each factor of interest. The statistical significance of clusters of voxels that exceed an uncorrected statistical threshold in the SPM is evaluated within the Gaussian random fields (GRF) theory which allows correcting p-values for multiple testing in the search volume and correlation among neighbouring voxels (due to spatial smoothing). SPM also allows performing con-

junction analysis as defined in [Friston *et al.* (2005), Friston *et al.* (1999)] to test the global null hypothesis that there is a conjunction of one or more effects, i. e. the factors of interests were consistently and jointly significant.

We performed a 'one-way ANOVA' based on the following four factors of interest: patient junction map, control junction maps, patient extension map, and control extension maps. We used two contrasts [1,-1,0,0] and [0,0,1,-1] to test for significant increases in the patient junction and extension maps compared to controls. A first analysis consisted in thresholding the two resulting T-score maps using a p-value of $p = 0.001$, to produce a cluster map where each cluster was characterized by its size and the maximum T-score value of its constituting voxels. Clusters with the highest T-scores were considered as most suspicious. A conjunction analysis of these two effects (junction and extension), as defined above, was also performed using the same p-value of 0.001.

For a fair comparison of the SPM analysis against the proposed CAD system, we used only the junction and the extension features as inputs for OC-SVM instead of all six features.

5.5.2 Evaluation on simulation data

We simulated two types of typical epileptogenic abnormalities described in Chap. 1, firstly a focally blurred junction between grey and white matter and secondly, heterotopion-like lesions resulting from the presence of GM in the white matter. Five MRI scans of healthy subjects from the simulation group were used to perform all simulations.

Simulation of blurred junctions The simulation of subtle junction alterations included the following steps: **1-** in the native T1-weighted MRI of a simulation subject, 2D U-shaped regions of interest (ROI) around the grey matter at the bottom of a sulcus were drawn on six consecutive slices to capture 3D information (see Fig. 5.6). **2-** the grey-level value histogram of the voxels within the junction ROI was computed. **3-** the voxels with grey-level values ranging in the GM/WM interface defined by $I = [mean_{GM} + \frac{1}{2}SD_{GM}\ ,\ mean_{WM} - \frac{1}{2}SD_{WM}]$ (with $mean_{WM} > mean_{GM}$) were isolated. **4-** a mean filter, a disk of radius 3 mm ($\approx$ 28 voxels) was applied to the entire original MRI. **5-** In the original MRI, values of the voxels selected in step **3** and within the ROI were substituted by their corresponding values in the filtered image. This processing ensures that only the voxels within the GM/WM interface are altered, closely resembling epileptogenic FCDs with blurred junctions [Huppertz *et al.* (2005)].
We simulated one lesion for each simulation subject, resulting in a total of 5 FCD-type junction lesions. Fig. 5.6-A, 5.6-B, and 5.6-C show an example of a very subtle U-shaped lesion.

Simulation of heterotopion-like lesions For each of the five simulation subjects, six heterotopion-like lesions were simulated at different locations in the WM. To control the

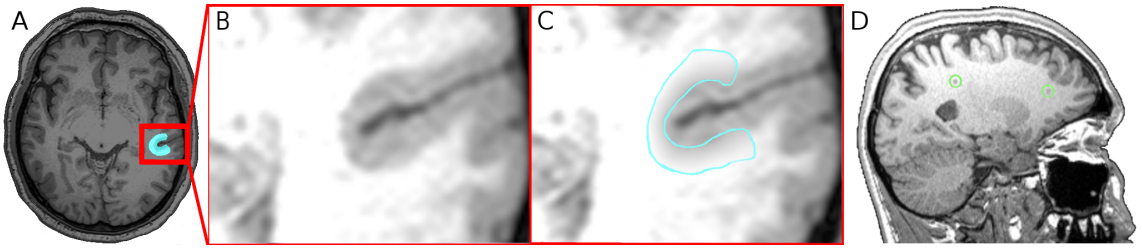


Figure 5.6: **Realistic simulations:** *(A) Example slice of the original MRI where the alteration location is highlighted in blue, (B) zoom on the original MRI before introducing the alteration, (C) and zoom on the introduced junction alteration. The introduced lesion has a very low contrast and is almost impossible to detect with the naked eye. (D) Example of a simulation subject sagittal MRI slice showing two heterotopion-like lesions (within the green circles) that were simulated using GM values selected within the range $C3$ of grey-level values.*

inter-subject variability due to lesion location, these lesions were simulated at the same approximate location in all five subjects. A binary mask containing six spherical lesions of radius 2 mm was drawn in the MNI space. This mask was then mapped to each individual subject native space to set the location and size of the six simulated lesions. Three lesion contrast levels were selected to sample a range of detectability. The standard values (*mean* and *standard deviation*) for each tissue class (GM and WM) were extracted from the grey-level histogram of each individual subject's MRI. The three contrast intervals were then set as follows:

$$
\begin{aligned}
C1 &= [mean_{GM} - \tfrac{1}{2}SD_{GM}\ ,\ mean_{GM}],\\
C2 &= [mean_{GM}\ ,\ mean_{GM} + \tfrac{1}{2}SD_{GM}],\\
C3 &= [mean_{GM} + \tfrac{1}{2}SD_{GM}\ ,\ mean_{WM} - \tfrac{1}{2}SD_{WM}].
\end{aligned}
$$

For each contrast level, grey-level values of the heterotopion-like lesions were selected by randomly sampling values with the considered contrast level. The contrast to surrounding WM was thus greatest for contrast level $C1$ and smallest for $C3$ that also corresponds to the contrast used for blurred junction-like lesions.

A total of 5 (number of simulation subjects)$\times$6 (number of different locations) $\times$3 (number of contrasts) = 90 heterotopion lesions was simulated. Fig. 5.6-D gives an example slice of the heterotopion lesions obtained with contrast $C3$ for a simulation subject displayed in native space.

Comparison with the ground truth for synthetic data A ROC analysis [Hanley and McNeil (1982)] was performed based on OC-SVM and SPM score maps (T-scores for SPM and distances to the hyperplane for OC-SVM). The ROC curve reports coupled values of the true positive rate (TPR) and false positive rate (FPR) for different values of the decision threshold. In this study, the ROC curve was computed at the voxel level, thus meaning that a voxel was recorded as a true positive if its score value exceeded the threshold and was located in one of the simulated lesions. This type of voxel analysis avoids defining mark-labelling rules that may bias the performance assessment [Petrick

et al. (2013)]. For a test patient, the threshold was varied to cover the entire range of T-scores (SPM) or OC-SVM scores for a test patient so as to homogeneously sample the ROC curve. Performance was also summarized by the standard area under the ROC curve (AUC). Confidence intervals on the AUC estimates were computed by using bootstrap estimates at the patient level for both the SPM and OC-SVM methods. One hundred and twenty bootstrap samples were obtained by choosing with replacement among all simulated patients repeated for each anomaly. For each bootstrap sample, the score maps (SPM T-scores or OC-SVM scores) of the selected simulated patients were concatenated to form a global score vector that was used to compute one bootstrap AUC estimate. Confidence intervals for the differences in AUC estimates between the different CAD configurations were finally derived according to the bias-corrected percentile method described in [Manly (2006)].

For clinical applications with highly imbalanced datasets (in our case, lesions have a small size compared to the entire volume ($< 1\%$)), the TPR value for a fixed low value of the FPR should also be reported [He and Garcia (2009)]. In this study, a FPR of 0.1 corresponds to the detection of 150 000 voxels (10% of the 1.5 million voxels) which is already more than 100 times the size of the true lesion. We thus chose to compare the performance at three fixed FPR values of 0.01, 0.05 and 0.1, respectively.

5.5.3 Evaluation on clinical data

The clusters detected either by the OC-SVM or by the SPM analysis were compared with the FCD lesion visually detected by expert assessment and other data (e.g. FDG-PET, Magnetoencephalography) for each of the eleven patients considered in this study. Clusters were designated by an expert as false-positive detections, or as identifying an abnormal region comprising the lesion. Considering the expected size of clusters given by the SPM analysis (82 voxels = 82 mm^3), we only reported clusters that were superior to 82 voxels for the SPM and the OC-SVM map as both analyses were performed using the same input feature maps. Sensitivity was defined as the fraction of detected clusters that correctly colocalized with the expert report. In lieu of specificity, we report the mean number of false positive (FP) clusters per patient scan. Separate evaluations were performed for the 3 MRI+ and the 8 MRI- cases.

5.6 Results

5.6.1 OC-SVM parameter optimization

The standard deviation of the RBF kernel σ and the parameter ν were optimized by randomly selecting 4000 voxels from the whole brain volume, and performing a leave-one-out optimization procedure for each voxel based on the 37 healthy control subjects of NDB1 and independently on the 40 healthy control subjects of NDB2. Fig. 5.7 shows some slices of the entire volume of interest where the voxels that were used for optimizing

the hyper-parameters are highlighted.

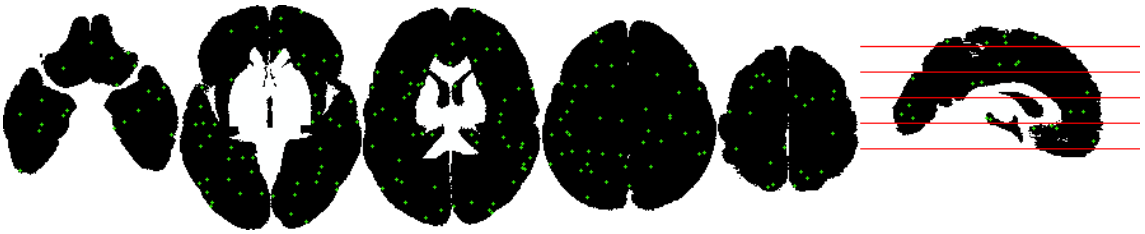


Figure 5.7: Example axial slices showing the randomly selected voxels (green plus mark) for optimizing the hyper-parameters (ν, σ).

In Chap. 4, we discussed strategies for tuning OC-SVM and SVDD hyper-parameters. In this study, we used the target error estimate proposed by [Tax and Duin (2004)] to optimize the OC-SVM hyper-parameters (see 4.4). This error estimate was obtained by averaging the leave-one-out error over the 4000 voxels. The value of ν was varied in the interval $[0.01 \ldots 0.56]$ using eight intervals on a log_{10} scale. The value of σ was varied in the interval $[2^{-4} \ldots 2^{4}]$ using eight intervals on a log_2 scale. Using UniSeg registration method, the values $\nu = 0.03$ and $\sigma = 4$ for NDB1, $\nu = 0.05$ and $\sigma = 3$ for NDB2 were shown to produce the smallest average leave-one-out error while balancing model complexity (in terms of dimensionality and memory load) and over-fitting.

Fig. 5.8 shows the resulting mean error curves for different ν and σ values and the UniSeg registration method. As expected the leave-one-out cross validation error is always superior to the minimum error achievable ν which depends on the number of observations and the number of support vectors (the ν-property discussed in Chap. 4). The optimization procedure was also performed for DARTEL using NDB1. The values $\nu = 0.05$ and $\sigma = 4$ were retained.

This optimization procedure is aimed at finding a compromise between computational efficiency and accuracy, as performing the grid search analysis for the entire volume of interest (1.5 million of voxels) would have been computationally prohibitively expensive.

5.6.2 Influence of the registration method on the CAD performance

The influence of the two registration methods (UniSeg and DARTEL) was evaluated for using the two types of simulated lesions.

Blurred junction Fig. 5.9 presents the results of the OC-SVM performance for both preprocessing methods. The ROC curves were obtained by pooling the scores of all five simulation subjects. Despite similar mean AUC values (0.94 and 0.95 for DARTEL and UniSeg), comparison of the sensitivity values at low FPR indicates that the UniSeg method yields the best performance for this type of lesions. The three coupled values of (TPR, FPR) are {(0.68, 0.01), (0.84, 0.05) and (0.88, 0.1)} for UniSeg and {(0.2, 0.01), (0.74, 0.05) and (0.85, 0.1)} for DARTEL. To explain the performance difference reported in Fig. 5.9, we examined the junction templates (maps of mean and standard deviation of

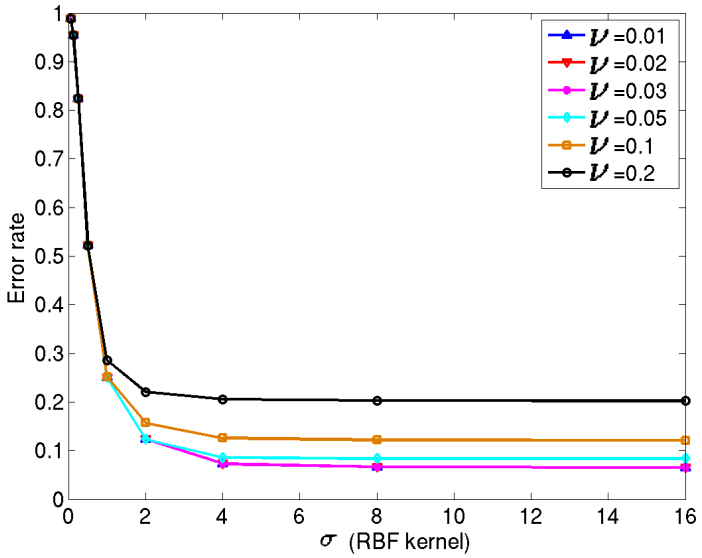


Figure 5.8: Leave-one-out mean error rate estimate for different vales of (ν, σ). All curves were obtained with the UniSeg registration method.

the junction feature) used in deriving the individual junction maps for both preprocessing methods, as explained in Section 5.2.2. Fig. 5.3 shows the junction standard deviation template obtained with UniSeg (Fig. 2.C) and DARTEL (Fig. 2.D). Comparison of these two figures shows that the DARTEL standard deviation image delineates gyri and sulci more precisely than the UniSeg image, indicative of the known higher registration performance of the DARTEL algorithm for these structures. As the introduced alteration is located near a gyrus or a sulcus, any deviation of the patient's anatomy (anatomical variability in gyrus position for instance) from the DARTEL standard deviation template is likely to produce high values of the normalized junction feature (division by a value of the standard deviation close to zero) and thus to trigger a false positive detection. This is likely to explain the higher FPR obtained with the DARTEL preprocessing method (see ROC curves in Fig. 5.9).

Heterotopion-like lesions Fig. 5.10a, 5.10b and 5.10c represent the ROC curves obtained for each of the three simulated contrasts with both preprocessing methods (UniSeg and DARTEL). The CAD system combined with the DARTEL preprocessing method yields the best AUC performance for well contrasted lesions ($C1$ and $C2$) and similar performance to UniSeg for the low contrast level ($C3$). The sensitivity values at low FPR (<0.1) are similar for all contrasts.
Fig. 5.10d shows the results when all types of heterotopion-like lesions are pooled together. In this case, the system achieved a global AUC value of 0.93 for UniSeg and 0.95 for DARTEL. The UniSeg method, however, allows achieving higher sensitivity values at low FPR; the three coupled values of (TPR, FPR) are {(0.80, 0.01), (0.86, 0.05) and (0.87, 0.1)} for UniSeg and {(0.74, 0.01), (0.87, 0.05) and (0.90, 0.1)} for DARTEL.

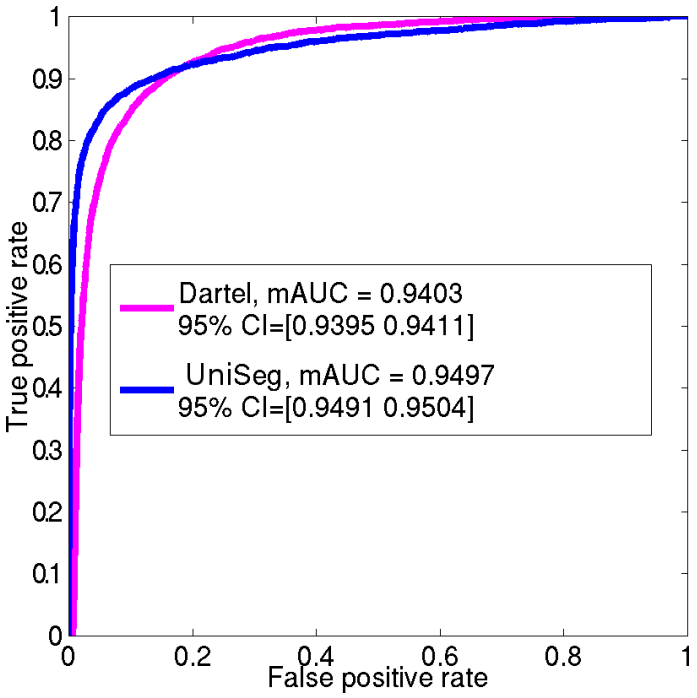
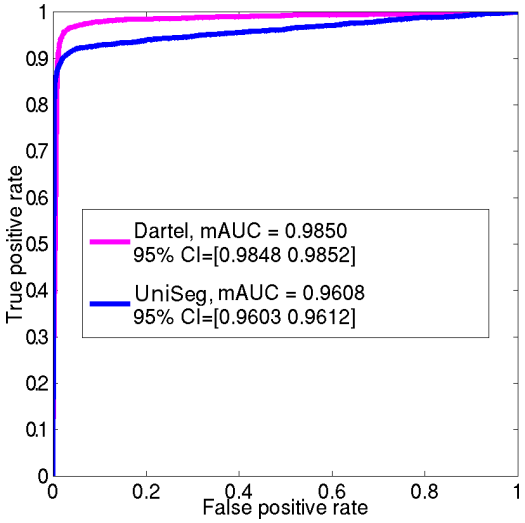


Figure 5.9: ROC curves for the UniSeq and DARTEL registration methods for the simulated blurred junction lesions.

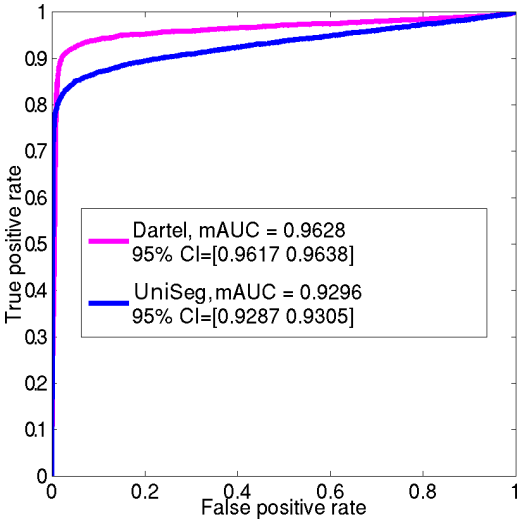
The visual analysis of the extension standard deviation templates of both preprocessing methods (Fig. 5.11) indicates that at the simulated lesion locations, the variability of the extension feature among the 37 normal subjects is very low (almost null) for both UniSeg and DARTEL. The spatial distribution of the extension variability is however more uniform for DARTEL than for UniSeg. The normalized individual extension map is computed by subtracting the average extension map from the raw extension map and dividing by the standard deviation map. Consequently, unlike for DARTEL, the normalized individual extension map for UniSeg presents a non uniform spatial distribution at the lesions' location. This is likely to result in broadening the distribution of the OC-SVM distances at the location of the simulated heterotopion-like lesions which consequently requires to increase the threshold value (see Fig. 5.5) to guarantee retrieving all lesions. This will be at the cost of an increased number of false positive detections and thus explains the smaller AUC value achieved with the UniSeg method on Fig. 5.10.

The simulation results show that the two preprocessing approaches (DARTEL and UniSeg) perform almost equally well with a slight advantage for the UniSeg approach for junction related abnormalities (see Fig. 5.9) and a slight advantage for the DARTEL approach for well contrasted heterotopion-like lesions (see Fig. 5.10a).

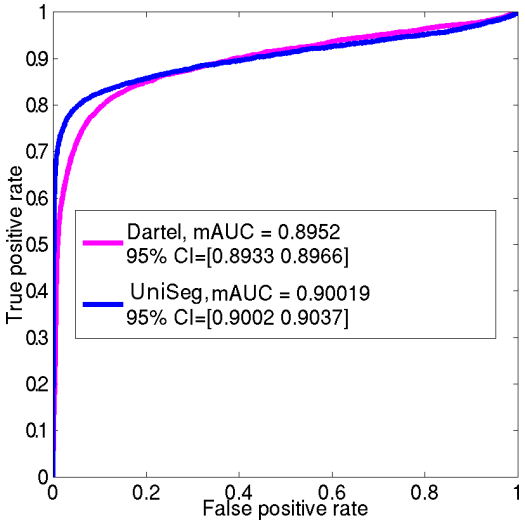
As no *a priori* knowledge about the lesion type is available, one cannot adjust the preprocessing approach to the lesion at hand. For the rest of this study we thus chose to use only UniSeg in the preprocessing steps. This choice was motivated by first considering the slight difference in performance between the two methods and second the larger computational load of DARTEL which requires an initial segmentation that is usually obtained



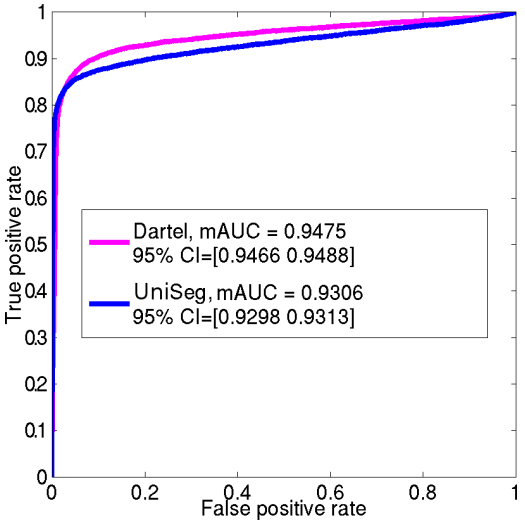
(a) UniSeg vs DARTEL: heterotopion-like lesions contrast C1 (high contrast)



(b) UniSeg vs DARTEL: heterotopion-like lesions contrast C2 (middle contrast)



(c) UniSeg vs DARTEL: heterotopion-like lesions contrast C3 (low contrast)



(d) UniSeg vs DARTEL: heterotopion-like lesions over contrasts

Figure 5.10: CAD system performance for heterotopion-like lesions with both UniSeg and DARTEL preprocessing methods.

from the UniSeg algorithm.

5.6.3 Comparison of OC-SVM and SPM detection performance

Simulated data results Fig. 5.12-a shows the ROC curves corresponding to the detection of the five blurred junctions. For these lesions, the OC-SVM approach and the SPM analysis based on the junction contrast perform equally in terms of AUC (AUC = 0.95) while the SPM conjunction analysis has an intermediate performance with AUC values of 0.84. As expected the SPM analysis based on the extension contrast yields very poor detection performance (AUC = 0.65) for this detection task. These results corroborate the 95% confidence intervals on difference in AUCs reported in the second column of Table 5.1.

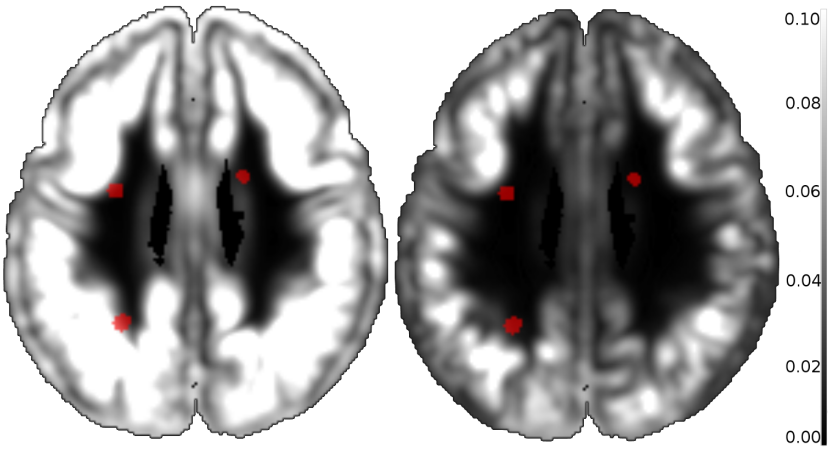
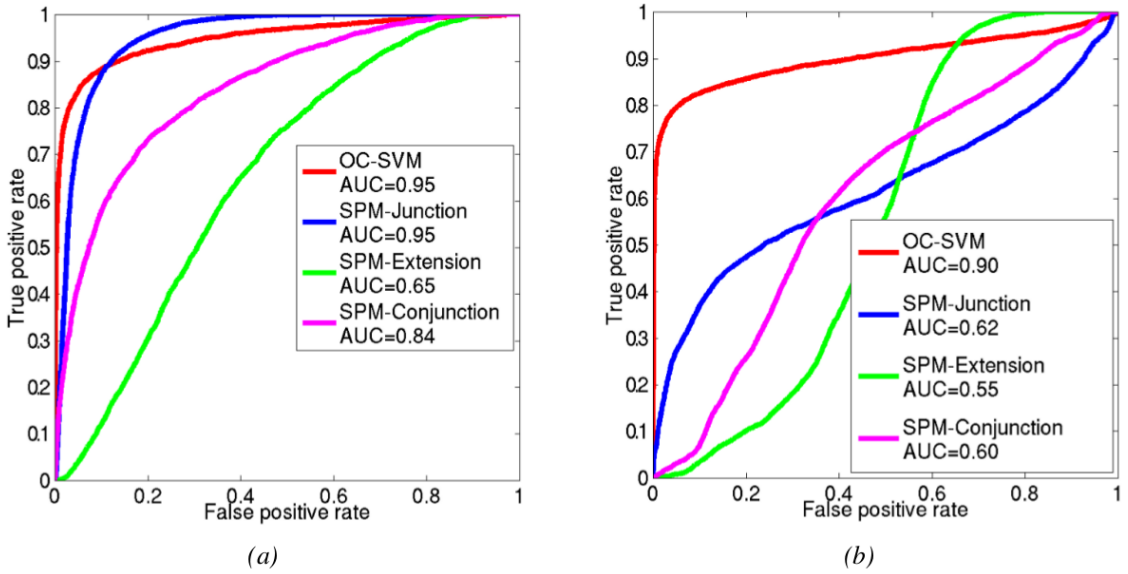


Figure 5.11: Example axial slice of the extension map SD template obtained using UniSeg (left) and DARTEL (right). The red areas illustrate the location of the simulated heterotopion-like lesions.



(a) *(b)*

Figure 5.12: **Comparison of OC-SVM and SPM performance for the simulated blurred junction and heterotopion-like lesions.**

The three coupled values of (TPR, FPR) are {(0.73, 0.01), (0.85, 0.05) and (0.88, 0.1)} for the OC-SVM and {(0.23, 0.01), (0.75, 0.05) and (0.87, 0.1)} for the SPM-junction classifier, thus underlining the higher performance of the OC-SVM classifier at low FPR which would be needed in clinical practice.

Fig. 5.13 gives an example of the detection maps obtained by OC-SVM and SPM analysis based on the junction map after thresholding the detection map at the same p-value of 0.001. This example illustrates that, for a reasonable false positive detection rate (0.1 %), the SPM approach fails to retrieve the blurred junction lesion while the OC-SVM approach detects most of the lesion with high specificity. The ability of the OC-SVM to

Table 5.1: **Comparison of OC-SVM and SPM classification performance. Data are differences in AUCs, with 95 % confidence intervals in brackets. All differences are significant and in favour of OC-SVM, except for the detection of the blurred junction where no difference between the techniques can be shown.**

	Blurred junction	heterotopy
OC-SVM vs SPM-junction	-7.622×10^{-4} $[-2.2 \times 10^{-3}$, $6 \times 10^{-4}]$	0.2737 $[0.2533, 0.2940]$
OC-SVM vs SPM-extension	0.2888 $[0.2835, 0.2951]$	0.3478 $[0.3445, 0.3507]$
OC-SVM vs SPM-conjunction	0.1053 $[0.1015, 0.1099]$	0.2892 $[0.2703, 0.3042]$

detect very subtle blurred junction lesions that are highly likely to be missed by standard visual inspection of the MR image is a very promising result.

For the heterotopion-like lesions, Fig. 5.12-b shows the ROC curves corresponding to the multivariate OC-SVM analysis as well as the SPM univariate analysis considering the junction and extension contrasts separately or the conjunction of both contrasts. The OC-SVM approach (AUC = 0.90) outperforms the SPM analyses based on the junction contrast (AUC = 0.62) and the conjunction analysis (AUC = 0.60) which both outperform the SPM analysis based on the extension contrast (AUC = 0.55) (see third column of Table 5.1). The ROC curve corresponding to the OC-SVM performance for heterotopion-like lesions in Fig. 5.12-b shows that the system was able to retrieve almost 70% of the global volume of the simulated lesions without any false positive detection. The coupled (TPR, FPR) values of the four methods were {(0.72, 0.01), (0.80, 0.05) and (0.83, 0.1)} for OC-SVM, {(0.09, 0.01), (0.28, 0.05) and (0.37, 0.1)} for SPM-junction, {(0, 0.01), (0.01, 0.05) and (0.03, 0.1)} for SPM-extension, and {(0, 0.01), (0.03, 0.05) and (0.07, 0.1)} for the conjunction of both SPM contrasts.

Clinical data results Patients from each test group were tested using OC-SVM and SPM models estimated based on their respective normal control database. The OC-SVM distance map and all three SPM maps (based on the junction or extension contrasts or the conjunction of both contrasts) of each of the eleven patients were thresholded at the same p-value of 0.001. Table 5.2 summarizes the results obtained for the eleven patients with all four approaches.

Results for MRI + patients (#1 to #3): The OC-SVM approach succeeded in detecting 3/3 lesions reported by the neurologist, yielding a sensitivity of 100% for MRI+ lesions with an average of 1.7 FP detections per patient. The SPM junction based analysis achieved equivalent sensitivity with an average of 6.3 FP per patient. As for the simulation examples, the OC-SVM approach produced fewer false positive detections than all SPM analyses (>5.7 FP). Fig. 5.14 compares the maximum intensity projection (MIP) of the classification results obtained by OC-SVM and SPM analyses based on the junction and extension map for patient #2 (see Table 5.2). Both methods succeeded in identifying the confirmed FCD lesion located in the left fronto-basal area. Comparison of cluster shapes

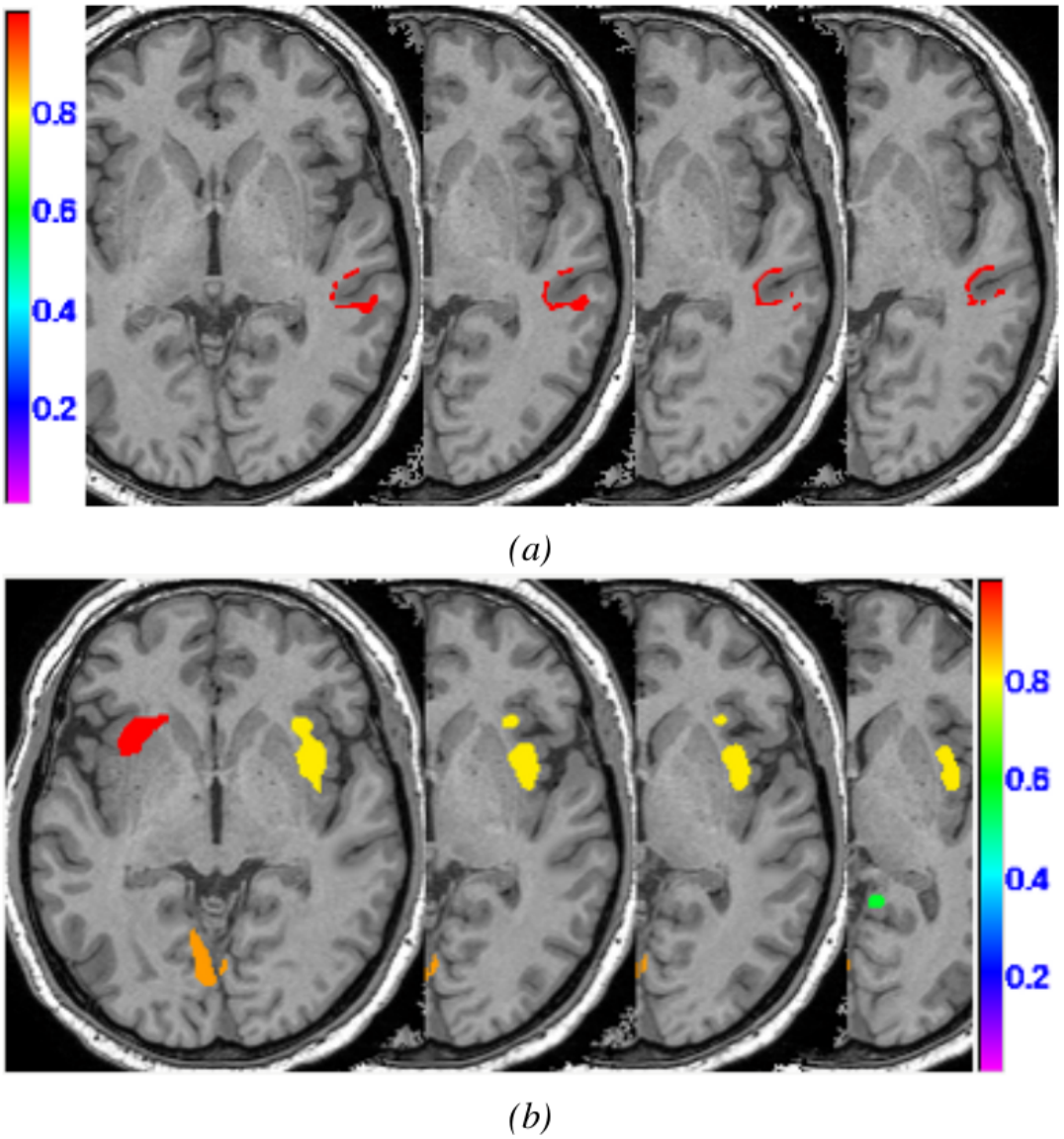


Figure 5.13: **Example of OC-SVM and SPM labelled cluster maps for the blurred junction simulation.**

and sizes in Fig. 5.14-a, 5.14-b, 5.14-c and 5.14-d illustrates that the OC-SVM approach is more specific in lesion localization than the SPM approach. For this lesion resulting from a GM extension, the cluster detected by OC-SVM is indeed located exactly at the bottom of a sulcus, whereas the cluster detected by the extension based SPM analysis has a bigger extent and is smoother. The OC-SVM analysis produced 3 FP whereas the SPM analyses based on the junction and extension contrasts produced 4 FP and 34 FP respectively. The conjunction analysis of both contrasts produced 7 FP.

Results for MRI- patients (#4 to #11): The OC-SVM approach succeeded in detecting 7/10 lesions, resulting in a sensitivity of 70% for MRI- lesions with an average of 3.7 FP detections per patient. It missed the lesions in patients #7 and #11. Two clusters were

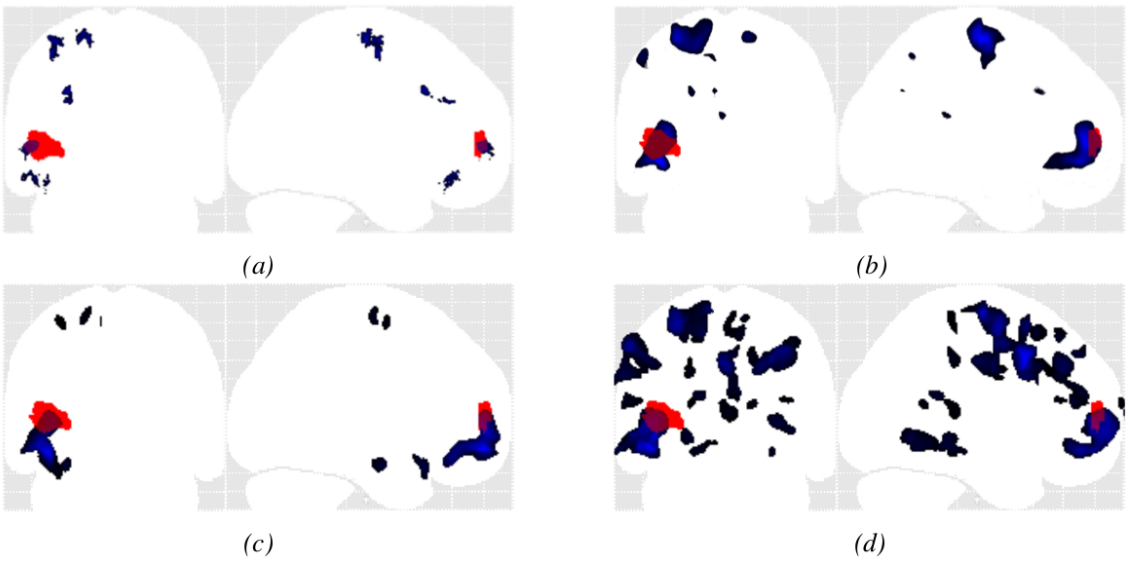


Figure 5.14: **Example MIP of the detected cluster maps (blue) for patient #2 (MRI+) overlaid on the MIP of the expert delineated lesion (red):** *(a) OC-SVM distance map thresholded at $p < 0.001$; (b) SPM analysis based on the T-score map from the conjunction of both contrasts thresholded at $p < 0.001$ (c) SPM junction-based T-score map thresholded at $p < 0.001$; (d) SPM extension-based T-score map thresholded at $p < 0.001$.*

correctly detected at these two lesion locations from the thresholded OC-SVM score map but they were discarded because of their small size (size <82 voxels). When selecting a higher p-value of 0.005, these two lesions are well detected at the price of a decreased specificity (13 FP for patient #7 and 15 FP for patient #11 against 3 FP and 1 FP for p=0.001, data not shown). Among all SPM analyses, the conjunction of both contrasts achieved the best detection performance in terms of sensitivity and the number of false positive detections, by detecting 5/10 lesions (50% sensitivity) with an average of 5.7 FP detections per patient. The conjunction of both contrasts allowed detecting the lesion in patient #7 that was initially missed by both individual contrasts. All SPM analyses missed the lesion in patient #4 located in the right amygdala. Fig. 5.15 compares MIPs obtained for patient #10. The lesion for this patient was not spotted after visual inspection of the MRI. The other exams including SEEG, VEEG as well as FDG-PET and MEG, however, all colocalized the presumed lesion in the left parietal lobe. Fig. 5.15-a, 5.15-b, 5.15-c and 5.15-d illustrate the higher specificity obtained using the OC-SVM approach.

Overall performance: Considering all eleven patients, the OC-SVM approach detected 10/13 lesions (77% overall sensitivity) with an average of 3.2 FP. The SPM conjunction analysis detected 7/13 lesions (54%) with an average of 6.3 FP. Both SPM analyses based on individual contrasts detected 6/13 lesions (46%) with an average of 7 FP for the junction contrast and 21 FP for the extension contrast. All approaches considered allowed a better sensitivity than the visual inspection of the T1-weighted MR scans that only allowed the detection of 5/13 lesions (39% sensitivity).

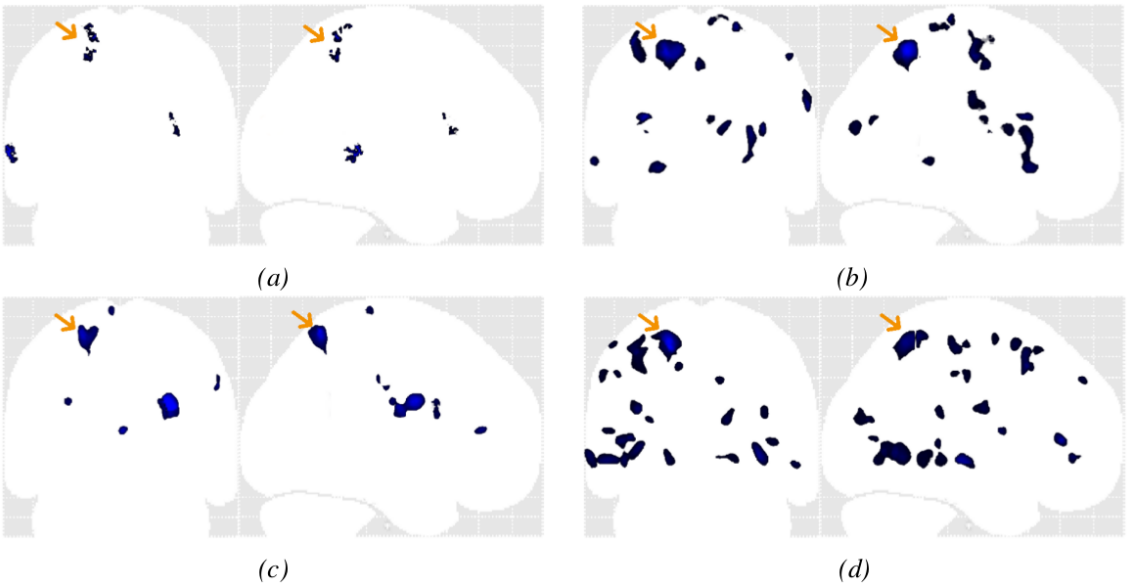


Figure 5.15: **Example MIP of the detected cluster maps (blue) for patient #10 (MRI-), the presumed lesion is indicated with the yellow arrow:** *(a) OC-SVM distance map thresholded at* $p < 0.001$*; (b) SPM analysis based on the T-score map from the conjunction of both contrasts thresholded at* $p < 0.001$ *(c) SPM junction-based T-score map thresholded at* $p < 0.001$*; (d) SPM extension-based T-score map thresholded at* $p < 0.001$*.*

5.7 Computation time

All results were obtained on a grid featuring 14 Intel Xeon quad-core L5420 2.50 GHz with 16 GB RAM memory. The main workload was associated with learning the 1.5 million OC-SVM models. If treated in sequential mode, this learning time was $\approx$ 28 hours while the feature extraction step and the score map computation for any test patient took about 12 and 15 min respectively. We therefore developed a parallel implementation based on the octree method [Meagher (1980)]. This method allows partitioning a three dimensional space by recursively subdividing it into eight octants. The final number of octants was set as to get a fixed maximum number of voxels in each octant. Each octant was then processed as an independent thread. In accordance with our grid specifications, each octant was composed of about 35000 to 60000 voxels. To process the whole 3D volume, 30 threads were executed in parallel on the grid. This allowed reducing the model learning time to less than 20 min and the two other processing times (feature extraction and score map computation) to less than 5 min.

5.8 Conclusions and perspectives

We designed and implemented an automated voxelwise detection system of common epileptogenic lesions in MR brain imaging based on machine learning of specific features of these lesions. The main contribution is 1) to consider this diagnostic task as a novelty detection problem that is solved using the OC-SVM methodology and 2) to perform a

multivariate voxel-based analysis yielding a cluster map quantifying the probability of a cluster of being abnormal. Performance of the multivariate OC-SVM based system was compared with that of the reference mass univariate SPM method. This comparison was first performed using simulated MR data mimicking standard epileptogenic lesions, and then using eleven patients, including 3 MRI+ and 8 MRI- exams.

The choice to perform a voxel-based analysis was motivated by the clinical need to develop an automated diagnostic system that can assist clinicians to non-invasively and accurately detect and delineate epileptogenic lesions of varying size anywhere on MRIs of the brain. The intuitive way to address this question would be to consider a standard binary classification problem.As underlined in the introduction, this however raises the issue of learning with imbalanced data [Sotiris *et al.* (2006), He and Garcia (2009)] and collecting enough samples from the pathological class to accurately model the intra-class variability. These led us to consider this challenging task as a novelty detection problem, which consists in constructing the predictive model from normal training samples only. Our choice to define one predictive OC-SVM model for each of the 1.5 millions voxels was aimed at accounting for the tissue and region specific feature distribution mentioned above. It was also inspired by the standard mass univariate statistical analysis developed in the neuroimaging community over the past few years. Our hypothesis was that the OC-SVM analysis could outperform the standard SPM analysis by enabling multivariate instead of mass univariate analysis.

Another novel contribution of this study is the method that we propose for converting the OC-SVM scores for a given test image into probabilities. This relies on the observation that typical epileptogenic lesions are very small in comparison with the entire volume of interest (less than 1%) and that we obtained similar distributions in terms of shape and skewness for the OC-SVM scores for patients and for healthy controls (see example in Fig. 5.5). This method that considers pathological samples as outliers of the normal score distribution may also be suitable for other pathologies based on subtle variations of the normal pattern. The proposed system was cross validated with a SPM analysis based on junction and extension maps.

For the simulation data, the CAD system based on the OC-SVM had an overall higher AUC than the SPM analysis for the two kinds of simulated lesions (heterotopion and blurred junction) and a higher sensitivity at very low FPR. For blurred-junction like lesions, unlike the SPM analysis, the OC-SVM approach was able to detect, with both high specificity and sensitivity, very subtle lesions that are very likely to be missed by standard visual inspection of the MRI (see Fig. 5.13).

For the clinical data, the proposed approach successfully detected 100% of the lesions in the MRI positive patients with less than 2 FP per scan while the SPM analysis based on the junction contrast achieved similar sensitivity at the price of about four times more FP per scan. For the MRI negative cases, the OC-SVM based approach outperformed all SPM analyses by combining both highest sensitivity (70%) and highest specificity (less than 4 FP per scan).

Performance achieved by OC-SVM compares well with the-state-of-the-art achieved in the recent study by Hong et al. [Hong *et al.* (2014)] that presented an automated algorithm for the detection of FCD type II in MRI- patients based on a two step classification scheme. The first step combined a linear discriminant analysis (LDA) with six surface based features (including cortical thickness) derived from the extraction of the inner and outer cortical surface from T1-weighted MR images. The resulting series of detected clusters (connected vertices) including true positive and false positive clusters (about 30 false positives per patient) were then passed through a second cluster-based LDA to remove the residual false detections. Results reported in their study showed a high sensitivity ($\approx$ 14 detected lesions out of 19 annotated lesions) and a good specificity with an average of 1-3 false positive detections per patient. These results are comparable with those achieved by our OC-SVM system in a one step procedure which successfully detected 7/10 MRI- lesions in eight patients with an average of 3-4 false positive detections per patient. In our study, unlike in Hong et al. [Hong *et al.* (2014)], we did not evaluate the specificity of the proposed system in healthy control subjects. However, we can deduce the achievable performance from the simulation study. For all simulation subjects, the proposed system identified the simulated lesion with no detections outside the simulated lesion location which suggests high specificity of our system in healthy controls. In another recent study, Ahmed et al. [Ahmed *et al.* (2015)] also proposed to combine five surface-based measures of cortical thickness at the vertex level with an ensemble classifier consisting of bags of 10 base-level classifiers trained using logistic regression. The authors evaluated the performance of this CAD configuration on 31 patients with FCD (7 MRI+ and 24 MRI- scans). Their approach detected 86% of the FCD lesions in the MRI+ group and 58% of the FCD lesions in the MRI- group. The author's approach was more sensitive than a mass univariate statistical analysis (SPM) based on the cortical thickness feature alone, but had a lower specificity.

The drawback to considering a voxel-wise classification scheme is that it requires registering all subject's brain images into a common space. In this study, we used the unified segmentation algorithm [Ashburner and Friston (2005)] to register all subject's images to the MNI space. A recent study by [Klein *et al.* (2009)] compared several registration algorithms of healthy control subjects brain images. In this comparison, the DARTEL method that was introduced by Ashburner in 2007 [Ashburner (2007)] as an alternative to the unified segmentation algorithm was found to be more accurate. In our present study, we tested both registration methods on our data. No significant gain in performance was achieved by using DARTEL instead of the unified segmentation approach.

It is difficult to establish a fair comparison between the performance achieved by the SPM analyses in our study and those published over the last few years [Srivastava *et al.* (2005), Bruggemann *et al.* (2007), Thesen *et al.* (2011)] because of the heterogeneity of the patient populations, and annotation and evaluation protocols. Thesen et al. [Thesen *et al.* (2011)], for instance, performed an SPM analysis using surface-based features, cortical thickness and GM/WM contrast. The best performance was found for the cortical thickness and GM/WM contrast. Further comparison of the ROC curve analysis, however,

is difficult because [Thesen *et al.* (2011)] used a definition of TP (a patient with one detected cluster in the lesion area) and FP (healthy control subject with detected clusters) different from that defined in the present study. While we used a fixed Gaussian kernel width of 6 mm to smooth the feature maps, it is unlikely that this parameter is important in the context, as only a marginal influence was found in [Thesen *et al.* (2011)] for typically used values between 5 and 12 mm. Results of the SPM analyses in our study are also in accordance with those obtained by [Bruggemann *et al.* (2007)] showing that conjunction analyses based on GM and WM maps allowed significant performance improvement as compared with the SPM statistical maps derived from a single contrast (WM or GM). It should be noted that the OC-SVM method presented here can easily be extended to any number of additional features.

One drawback of the novelty detection algorithm is that the model does not learn specific patterns of the lesions. We tried to alleviate this limitation by incorporating only features that were previously reported to be discriminant in the task of identifying FCD lesions. [Wagner *et al.* (2011)], for instance, demonstrated a 29% sensitivity gain induced by the use of MRI based junction and extension feature maps to visually detect FCD type IIa lesions in addition to the conventional MR T1-weighted images.

In the present study, we hypothesized that these two features will allow discriminating malformative lesions including FCD and heterotopia from normal brain tissue; choosing just two feature maps also allowed an easy comparison with SPM. The OC-SVM framework, however, offers a great flexibility in adding other types of features such as surface-based measures of cortical thickness investigated in recent studies [Hong *et al.* (2014), Ahmed *et al.* (2015)]. An improvement in performance is however not necessarily guaranteed as it depends heavily on the correlation between the added features and the target lesions and on the number of available training samples.

A future direction is to investigate the possibility of going towards a multi-modal CAD system for intractable epilepsy detection by incorporating additional features extracted from other imaging modalities such as the FLAIR sequence in MRI or FDG-PET scans. While the CAD system developed in this study was designed to detect epileptogenic lesions, the framework depicted in Fig. 5.1 is in principle suitable for detecting other pathologies characterized by small lesions on brain MRI. Examples would include the detection of plaques in multiple sclerosis, vascular white matter lesions in normal ageing and as a risk factor for stroke, focal atrophy in dementia, etc. The feature selection and extraction steps can be adjusted to the specific pathology. Importantly, the OC-SVM outlier detection step and the method to control false positives developed in this chapter could be used for such applications.

Table 5.2: **OC-SVM and SPM classification results for clinical data for a p-value of 0.001. The third column indicates the location of the epileptogenic lesions reported by the clinician for each patient. Columns 4 to 7 report the detection results of OC-SVM and the three SPM analyses. For each method, the number of false positive clusters is indicated in parentheses. The * in column 2 indicates the FCD lesions that were confirmed by histology.**

Patient	**Lesion**	**Location**	**OC-SVM**	**SPM junction**	**SPM extension**	**SPM Global null**
#1 (MRI+)	#1*	precentral gyrus R	✓ (2)	✓ (3)	X (17)	X (1)
#2 (MRI+)	#2*	middle frontal gyrus L	✓ (3)	✓ (4)	✓ (34)	✓ (7)
#3 (MRI+)	#3	superior frontal gyrus R	✓ (0)	✓ (12)	✓ (24)	✓ (9)
#4 (MRI-)	#4	hippocampus R	✓ (2)	X (5)	X (24)	X (5)
	#5*	amygdala R	✓	X	X	X
#5 (MRI-)	#6*	middle frontal gyrus L	✓ (1)	X (3)	X (3)	✓ (0)
#6 (MRI-)	#7	hippocampus R	X (7)	X (11)	X (13)	X (5)
	#8*	temporal R	✓	X	✓	✓
#7 (MRI-)	#9	middle frontal gyrus L	X (3)	X (3)	X (9)	✓ (4)
#8 (MRI-)	#10	frontal L	✓ (10)	X (14)	✓ (56)	X (18)
#9 (MRI-)	#11	anterior temporal lobe R	✓ (4)	✓ (14)	✓ (3)	✓ (2)
#10 (MRI-)	#12	parieto occipital L	✓ (2)	✓ (5)	✓ (25)	✓ (14)
#11 (MRI-)	#13	temporal lobe L	X (1)	✓ (3)	X (23)	X (4)
Sensitivity MRI+			3/3	3/3	2/3	2/3
Mean FP MRI+			1.7 FP	6.3 FP	25 FP	5.7 FP
Sensitivity MRI-			7/10	3/10	4/10	5/10
Mean FP MRI-			3.7 FP	7.2 FP	19.5 FP	6.5 FP
Overall Sensitivity			10/13	6/13	6/13	7/13
Overall Mean FP			3.2 FP	7.0 FP	21.0 FP	6.3 FP

III Optimized outlier detection

Chapter 6

Robust outlier detection

In the previous Chap. 5 we presented a CAD system for the detection of lesions underlying epilepsy. The proposed pipeline built upon outlier detection algorithms such as OC-SVM and SVDD. Experiments using both realistic simulations and clinical patient data were used to show the higher performance of the proposed CAD system in comparison with optimized SPM analyses. It also compared favourably with 2 step state of the art methods. In this chapter, we will investigate the proposed framework sensitivity to noise. We start by showing the behaviour of the CAD system when labelling errors are present. To handle these uncertainties in the training data, we propose a reformulation of the SVDD problem based on an l_0 cost term. The solution to the new formulation is then obtained by iteratively solving a reweighted l_1 penalized problem. This formulation is compared to a state of the art method that also proposes a reweighting scheme using simulation data and datasets from the UCI repository.

6.1 Motivation

Fig. 6.1 shows the output of the CAD system presented in 5 for two distinct patients. In this example, the OC-SVM classifier was replaced by the SVDD classifier. Like for OC-SVM, the optimal hyper-parameters of the SVDD model were derived by using the leave-one-out error estimate introduced in 4.4 and averaged over 4000 voxels.

When examining the clusters outputted by the CAD system for the two patients, it is clear that some of the clusters and in particular the cluster at the cross-hairs center is detected in both scans. In both cases, these clusters correspond to false positive detections (do not coincide with the epileptogenic zone). It is very unlikely to have two distinct

patients presenting the exact same deviation from normal patterns at the same location and with the same extent. We can therefore conclude that these clusters can correspond to false positive detections that are detected regardless of the considered test patient. This also suggests that the SVDD model that was leaned at these voxel locations did not capture well enough the domain of the normal training observations.

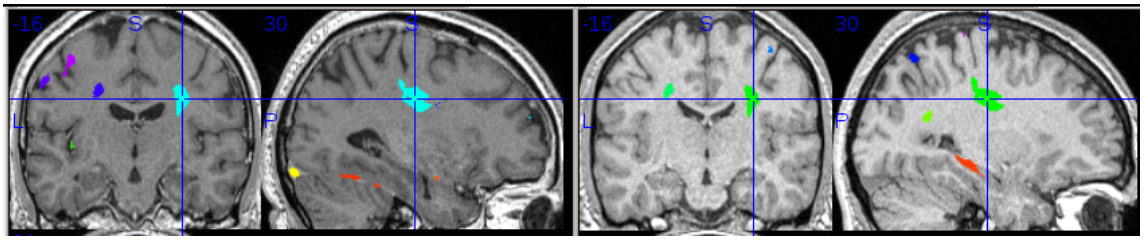
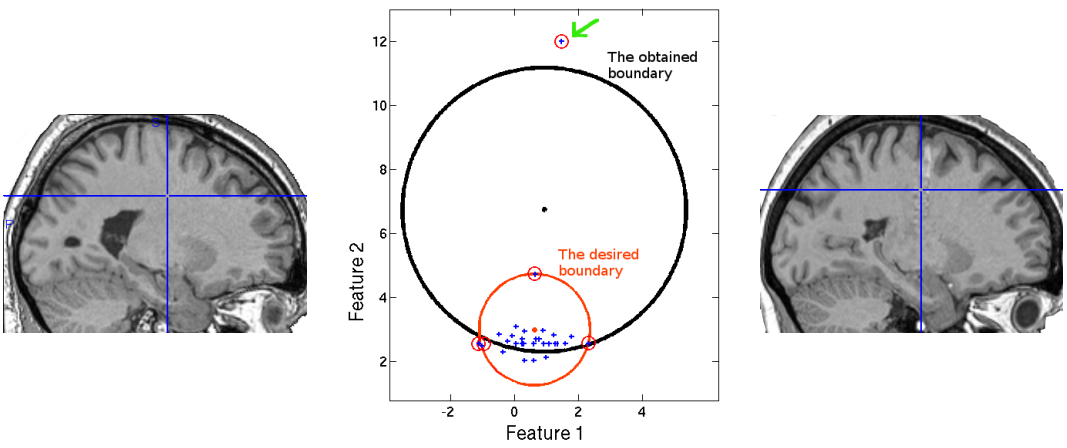


Figure 6.1: Sagittal and coronal example slices of the labelled cluster map outputted by the CAD system for two distinct patients.

To explain the systematic presence of these clusters, we visualized the SVDD decision boundary at the voxel location indicated with the cross-hairs in Fig. 6.2. This decision boundary was computed using only the two most discriminant features (the junction map and the extension map) and training observations corresponding to healthy control subjects.



(a) (b) (c)

Figure 6.2: (a) Sagittal slice from a healthy control subject centered on the cluster indicated by the cross-hairs in Fig. 6.1, (b) the distribution of training observations at the voxel location indicated by the cross-hairs in (a), the black decision boundary correspond to the decision obtained by training the SVDD classifier, the boundary in red is the desired decision boundary that best describes the target data domain. (c) Sagittal slice from the healthy control subject scan corresponding to the outlying observation marked with the green arrow in (b).

The training observation distribution in Fig. 6.2-b shows the presence of an 'abnormal' normal observation indicated with the green arrow. When examining the healthy control subject scan corresponding to this outlying normal observation (see Fig 6.2-c), we found that the MRI scan has an artefact at this location and should therefore not be con-

sidered as a normal observation. This indicates that our training dataset contains falsely labelled observations that were not detected during the visual inspection of the healthy control subject's dataset. In the literature this type of noise is refereed to as label noise. This type of noise has different sources and naturally occurs when human experts (or not) are involved in the labelling process. In some cases such as in medical applications, the subjectivity of the labelling task and using different experts also introduces label noise in the training datasets.

One possible explanation of the poor performance obtained by the CAD system can be the fact that the SVDD algorithm does not allow handling falsely labelled observations. Another possible explanation could be that the hyper-parameter optimization procedure (*i.e.* selecting only 4000 voxels from the entire volume of interest and optimizing the average detection error) is suboptimal. In the next section, we will try to explain this bad performance by going back to the original formulation of the SVDD algorithm.

6.2 Hinge loss and sensitivity to label noise

As explained in Chap. 4, domain-based algorithms for outlier detection consist in learning the compact representation domain of a *normal* class (also called the target class), in view of predicting whether a test sample belongs to this compact description. In this category, the SVDD algorithm [Tax and Duin (2004)] hypothesizes that the normal data belong to a hypersphere characterized by a center $\mathbf{a}$ and a radius R. Let us recall the constraint-based optimization problem associated with finding optimal $\mathbf{a}$ and R for properly chosen positive parameter C:

$$\begin{cases} \min\limits_{R,\mathbf{a},\boldsymbol{\xi}} & \underbrace{R^2}_{\text{structural error}} + C \underbrace{\sum_{i=1}^{n} \xi_i}_{\text{empirical error}} & \\ \text{subject to} & ((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i, & i \in [1, n] \\ \text{and} & \xi_i \geq 0, & i \in [1, n]. \end{cases} \tag{6.1}$$

The empirical error in problem 6.1 corresponds to a Hinge loss. Minimizing the Hinge loss results in penalizing more heavily large error values ξ_i than smaller ones. The cost of rejecting an observation $\mathbf{x}_i$ being proportional to $C\xi_i$.

The presence of uncertain samples or wrongly labelled observations x_j will likely generate high values of ξ_j and result in a significant increase of the structural error term in equation 6.1. Of course the C parameter can be used to balance the two errors in the objective function. Tuning this parameter is however very difficult especially when no *a priori* knowledge about the presence of such uncertain observations is available. Fig. 6.3 shows an illustrative example of the impact of such an *outlier* normal point on the prediction of the hypersphere decision boundary. Choosing very low values for C ($C = \frac{1}{16}$

and $C = \frac{1}{8}$) allows suppressing the influence of the outlying observation on the decision boundary. However, it also results in increasing the number of support vectors and in a very tight decision boundary that would not generalize well when tested with new observations. On the contrary, high values of C will favour including the outlying observation inside the decision boundary and often times result in a hypersphere whose center $\mathbf{a}$ is not representative of the center of mass of the target class (see also Fig. 6.2-b).

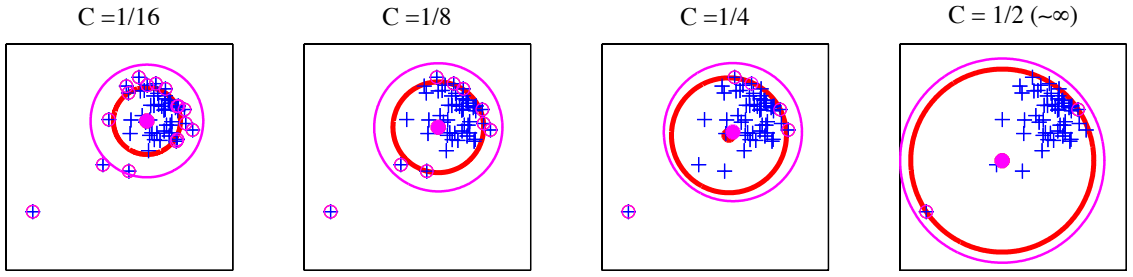


Figure 6.3: Example of SVDD solutions with different C values, $m = 0$ (red) and $m = 5$ (magenta). The circled data points represent support vectors for both m.

Learning from observations without putting much effort into handling label noise often results in a decrease in detection performance and incorrect data description boundaries. In the motivation example in Sec. 6.1, the presence of an outlying observation in the training dataset resulted in having a too large hypersphere that do not allow defining a good representation of the normal class. Two kinds of detection errors can be made when using this unrepresentative boundary: 1) as illustrated in Fig. 6.1, false positive detections which can result from having a hypersphere that is shifted towards the outlying observation, and/or 2) false negative detections which are mainly due to having a too large hypersphere. For instance, in the example in Fig. 6.2, a lesion represented in the feature space by a test data point lying inside the black decision boundary but far from the majority of the training observations will not be detected by the model.

In the motivation example, a hyper-sphere was directly fit into the data without projection using a kernel function. One can argue that potentially, the influence of the outlying observation can be reduced by considering more flexible descriptions that would reduce the domain captured by the boundary. This would in turn require fine tuning of the kernel hyper-parameter and would in some cases lead to too complex decision boundaries.

6.3 State-of-the-art methods for handling label noise

Dealing with the presence of label noise in the training observations has recently attracted a lot of interest in the machine learning community and especially in classification related tasks [Frénay and Verleysen (2014)]. In the literature, three main approaches have been investigated to handle label noise.

- Probabilistic label noise tolerant models: correspond to probabilistic methods that explicitly model label noise and learn it from the training observations. These include

Bayesian approaches where priors on the mislabelling probabilities are used and mixture models in which an observation is assumed to be generated either from a normal distribution or from an outlying distribution. To some extent this is very related with outlier detection methods that are based on probabilistic models.

- Data cleansing methods: correspond to 'preprocessing' of the data to identify or remove mislabelled observations prior to learning the model. Filter methods are most commonly used. They often operate in two steps. First, the output of a classifier, an outlier measure or an ensemble of classifiers (also boosting-based methods) learned using the dataset with label noise is thresholded to identify and/or remove the mislabelled training observations. A second classifier is then learned using the cleansed training dataset.

- Robust loss functions: standard losses used in well-known machine learning algorithms are not robust to the presence of label noise. For instance, the log loss associated with logistic regression and the Hinge loss, which corresponds to SVMs and SVDD are not completely label noise robust (see Sec.6.2). More robust losses have been proposed. In [Lin *et al.* (2004), An and Liang (2013)] the hinge Loss in the SVM formulation was adapted by weighting the contribution of each training observation. The weights however are either computed using heuristics or correspond to a confidence score provided by an expert [Niaf *et al.* (2011), Niaf *et al.* (2014)].

In their paper, [Frénay and Verleysen (2014)] argue that considering probabilistic models (*e.g.* mixture model) or two step approaches for the identification of mislabelled observation amounts to outlier detection. This is not necessarily true because most outlier detection methods make assumptions about the nature of outliers. The most common assumptions are that outliers lie in low density regions or outliers are associated with low probability events. The same assumptions cannot necessarily be made about mislabelled observations.

There have been a few attempts to improve the SVDD performance in the presence of label noise. [Shawe-Taylor and Cristianini (2004)] proposed to introduce a margin parameter m in the first constraint in problem 6.1 to deal with uncertain observations. The new constraint is given by: $((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq m + R^2 + \xi_i, i \in [1, n]$ It allows setting a confidence margin around the target data. The authors even provide a generalization error bound on the probability of an observation being outside of the decision boundary (p. 200 theorem 7.5 in [Shawe-Taylor and Cristianini (2004)]). In practice, this theoretical bound is too loose and cannot be used to select the margin parameter. In Fig. 6.3, two values of the m parameter were used. $m = 0$ correspond to the standard formulation of the SVDD algorithm. For small values of C and a well chosen value of m, the solution provided by the m-SVDD successfully handle the presence of the outlying observation but at the cost of a large number of support vectors and thus does not scale.

More recently, [Lee *et al.* (2007)] and [Liu *et al.* (2013)] proposed to handle the

presence of uncertain observations by considering a different empirical error (loss) instead of the Hinge loss. They introduced a weighting term for each training observation. The weights are intended to constrain the impact of the ξ_i in problem 6.1. Keeping the same constraints as in problem 6.1, the new objective function is given by:

$$\min_{R,\mathbf{a},\boldsymbol{\xi}} R^2 + C\sum_{i=1}^{n} w_i\xi_i$$

In both studies, the weight terms are set beforehand. In density induced SVDD (DI-SVDD) [Lee *et al.* (2007)], the weight w_i for observation $\mathbf{x}_i$ corresponds to a measure of the relative density degree for observation $\mathbf{x}_i$ computed either by considering a nearest neighbourhood approach or a kernel density estimator (Parzen-window). In Liu's formulation [Liu *et al.* (2013)], the weight w_i for observation $\mathbf{x}_i$ correspond to a confidence score computed by considering the kernel-based distance (*i.e.* in the feature space) between observation $\mathbf{x}_i$ and the center of mass of all training observations.
An interesting aspect of both reformulations is that for given weights w_i, the corresponding dual problem is very similar to the standard SVDD dual problem. The only difference is in the upper bound of the inequality constraint on the Lagrange multipliers: $0 \leq \alpha_i \leq w_i C$. Using the reference intrusion detection database and databases from the UCI repository, [Liu *et al.* (2013)] showed that their approach outperforms the standard SVDD and the alternate DI-SVDD [Lee *et al.* (2007)].

6.4 Our contribution: L_0-SVDD

Weighting schemes seem to be a good way of reducing the effect of uncertain observation in the learned decision boundary. One limitation however is that their performance relies heavily on the heuristic that was chosen to measure the confidence in the label of a given observation. For instance, in DI-SVDD, a density argument was considered. It is likely to work very well for datasets in which uncertainties lie in low density region and would potentially fail if the distribution of the uncertainties has a given structure.

We propose a reformulation of the SVDD algorithm in the context of outlier detection in presence of label noise. For this purpose, we go back to the most natural loss that is the 0-1 loss which gives a cost of 1 in case of error and 0 otherwise. We show that the new formulation can be approximated by using a log penalty and solved efficiently by iteratively solving a weighted convex l_1 penalized problem.

6.4.1 Formulation

We consider the problem of detecting outliers from a set of n observations of p dimensional vectors stored in $\mathcal{X}$ a $n \times p$ matrix. Regarding the outlier detection problem in presence of label noise, it looks relevant to consider the zero-norm cost function [Forero *et al.* (2012)]. The l_0 cost is defined as $\|\mathbf{z}\|_0 = card\{i|z_i \neq 0\}$. The advantage of taking

such a non convex cost is well motivated for instance in [Antoniadis *et al.* (2011)] where it is shown that the resulting estimator is asymptotically unbiased.

Based on this cost design for outlier detection, we propose to define the L_0-SVDD problem as follows, for a given C:

$$\begin{cases} \min\limits_{\mathbf{a},R,\boldsymbol{\xi}} & R^2 + C\|\boldsymbol{\xi}\|_0 \\ \text{with} & ((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i \\ & \xi_i \geq 0 \qquad i \in [1, n]. \end{cases} \tag{6.2}$$

In this formulation, unlike for the Hinge loss, the empirical error term corresponds simply to the number of errors and is no longer proportional to the amplitude of the errors.

The use of an l_0 term in an objective function have previously been investigated in the context of variable selection (*i.e.* selection of a small subset of variables from a large set). Like classification, variable selection is also often formulated as a minimization problem where the objective function is the sum of a loss dependent on the data (the equivalent of the empirical error) and a regularization term (the equivalent of the structural error). As the selection problem amounts to finding a sparse variable vector with only a few non null entries, the l_0 penalty was considered in this context [Candès *et al.* (2008), Gasso *et al.* (2009)]. The main difference with the problem formulation in 6.2 is that we propose to use the l_0 term for the loss term (empirical error) instead of the regularization term (structural error).

Unfortunately the l_0 loss is non differentiable, combinatorially hard, and does not lead to an effective algorithmic approach.

6.4.2 Logarithmic relaxation and DC programming

To obtain an efficient technique for solving problem 6.2, two key insights are needed. The first key step is to approximate the l_0 loss by another loss function. Depending on the context, various approximations have been proposed (see examples in [Gasso *et al.* (2009)]).

Following [Weston *et al.* (2003)],we propose to replace the l_0 pseudo-norm by its logarithmic approximation. Fig. 6.4 compares the l_0 to its logarithmic approximation. Note that in our case $\xi_i \geq 0$.

In the context of variable selection and sparse signal approximations, [Candès *et al.* (2008)] have also investigated the use of this log penalty and empirically prove the nice capability of the resulting estimator for recovering sparsity.

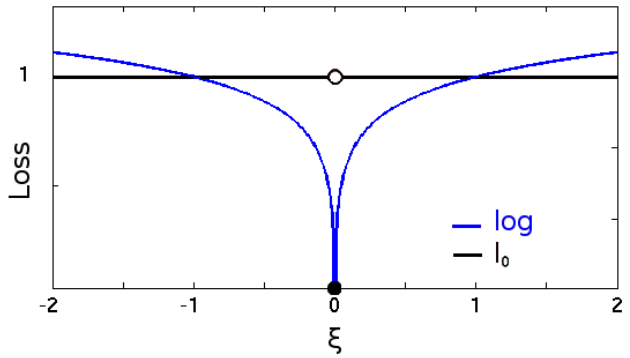


Figure 6.4: l_0 loss versus its logarithmic approximation. The log approximation was scaled for a better visualization of the two losses.

This leads to the following problem, for given parameters C and γ:

$$\begin{cases} \min\limits_{\mathbf{a},R,\boldsymbol{\xi}} & R^2 + C\sum\limits_{i=1}^{n} \log(\gamma + \xi_i) \\ \text{with} & ((\phi(\mathbf{x}_i) - \mathbf{a}) \cdot (\phi(\mathbf{x}_i) - \mathbf{a})) \leq R^2 + \xi_i \\ & \xi_i \geq 0 \qquad i \in [1, n]. \end{cases} \tag{6.3}$$

Problem 6.3 remains non convex. Our second key idea is to solve this problem by using an iterative procedure solving at each iteration a convex QP problem, based on the decomposition of the non-convex function as a difference of convex functions (DC) [An and Tao (2005), Gasso *et al.* (2009)]. Difference of convex functions programming (DC) is a generic and principled way for solving non-smooth and non-convex optimization problems. Recently [Rakotomamonjy *et al.* (2016)] proposed a new algorithm based on the DC framework for solving optimization problems where both the loss function and the regularization are non-convex but can be expressed as a difference of convex functions. The authors also provide more theoretical guarantees.

In our case, the regularization (structural error) is convex and only the loss function (the logarithmic approximation) is non convex. The decomposition is rather simple and is of the form:

$$\log(\gamma + t) = f(t) - g(t) \qquad \text{with } f(t) = t \quad \text{and} \quad g(t) = t - \log(\gamma + t),$$

both functions f and g being convex.

The DC framework consists in minimizing iteratively (R^2 plus a sum of) the following convex term:

$$f(\boldsymbol{\xi}) - \nabla g(\boldsymbol{\xi}^{\text{old}})\boldsymbol{\xi} = \boldsymbol{\xi} - \left(1 - \frac{1}{\gamma + \boldsymbol{\xi}^{\text{old}}}\right)\boldsymbol{\xi} = \frac{\boldsymbol{\xi}}{\gamma + \boldsymbol{\xi}^{\text{old}}},$$

where $\boldsymbol{\xi}^{\text{old}}$ denotes the solution at the previous iteration and function g have been replaced by its linear approximation locally.

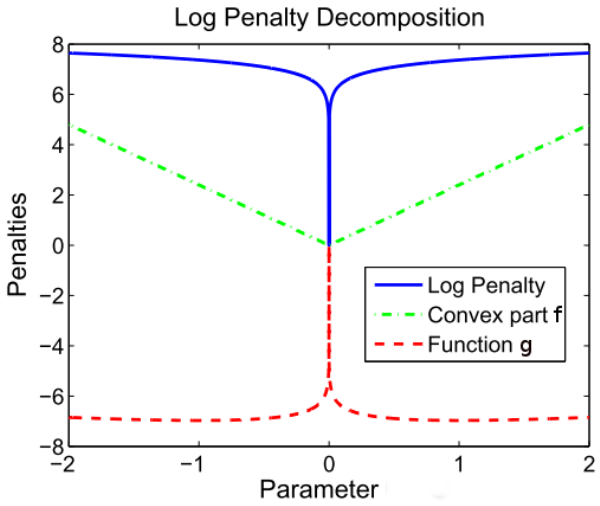


Figure 6.5: DC decomposition for the logarithmic approximation [Gasso et al. (2009)].

The DC idea applied to our L_0-SVDD approximation consists in building a sequence of solutions of the following adaptive SVDD:

$$\begin{cases} \min\limits_{\mathbf{a},R,\boldsymbol{\xi}} & R^2 + C\sum\limits_{i=1}^{n} w_i\xi_i \\ \text{with} & ((\phi(\mathbf{x}_i)-\mathbf{a})\cdot(\phi(\mathbf{x}_i)-\mathbf{a})) \leq R^2+\xi_i \qquad \text{with} \qquad w_i = \dfrac{1}{\gamma+\xi_i^{\text{old}}}. \\ & \xi_i \geq 0 \qquad i\in[1,n] \end{cases} \tag{6.4}$$

Stationary conditions of the KKT give: $\mathbf{a} = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{n}\alpha_i = 1$, where the α_i are the Lagrange multipliers associated with the inequality constraints $((\phi(\mathbf{x}_i)-\mathbf{a})\cdot(\phi(\mathbf{x}_i)-\mathbf{a})) \leq R^2+\xi_i$. The dual of this problem is [Liu *et al.* (2013)]:

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \boldsymbol{\alpha}^\top K\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top diag(K) \\ \text{with} & \sum_{i=1}^{n}\alpha_i = 1 \qquad\qquad 0\leq\alpha_i\leq Cw_i \qquad i\in[1,n], \end{cases} \tag{6.5}$$

where $diag(K)$ corresponds to the diagonal of the matrix K of elements $K_{ij} = (\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)) = K(\mathbf{x}_i,\mathbf{x}_j)$.

This is a classical QP with box constraints that can be solved efficiently using for instance an effective active set solver[1] [Loosli and Canu (2007)]. Note that if λ denotes the Lagrange multiplier associated with the equality constraint $\sum_{i=1}^{n}\alpha_i = 1$, we can see that $R = \lambda + \mathbf{a}^\top\mathbf{a}$ by calculating the dual of (6.5), that is the bidual.

It is worth noting that formulation 6.4 corresponds exactly to the formulations proposed by [Lee *et al.* (2007),Liu *et al.* (2013)]. In our case however the weights w_i associated with each training observation are adapted iteratively.

[1] available at asi.insa-rouen.fr/enseignants/~arakoto/toolbox

6.4.3 L_0-SVDD algorithm

Put all together, this leads to the following algorithm 1.

Data: X, C , γ

Result: R , $\mathbf{a}$, $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$

$w_i = 1; \quad i = [1, n];$

while *not converged* **do**

$(\boldsymbol{\alpha}, \lambda) \leftarrow$ solve_QP$(X, C, \mathbf{w})$ % solve problem (6.5)

$\mathbf{a} \leftarrow X^\top \boldsymbol{\alpha}$;

$R \leftarrow \lambda + \mathbf{a}^\top \mathbf{a}$;

$\xi_i \leftarrow \max(0, \|\mathbf{x}_i - \mathbf{a}\|^2 - R^2) \quad i \in [1, n]$;

$w_i \leftarrow 1/(\gamma + \xi_i) \quad i \in [1, n]$;

end

Algorithm 1: L_0-SVDD for the linear kernel

6.5 Experiments

6.5.1 Synthetic data results

Learning dataset In the first experiment, we generated learning examples for the normal class by drawing $n = 25$ pseudo-random normally distributed samples of dimension $p = 2$. The mean was set to 1 for the first dimension and to 2 for the second dimension. To simulate the presence of outliers in the learning dataset, we added examples that fall outside the range of the distribution of the normal class (see Fig. 6.6 left). The hyper-parameters were fixed as follow: linear kernel, $\gamma = 1$, $nb\_iter = 3$ and $C = 0.4$.

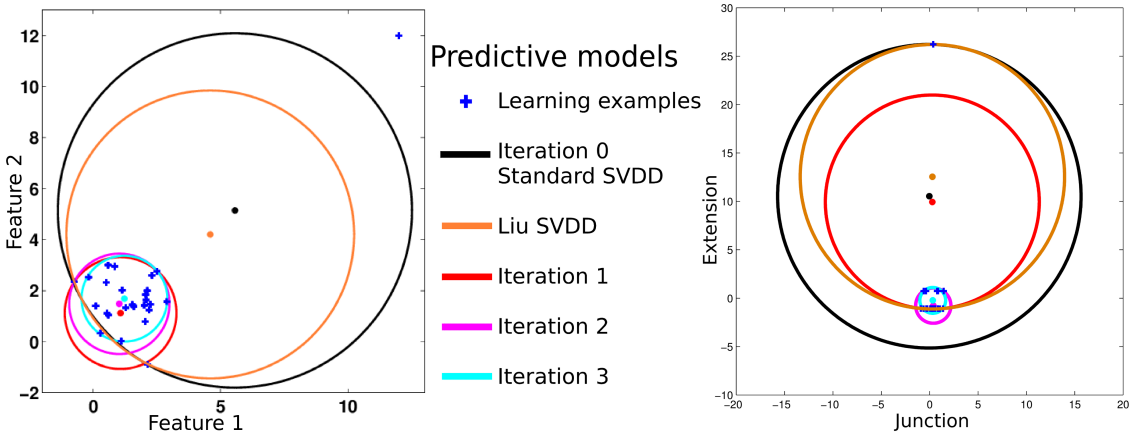


Figure 6.6: Predictive models obtained using synthetic and realistic learning examples with an outlier.

Prediction results The predicted models, given in Fig. 6.6 (left), show that 1) the L_0-SVDD decision boundary is the one that best fits the learning data 2) Liu-SVDD sphere center is still influenced by the presence of an outlier among the learning examples.

6.5.2 Realistic data

In the second experiment, we compared the performance of the Liu-SVDD and the L_0-SVDD methods in detecting abnormalities in magnetic resonance images (MRIs) of patients suffering from intractable epilepsy on a voxelwise basis.

Learning database In Chap. 5 we proposed a CAD system for the detection of epilepsy lesions based on the SVDD algorithm. Two parametric maps extracted from the MR scans, namely the junction and extension maps were used to discriminate between controls and patients suffering from intractable epilepsy. The same learning database, pre-processing steps of the MRIs and feature extraction step were used in this experiment. Two classifiers, L_0-SVDD and Liu-SVDD, were learned for each voxel k using the matrix $M^k \in M_{n,p}(\mathbb{R})$, $n = 29$ and $p = 2$. The value $C = 0.6$ was obtained from a grid search analysis performed on a set of 4000 voxels randomly selected from the 1.5 million voxels of the brain MRI. The analysis of the training data in M^k indicated that some clusters of voxels contained uncertain data resulting from artefacts in the original scan or from image processing issues. Fig. 6.6 (right) illustrates the presence of such uncertain data in a voxel belonging to the cluster highlighted in green in Fig. 6.7 (left). In this illustration, the distribution of the two features (junction and extension) was computed over the 29 control subjects from the learning database.

Test data In the MRI of a control subject, at the known location of one of the clusters containing uncertain 'normal' data, we simulated an heterotopy like lesion, that is an abnormal extension of the grey matter into the white matter. We locally changed the grey level values of the voxels within the white matter in the original MRI, to make them correspond to the grey matter distribution. The simulation procedure was described in more details in Chap. 5. Fig. 6.7 (left) shows the obtained lesion. The outlier cluster location was found while testing a standard SVDD built on the learning examples and noticing the systematic detection of the same cluster regardless of the considered test data. The suspected presence of outliers in the learning database was then confirmed by inspecting the values of the features at the cluster location (see the motivation example in Sec. 6.1).

Prediction results Fig. 6.7 shows that the L_0-SVDD classifier detected most of the lesion (DICE of 80% for the example slice in Fig. 6.7 and 53% for the whole lesion) while Liu-SVDD classifier failed in retrieving the lesion (DICE of 0%). The lesion detected by L_0-SVDD is bigger than the simulated one because the images were smoothed in the pre-processing step to increase the signal to noise ratio and thus to reduce the feature noise.

6.5.3 UCI datasets

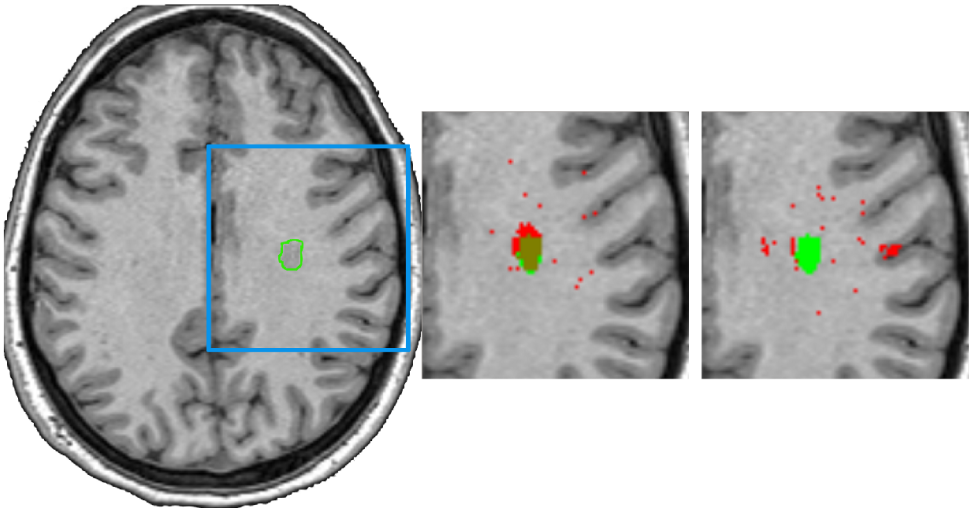


Figure 6.7: Transverse MR slice of the control subject showing: the simulated lesion highlighted in green in all images, L_0-SVDD (middle, in red) and Liu-SVDD (right, in red) classification results.

Data description We used four datasets from the UCI repository[2] to compare the performance of L_0-SVDD with the original SVDD algorithm. Tab. 6.1 gives the properties of the four datasets. These datasets are often used to evaluate the performance of binary or multi class classification algorithms. [Tax and Duin (2004)] proposed using these datasets in the context of outlier detection by considering for each dataset, one of the classes as the target class and the remaining classes as the outlier class.

Data partitioning We propose to use the majority class (the class with the highest number of examples) as the target class and the other class as the outlier class. For the Balance dataset, we tested two configurations: (class 1 + class 2) as the target class (referred to as Balance 12vs3) and (class 1 + class 3) as the target class (referred to as Balance 13vs2).

In all our experiments, we used 10 fold cross validation on the target class. This corresponds to partitioning the target class data into 10 folds, 9 folds are used to train the SVDD and the L_0-SVDD models, and the remaining fold and the outlier class are used to test the model.

Parameter search and performance evaluation For both SVDD and L_0-SVDD, the Gaussian kernel was used for all datasets. The kernel width σ was varied in the interval $[2^0, 2^1, ..., 2^{15}]$ and the hyper-parameter ν was varied in the interval $[2^{-4}, 2^{-3.5}, ..., 2^{-0.5}, 2^0]$. For L_0-SVDD, the number of iterations was set to 4 and the parameter γ was set to 1.

The area under the ROC curve (AUC) was used to evaluate the performance of the SVDD and L_0-SVDD algorithm.

[2]available at http://archive.ics.uci.edu/ml/

Dataset	Class	Nb of examples	Nb of features
Balance	class 1	288	4
	class 2	49	
	class 3	288	
Blood Transfusion	class 1	570	5
	class 2	178	
Breast Cancer	class 1	444	9
	class 2	239	
SPECTF Heart	class 1	212	44
	class 2	55	

Table 6.1: Description of the UCI datasets.

Experiment 1: performance under different parameter In this first experiment, we evaluate the average AUC performance of SVDD algorithm and L_0-SVDD over the 10 folds and the 9 values of parameter ν.

Comparison of the average performance of SVDD and L_0-SVDD was not conclusive as it depended on the considered dataset. In Fig. 6.8 we consider the results obtained for two distinct datasets: the Blood Transfusion dataset and the Breast Cancer dataset.

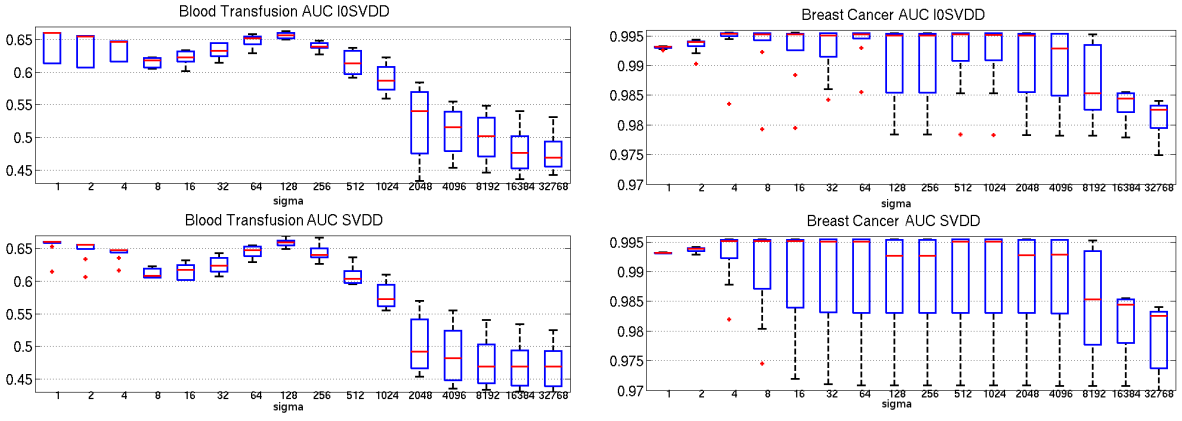


Figure 6.8: Performance of L_0-SVDD (top) and SVDD (bottom) for varying values of σ (x axis) for (a) the Blood Transfusion dataset and (b) the Breast Cancer dataset.

For the Blood Transfusion dataset (Fig. 6.8-a), no gain in performance was obtained by using the L_0-SVDD algorithm instead of the standard SVDD algorithm. For the Breast Cancer dataset (Fig. 6.8-b), L_0-SVDD shows less variability than SVDD with respect to the value of ν and the data partitioning.

Another interesting result is that the best performance for SVDD and L_0-SVDD is obtained for the same range of σ values. For instance, for the Blood Transfusion dataset, SVDD and L_0-SVDD give the highest performance with minimum variability for σ around 128, whereas for the Breast Cancer dataset, the highest performance is reached for low values of σ for both algorithms.

Experiment 2: sensitivity to label noise Experiment 1 showed that no notable gain in performance was observed after using the L_0-SVDD algorithm instead of the standard SVDD algorithm. This result suggests that label noise may not be naturally present in the considered UCI datasets.

In [Liu *et al.* (2013)], sensitivity to the presence of noise in the training observations was investigated by artificially adding a Gaussian noise component to each feature vector. To simulate the presence of label noise in the training dataset, we propose to make use of the outlier class by including observations from the outlier class in the training dataset after flipping their labels. The same cross validation framework is used in this experiment, the only difference being that a fraction of observations from the outlier class are added to the nine training folds to form the training dataset. This same fraction was excluded from the testing dataset.

Different noise levels were considered by varying the corresponding fraction of noisy observations in the interval [0, 0.01, 0.02, 0.05, 0.1, 0.2] when possible. To select the noisy observations we proceeded as follows: first we trained an SVDD model without noise. We then evaluated this model on the test observations from the outlier class. Next, we identify the test observations from the outlier class that were successfully rejected (*i.e.* true negative detections) by the learned model. Finally, a fraction from the rejected observations is selected and added to the training data to simulate the presence of label noise.

For each dataset, the ν value and the kernel width σ, were chosen based on the results of our first experiment. For all datasets, these values corresponded to the approximate values that allowed obtaining the best performance when no noise was added to the training dataset. Fig. 6.9 shows the results obtained for all five UCI datasets.

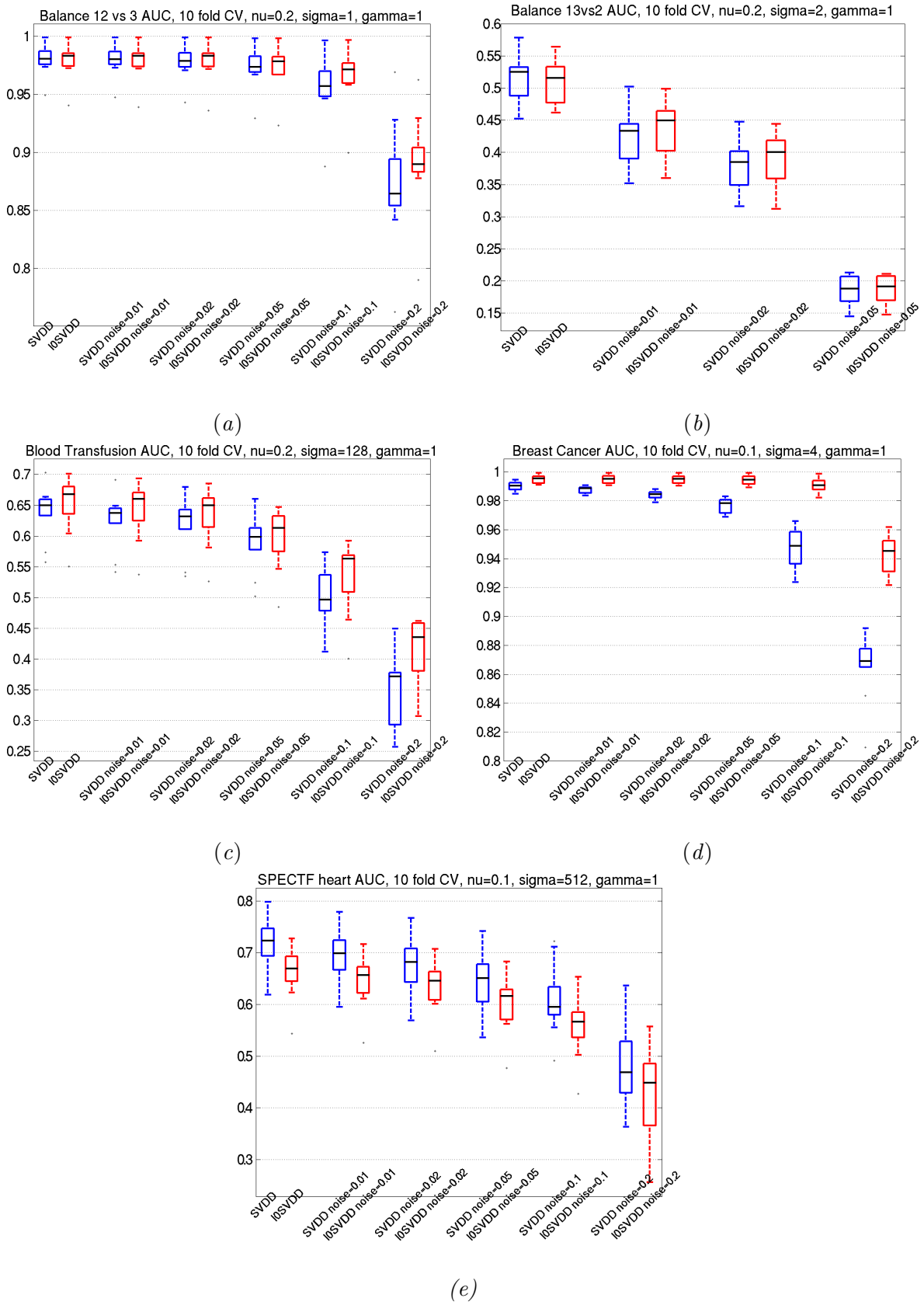
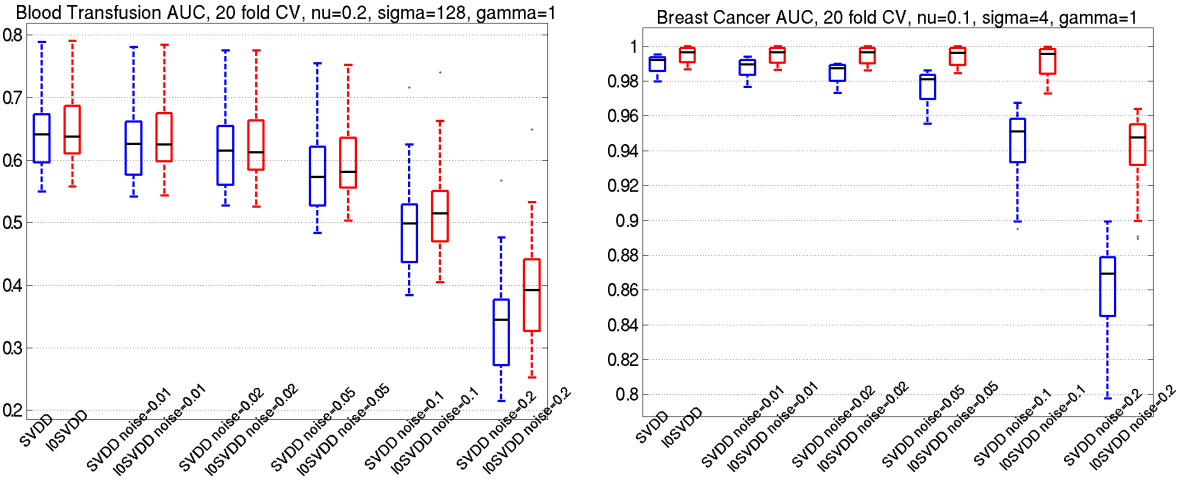


Figure 6.9: Sensitivity of SVDD and L_0 to label noise for both Balance dataset configurations (a) and (b), the Blood Transfusion dataset (c), the Breast cancer dataset, and (e) the SPECTF Heart dataset. For each dataset, the values of the hyperparameters are indicated at the top of the corresponding figure.

Overall performance of both SVDD and L_0-SVDD algorithms decreases when the noise level increases. For all dataset, except the SPECTF Heart dataset, the decay in performance for the L_0-SVDD is slower than that of the SVDD algorithm. For the SPECTF Heart dataset, performance of the L_0-SVDD algorithm is consistently lower than that of the SVDD algorithm for all considered noise levels. Compared with the other considered UCI datasets, the SPECTF Heart dataset has a higher dimension (44 features) and a lower number of training observations. This may explain the lower performance of L_0-SVDD and the rapid decay in performance of SVDD and L_0-SVDD algorithms when noisy observations are added.

To test the effect of the number of folds used in the cross validation, we repeated the same experiment with 20 folds instead of 10. Fig. 6.10 shows the results obtained for the Blood Transfusion dataset and the Breast Cancer dataset. In this case, the same decay in performance is observed with L_0-SVDD being less sensitive than SVDD to the presence of label noise in the training observations.



(a) (b)

Figure 6.10: Sensitivity of SVDD and L_0-SVDD to label noise for (a) the Blood Transfusion dataset and (b) the Breast Cancer dataset. The AUC estimates were computed using 20 fold cross validation.

6.6 Conclusion

In this chapter, we proposed a new SVDD formulation based on the l_0 loss to efficiently handle the presence of uncertainties in the training dataset. The associated problem was solved using difference of convex functions and quadratic programming. The conducted experiments on synthetic and realistic clinical data show that, unlike state of the art methods, the L_0-SVDD approach successfully reduces the effect of uncertain data on the predicted decision boundary. We also evaluated the performance of the proposed approach using five datasets from the UCI repository. The presence of label noise was simulated by adding observations from the outlier class to the training dataset. Our results showed that the proposed L_0-SVDD formulation is less sensitive to the presence of label noise in the training data. One of the limits of the experiments using the UCI datasets is the fact that the hyper-parameters were selected using the AUC value as a criterion. In a true outlier detection scenario, observations from the outlier class are not available during model selection to compute the AUC value. Our attempts at finding the optimal parameters by using only the leave-one-out estimate of the target class error used in Chap. 5 did not give satisfactory results. The unsuitability of using this error estimate to find the best parameters of the L_0-SVDD model is perhaps due to the fact that minimizing the target error is in contradiction with allowing more easily some observations to lie outside of the decision boundary. This is even more true when the number of training observations is low or insufficient for describing the whole target class distribution. Another limit of the present study is that the value of the γ parameter of the L_0-SVDD algorithm was set to 1 in all our experiments. Previous studies [Candès *et al.* (2008), Gasso *et al.* (2009)] that used an l_0 cost term, investigated the influence of this parameter. They showed that adapting the value of this parameter following an annealing design can help with convergence and reduce the effect of the initialization of the optimization problem.

Chapter 7

Multi-modal outlier detection

7.1 Motivation

In Chap. 1 we discussed the role played by neuroimaging data in the pre-surgical evaluation of patients suffering from intractable epilepsy. MRI is the mainstay imaging technique in finding potentially surgically removable abnormalities. Multi-parametric MRI associating T1 weighted (T1-w), FLAIR and more recently diffusion weighted sequences (DTI) improves the non-invasive in vivo diagnostic performance [Rugg-Gunn *et al.* (2001), Focke *et al.* (2008)]. The integration however of this large amount of visual information remains a complex task and often lacks sensitivity. Invasive intracranial EEG based on stereotactic implantation of depth EEG electrodes (SEEG) remains the reference exam for pre-surgical planning.

To help the radiologist facing the challenging task of non-invasively detecting the EZ on MR images, we proposed in Chap. 5 a CAD system based on multivariate analysis of different features extracted from T1-w sequences. The system is based on the OC-SVM algorithm associated with six image-based parameters modelling the clinical description of focal cortical dysplasia, a small brain malformation often causing focal epilepsy. These included the probabilities of belonging to one of the three main brain tissue classes (GM, WM and CSF) as well as three other features characterizing the GM extension, the GM/WM junction, and the cortex thickness. The CAD system produced an overall good classification and localisation performance, outperforming optimized SPM analysis (see Chap. 5). It successfully detected 10/13 lesions with a high sensitivity for MRI+ lesions (3/3), and an average of 1.7 false positive detections for MRI+ and 3.7 false positive detections for MRI- cases.

In this chapter, we focus on improving the diagnostic performance of the CAD system when facing difficult cases. We hypothesize that a significant diagnostic gain can be achieved by complementing the original feature set by other parameters derived from the FLAIR and/or DTI MR sequences. An optimal data fusion strategy that maximizes the detection performance is proposed. The detection results for three patients are then cross-validated with an epileptogenicity index map derived from the reference SEEG exam.

7.2 State-of-the-art: data fusion methods

Data fusion, can be viewed as a special application of ensemble learning [Polikar (2012)]. It corresponds to learning from measurements that were recorded using various sensors. Combining these measurements can provide complementary information and increase the performance of an automated decision making system [Atrey *et al.* (2010)]. In [Lahat *et al.* (2015)], the authors review different practical application domains in which multimodality data (data from different sources) is available and can be efficiently exploited to better extract knowledge from the data for various purposes. These application domains include multisensory systems, biomedical and health related systems and environmental studies. Multimodal information fusion is however subject to many challenges. Three key questions have to be addressed when investigating a fusion strategy. The first one is of course what to fuse? then comes the question when to fuse? and finally how to fuse? The answer to the first question is very much related to the application. For instance, in medical applications and in particular in disease diagnosis, the patient undergoes several exams (imaging, physical examination, biopsies, blood tests, etc.) that all potentially provide complementary information that is analysed to obtain a successful disease diagnosis. [Calhoun and Adali (2009), Lahat *et al.* (2015)] In most cases, *a priori* knowledge help greatly in deciding what information should be combined. When no such knowledge is available, the problem of finding which information to fuse becomes very similar to the feature selection problem discussed in Chap. 2. In this section we will focus on the two remaining questions.

7.2.1 Fusion levels

Fusion level refers to the level of the step consisting of merging the information coming from different sources. For simplicity we will consider only two main fusion levels: the early fusion level (also called the feature fusion level) and the late fusion level (also called the decision feature level). It is worth noting that some authors consider an intermediate fusion level for kernel methods in which the kernels are computed separately for each modality and then summed before the decision step [Noble (2004), Gönen and Alpaydin (2011)]. Multiple kernel learning (MKL) methods are well representatives of this intermediate fusion level [Bach *et al.* (2004), Sonnenburg *et al.* (2006)]. Fig. 7.1 extracted from [Noble (2004)], illustrates these three fusion levels for kernel methods. The same

taxonomy also holds for other learning approaches.

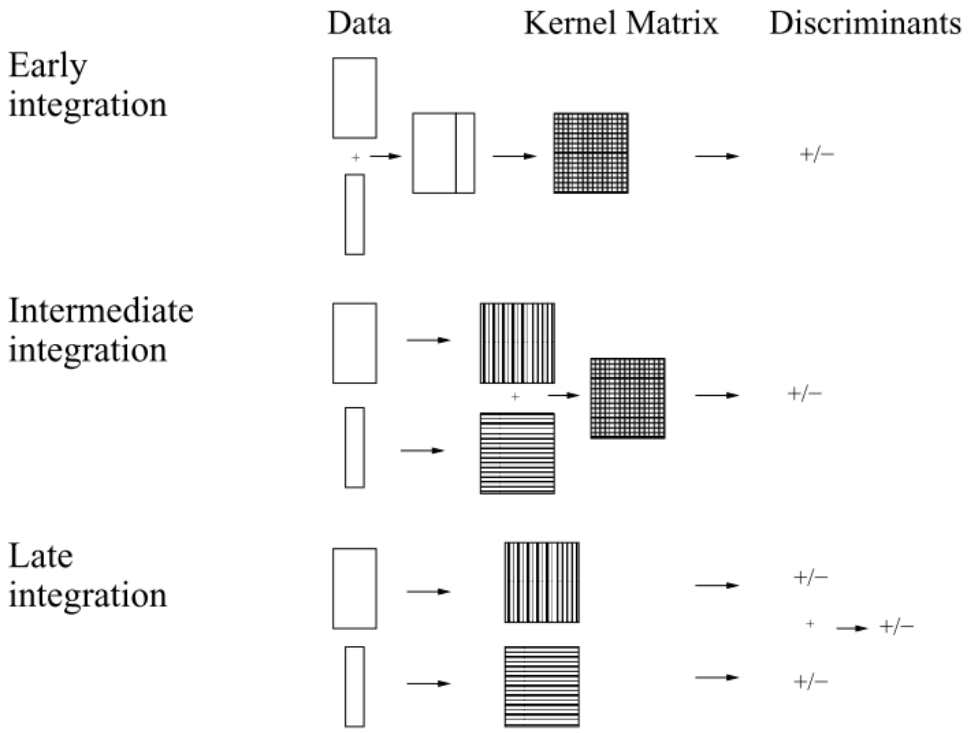


Figure 7.1: Example of the three fusion (integration) levels encountered in the context of kernel methods.

Early fusion As illustrated in the top of Fig. 7.1, this level of fusion corresponds to combining the features coming from different sources before the learning step. The most straightforward way of doing so is by concatenating theses features in a single feature vector. A global model is then learned using the series of this feature vectors corresponding to the available training observations. The final decision is in general obtained by thresholding directly the output of the learning algorithm or after a calibration step that is intended to increase the interpretability of the system output.

The early fusion level may seem as the optimal level. It should in theory allow the best decision as no information is lost in the fusion process. It also allows exploiting potential correlations that may exist between the different data sources and requires only one learning step. In practice, this fusion level raises various issues. First, from a learning point of view, we only have access to a limited number of observations to train the model and therefore, simply concatenating the features from each modality increases the dimensionality of the problem and makes it even more sensitive to the small n large p issue. In some cases, features from different sources can be very heterogeneous in their nature. In such cases, finding a global similarity measure or a unified representation can be difficult. For all these reasons, it may be more suitable in some cases to fuse the information from different sources at a higher level of abstraction.

Late fusion corresponds to combining the local decisions obtained from individual data sources. This fusion level is also sometimes referred to as data integration [Lahat *et al.* (2015)]. In Fig. 7.1, this fusion level is represented at the bottom. Features from individual modalities are used in parallel to learn single source models and their outputs combined to form the final system decision.

Late fusion presents many advantages over early fusion. An important advantage is that is allows using learning algorithms that are optimized for each single modality. It also avoids the problem of finding a common representation for all sources by combining only the decisions from each single model. Furthermore, in this fusion level, it is possible to include prior knowledge by using specific combination rules or weighting schemes.

7.2.2 Combination methods

These methods can be used in several ways to combine single features (at the early fusion level) or single model decisions (at the late fusion level). [Atrey *et al.* (2010)] distinguish between three families of combination methods (see Fig. 7.2).

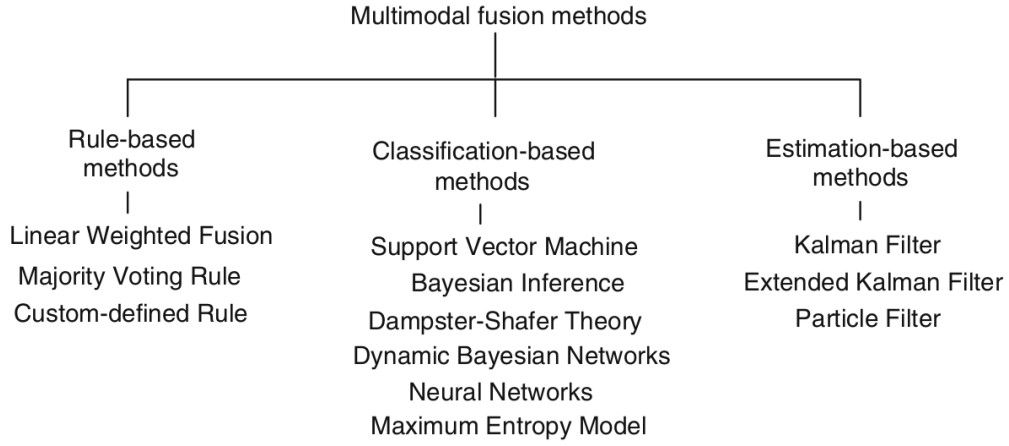


Figure 7.2: A categorization of combination methods.

Rule-based methods rely on the definition of a combination rule that is used to fuse the features or the outputs of single models. These include linear weighted fusion (sum and product), Min, MAX, AND, OR and majority voting (when all weights are set to be equal). It is also possible to use *a priori* knowledge to design custom combination rules. These methods in general are very simple to use and only require scaling of the features or the outputs of the different single models.

Classification-based methods corresponds to using the features or the outputs of single models to learn another classifier or detector. In supervised classification, the SVM algorithm is the most commonly used classifier. In general, introducing an additional classifier requires another step for optimizing the hyper-parameters of the classifier (*e.g.* the

C parameter for SVM, conditional probability priors for Bayesian inference) and avoiding complex models that do not generalize well.

Estimation-based methods correspond to filtering approaches that can be used to infer the value of a given attribute of interest. In general a linear dynamic (time dependence) system corrupted with Gaussian noise is assumed and the training observation are used to estimate the model parameters. The estimated model can then be used to predict given attributes of interest.

7.3 Application to epileptogenic lesion detection

To address the task of epilepsy lesion detection, we previously considered only patterns corresponding to healthy control subjects in the training phase (see Chap.5) and applied an outlier detection method to estimate the boundaries of this target class distribution. The method was applied on a voxel basis where each voxel from the volume of interest is independently associated with a one class support vector machine (OC-SVM) [Schölkopf *et al.* (2001)] classifier. The core idea of the OC-SVM is mapping the learning data $x_i,\ i \in [1, n]$ into a feature space via a feature map ϕ. The problem is then reformulated as a two class problem where the origin is considered as the only member of the outlier class and the learning data as representative of the target class. The algorithm tries then to find the optimal separating hyperplane $(\phi(x_i).w) - \rho$ that maximizes the separation margin between these two classes. This is done by solving the minimization problem below:

$$\begin{cases} \min\limits_{\mathbf{w},\rho,\boldsymbol{\xi}} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum\limits_{i=1}^{n}\xi_i - \rho \\ \text{subject to} & (\phi(\mathbf{x_i}).\mathbf{w}) \geq \rho - \xi_i, \qquad i \in [1, n] \\ \text{and} & \xi_i \geq 0, \qquad i \in [1, n] \end{cases} \tag{7.1}$$

ξ_i are slack variables which account for non-separable classes and ν is the upper limit for the fraction of permitted outliers. The decision variable is dependent on the signed distance separating x from the hyperplane. Giving a test voxel example, this signed distance relative to the hyperplane is positive if the example belongs to the target class and negative otherwise. We used the Matlab toolbox developed by Canu et al. [Canu *et al.* (2005)b] to solve the OC-SVM minimization problem. The OC-SVM parameter ν and the kernel parameter were optimized following the procedure described in Chap 5.

Applying these models to a test image produces an OC-SVM distance map. This distance map is thresholded at a certain level of significance to highlight the most suspicious areas in the image. The threshold is set by approximating the OS-SVM score distribution by a non-parametric distribution and finding the threshold value that corresponds to a given p-value. Clusters, connected voxels for the 26-connectivity rule, are then extracted from the thresholded distance map. The cluster map is transformed from the MNI space

to the patient native space by applying the inverse normalization transformation resulting from the segmentation preprocessing step.

In this section we extend this CAD system by investigating two approaches to multi-modality data fusion where multimodality data comes from three different MRI sequences, namely T1-w, FLAIR and DTI.

7.3.1 Data description

Study group The learning database is composed of 38 healthy control subjects aged 20 to 62 who had an MRI at the CERMEP imaging center. The proposed method was tested in three patients with medically intractable epilepsy (one lesion per patient) undergoing a pre-surgical evaluation at Lyon's Neurological Hospital. Abnormalities were detected by conventional visual inspection of the MRI in one patient (MRI+). The two other patients had a normal MRI (MRI-). The three patient all belong to the second patient database PDB2 introduced in Chap. 3.

MRI and SEEG protocol MR imaging was carried out on a 1.5T Sonata scanner (Siemens Healthcare, Erlangen, Germany). All subjects had a 3D anatomical T1-w sequence (TR/TE 2400/3.55; 160 sagittal slices of 192×192 1.2 mm cubic voxels) and a T2 FLAIR (176 sagittal slices of 196 × 256 milimetric cubic voxels). The DTI sequence consisted of 4 unweighted images and 48 diffusion weighted images. The reconstruction matrix was 96×96 and voxel size 2.5 mm. Figure 7.3 gives example slices of T1-w, FLAIR and DTI FA maps. SEEG electrodes were used in all three cases. The implantation strat-

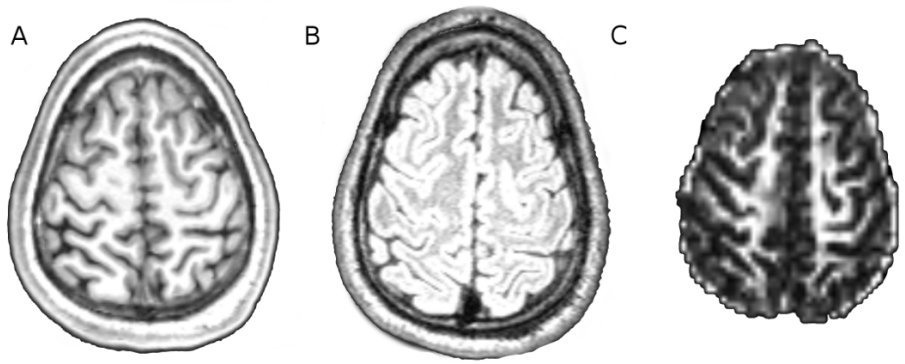


Figure 7.3: Example axial slices of (A) T1-w MRI (B) T2 FLAIR MRI (C) DTI-FA of patient 1 (compare figure 7.6).

egy was derived from the analysis of both clinical and imaging data, but not the results of the CAD system. After implantation, subjects underwent a 3D anatomical T1-weighted brain MRI allowing an accurate localization of the depth electrodes on the 3D MRI. The anatomical targeting of electrodes was established in each patient according to available non-invasive information and hypotheses about the localization of the epileptogenic zone.

7.3.2 Pre-processing

Spatial normalization To perform the voxelwise analyses, all subjects' T1-w MRIs were spatially normalized to the Montreal Neurological Institute (MNI) space using the 'Unified segmentation' algorithm within SPM8 [Ashburner and Friston (2005)]. This method also generates a transformation between the subject's native image and the normalized stereotactic MNI space. To use this transformation to normalize the individual T2 FLAIR and DTI parametric maps to the MNI space, these images were rigidly co-registered to the individual T1-w MRI and the registration parameters were applied to the feature maps which had been extracted in native space. Like in the preprocessing steps presented in Chap. 5, cortical and subcortical GM and WM regions were included, and cerebellum and brain stem excluded, to restrict the analysis to brain regions susceptible to harbour epilepsy lesions. The masking image in the reference space was derived from the Hammersmith maximum probability atlas [Hammers *et al.* (2003)].

Feature extraction For each subject of the learning database ($n = 38$), and for each patient of the test database ($m = 3$), we extracted a total of five features.

T1 features: we computed two parametric maps described in [Huppertz *et al.* (2005)] characterizing the GM extension, and GM/WM junction. These maps are derived from the individual tissue probability maps and from normative data computed on the healthy control dataset.

T2 FLAIR features: the individual FLAIR images were spatially normalized to the MNI space and intensity normalized by following the method described in [Focke *et al.* (2008)] using a manually drawn mask comprising the pons and the anterior frontal white matter.

DTI features: the DTI data were processed using FSL4.1 to correct for eddy currents and compute individual maps of fractional anisotropy (FA) and mean diffusivity (MD). These two parametric maps were then normalized to the MNI space. Feature scaling was performed by subtracting the mean parametric value of the controls from the parametric value of each individual, and dividing by the standard deviation of the parameters in the controls.

7.3.3 Multi-modal fusion

To find an optimal fusion strategy that efficiently exploits the information contained in the three MRI sequences, we propose to compared two fusion approaches with different fusion levels.

Early fusion strategy This approach corresponds to first combining the features extracted from all modalities before performing the classification task. In most cases, this is done by simply concatenating the different features to form a single feature vector.

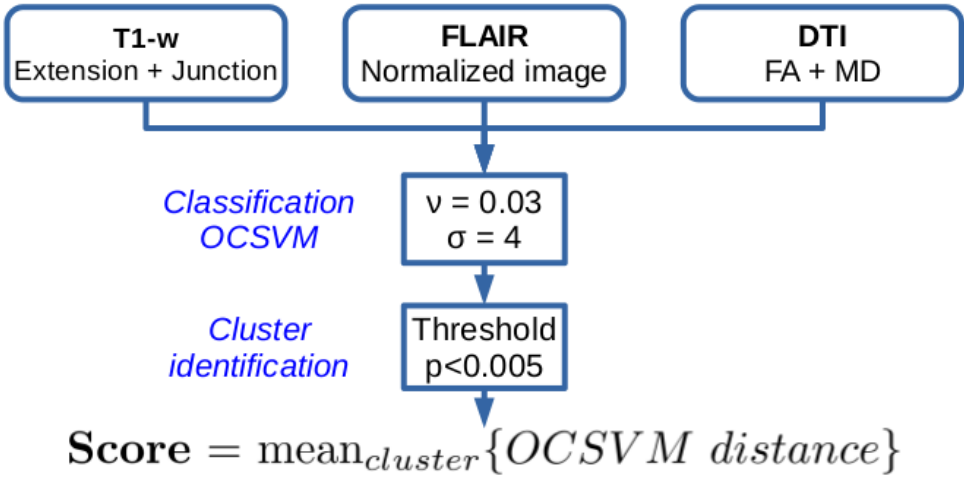


Figure 7.4: Overview of the early fusion approach.

In our case, we extracted five features from three modalities (T1, T2 FLAIR and DTI). For each voxel k, a single OC-SVM model (global model) is trained with the matrix $M^k \in M_{n,p}(\mathbb{R})$, $n = 38$ and $p = 5$ where each row of M^k is an instance of a five-component feature vector. As only one global model is derived we obtain a single cluster map. The clusters are ranked according to the mean value of the OC-SVM distances over the cluster.

Late fusion strategy This approach constructs different local models based on individual features and then combines the decisions associated with these models. Fig. 7.5 gives an overview of the proposed late fusion methodology.

We chose to define three local models associated each with one modality (*i.e.* one sequence). Depending on the considered local model (T1, T2 FLAIR or DTI), the OC-SVM classifier associated with a voxel k is trained using a matrix $M^k \in M_{n,p}(\mathbb{R})$, $n = 38$ and $p = 2$ for the T1-based model, $p = 1$ for the T2 FLAIR-based model and $p = 2$ for the DTI-based model. Three different cluster maps are thus derived. To produce a single labelled cluster map we proceed in two steps. First, we use the majority voting (MV) rule to combine the three maps and obtain a new cluster map referred to as 'MV map'. The scores assigned to each voxel by each of the three models are not normalized (signed distances), so that they can not be easily combined. To assign a label to each cluster of the MV map, in a second step, we perform a VBM analysis, considering each of the five features separately. A general linear model is applied considering two factors, the tested patient and the control group. The resulting five z-score maps are combined using Stouffer's method [Stouffer *et al.* (1949)], also called the z-transform, to produce a single z-score map. In this method, the final z-score is computed as a sum of the local z-scores, divided by the square root of the number of tests, k ($k = 5$). The mean z-score value of each cluster is then finally used to rank the detected clusters of the MV map.

The proposed combination strategy can be though of as a hybrid combination method between rule-based methods (in this case majority voting) and classification-based methods

(statistical z-score test). The second step that corresponds to performing a statistical test, that gives a z-score value, is aimed at avoiding the use of an arbitrary calibration method or making an assumption on the distribution of outliers to convert the output of single OC-SVM models into a posterior probability before combining their outputs.

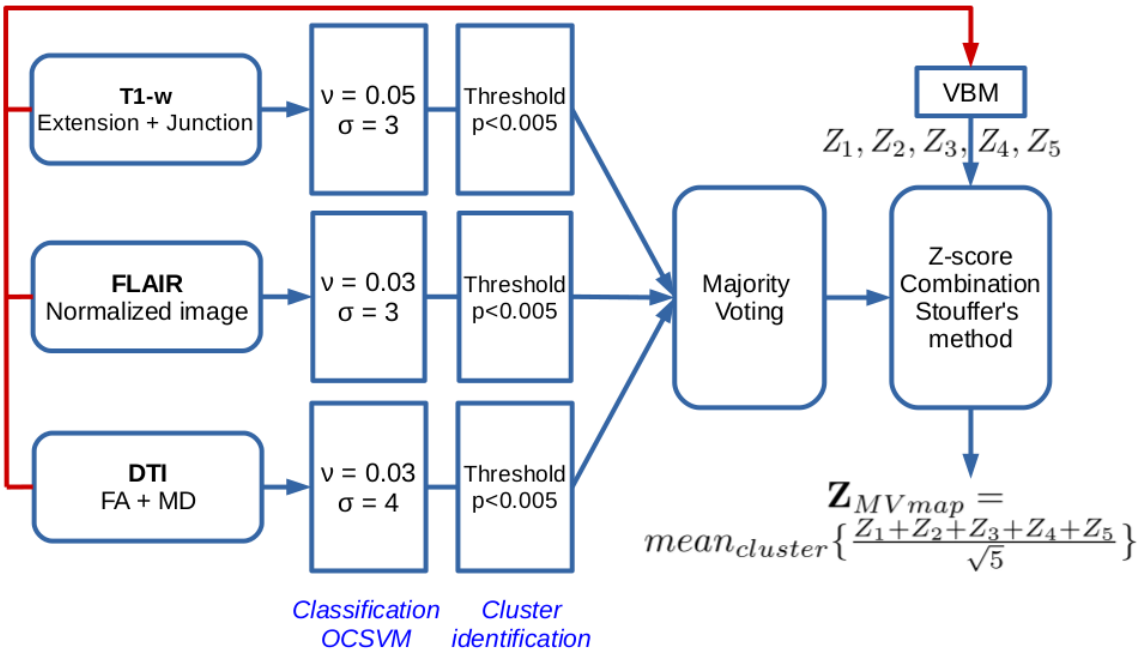


Figure 7.5: Overview of the late fusion approach.

7.3.4 Performance assessment: comparison against SEEG findings

For each model, the detected clusters were compared with the visually detected abnormalities and other data (e.g. FDG-PET, MEG). Clusters were designated by an expert as true positive or false-positive detections.

SEEG recordings for each patient were used to identify the brain regions generating seizures. Bartolomei et al. [Bartolomei *et al.* (2008)] described a method to compute an 'epileptogenicity index' (EI) that measures the epileptogenicity of the recorded brain structures. This index takes into account both the involvement of a given structure in the generation of rapid discharges at seizure onset and the delay in the appearance of discharges compared with the structures involved at seizure onset. For each patient and each SEEG contact, we computed the corresponding EI. These values were used to construct an image that can be superimposed on the patient's MRI after SEEG implantation. The patient's native T1-w MRI was rigidly co-registered to post-SEEG implantation MRI. The resulting transformation was applied to the different patient cluster maps. Cluster maps and EI map were both overlaid on post-SEEG implantation MRI to check for correlations. Based on this overlay, a cluster was designated as true positive (TP) if its location concurred with an activated SEEG electrode. Otherwise, the cluster was considered as a FP.

7.4 Results

7.4.1 Parameter optimization

A RBF kernel (with standard deviation σ) was used for all models to guarantee data separability from the origin in the feature space. Parameter optimization was performed for each model. Parameters (ν, σ) were optimized by randomly selecting 4000 voxels from the whole brain volume and performing a leave-one-out analysis on the grid $2^{[-1:1:4]} \times 10^{[-2:0.25:1]}$. The optimal parameters, in term of accuracy, for the T1, FLAIR, DTI and the global model were respectively $(\nu = 0.05, \sigma = 3)$, $(\nu = 0.03, \sigma = 3)$, $(\nu = 0.03, \sigma = 4)$ and $(\nu = 0.03, \sigma = 4)$. See also Fig. 7.5 and Fig. 7.4.

7.4.2 Clinical data results

Table 7.1 summarises the results obtained for the test group when tested with the different models.

Local model comparison Columns 2 to 7 in table 7.1 show that: 1) the T1 based model has a higher sensitivity than the T2 FLAIR based model (3/3 detections vs 2/3 detections). 2) the DTI based model failed to detect the EL in all three patients. 3) the T2 FLAIR based model has a better specificity (3 FPs on average) compared with the T1 based model (7 FPs on average).

	T1		**Flair**		**DTI**		**feature fusion**		**score fusion**		**VBM fusion**	
	count	**EL rank**	**count**	**EL rank**	**count**	**EL rank**	**count**	**EL rank**	**count**	**EL rank**	**count**	**EL rank**
patient 1 positive MRI	8	1	4	4	7	-	6	3	1	1	2*	2
patient 2 negative MRI	6	1	2	1	3	-	3	-	3	1	6*	-
patient 3 negative MRI	7	6	1	-	5	-	5	-	1	-	3*	-

*Table 7.1: Classification results for the test group with the different models. Count corresponds to the number of detected clusters and the rank is the EL rank assigned by the model. * very small lesion (size inferior to 5 voxels).*

Fusion strategies comparison Columns 8 to 11 in table 7.1 compare the two fusion strategies. Columns 12 and 13 show the results obtained by combining directly the local VBM z-score maps after thresholding at the same p-value of 0.005 and applying a majority voting rule. For patient 1, both fusion approaches succeeded in identifying the EL with a better ranking in the score fusion case. The score fusion approach successfully removed all FPs whereas the feature fusion approach produced 5 FPs. The fusion approaches outperform the three local models in terms of both specificity and sensitivity. For patient 2,

only the score fusion based approach succeeded in identifying the lesion while reducing considerably the number of FPs (2 FPs instead of 5 FPs for the T1 based model). For patient 3, both approaches failed to detect the lesion. The score fusion approach gives the best specificity compared with the feature fusion approach (1 FP vs 5 FPs). The best model for this patient is the T1 based model with the lesion ranked 6 out of 7 detected clusters. For all patients, the VBM based fusion lacks both sensitivity and specificity. Fig. 7.6 and Fig. 7.7 give example slices showing a good correlation between the activated SEEG contacts and the MV map obtained for patient 1 and 2.

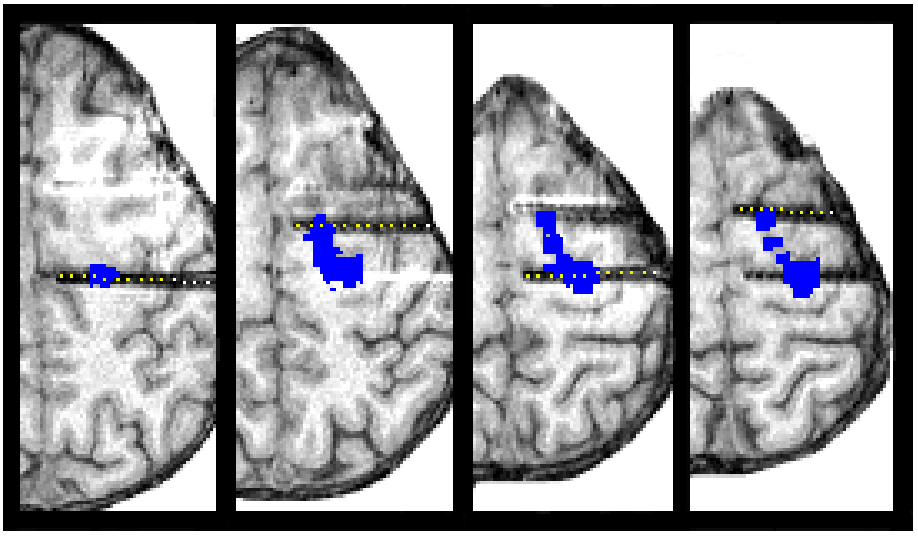


Figure 7.6: Patient 1 post-SEEG implantation MRI axial slices overlaid with the EI map (yellow dots) and the MV map (blue). The IE map was thresholded to highlight the most activated SEEG electrodes.

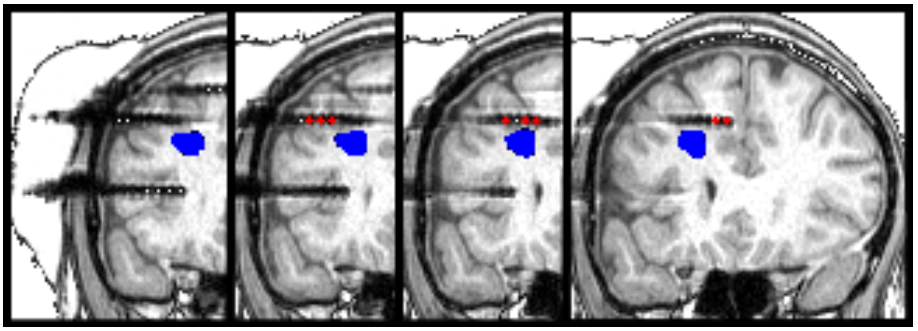


Figure 7.7: Patient 2 post-SEEG implantation MRI coronal slices overlaid with the EI map (red dots) and the MV map (blue). The IE map was thresholded to highlight the most activated SEEG electrodes.

7.5 Conclusion

In this chapter, we proposed an extension of our CAD system to deal with the multi-modal nature of the neuroimaging data used in the diagnosis of epilepsy. Our CAD system

originally based only on features extracted from T1-w MR images, was adapted to the multimodality scenario by integrating two additional MR sequences (FLAIR and DTI).

We investigated two fusion strategies with different fusion levels. The early fusion strategy consisted in concatenating features extracted from all three sequences in a single feature vector before learning a global model of normality. The late fusion approach consisted in learning three local models where each model is associated with a single MR sequence. We also proposed a hybrid approach for decision combination based on majority voting and Stouffer's method coupled with a univariate statistical test.

Detection results were validated against the reference SEEG exam. Our results show that multi-modal fusion can improve the CAD system specificity. The best performances were obtained with our novel late fusion approach based on majority voting and VBM z-score combination using Stouffer's method. For the test group, no correlation was found between the clusters found by the DTI based model and the SEEG epileptogenic zone. A previous study [Thivard *et al.* (2006)] demonstrated that DTI anomalies correlates more with the SEEG irritative or spreading zone than with the SEEG onset zone. One interesting perspective of this work could be to introduce a confidence score for each base classifier that can be used to weight its decision in an adaptive Stouffer z-score combination framework. This would allow incorporating *a priori* knowledge about the discriminative power of each model. For instance, giving higher weight for the T1-w model in comparison with the DTI model.

In this study, we only considered the late and early fusion levels. Another possibility would be to investigate the use of an intermediate fusion level method such as multiple kernel learning (MKL) methods. The applicability of MKL in the context of outlier detection has not been largely investigated due to difficulty of finding a principled methodology for hyper-parameter optimization. Recently, [Gaëlle, Loosli and Aboubacar (2014)] proposed an adaptation of the SVDD method to multiple kernel learning, based on SimpleMKL [Rakotomamonjy *et al.* (2007)] algorithm. The authors also proposed to modify the objective function to favour tight boundary decisions by maximizing the number of support vectors. Hyper-parameter optimization is however performed using cross validation and optimizing the AUC criterion. As discussed in Chap. 4 such performance measures cannot be used in an outlier detection scenario where no examples of the outlier class are available for model training and selection like it is the case in our application.

Chapter 8

Probabilistic outputs for outlier detection

8.1 Motivation

In Chap. 2 we gave a description of outlier detection. It mainly consists in identifying observations in the data that appear to not conform to the expected behaviour. These observations are designated as anomalies or outliers. In the recent years, the challenging topic of outlier detection has been extensively studied and many algorithms have been proposed for outlier detection depending on the nature of the data, the labels and the type of anomalies [Chandola *et al.* (2009), A.F. Pimentel *et al.* (2014)]. Most reviews on the subject classify these algorithms according to five major categories: (i) probabilistic or statistical, (ii) distance-based, (iii) reconstruction based, (iv) domain or classification based, and (v) information theoretic. The choice of a specific method depends greatly on the nature of the data, the available labels and the type of anomalies. For instance, statistical are reconstruction based approaches are best suited when a large number of training data is available. Information theoretic methods operate in an unsupervised setting, still the presence of a significant number of outlier examples in the training data is needed.

In this chapter we will focus on the output of these methods. Typically, this output is either a label or a numeric score representing the degree of outlierness of a given observation. Scores outputted by these approaches, with the exception of statistical methods, are in general very hard to interpret and to compare as they vary widely from one dataset to another and also between methods. Having meaningful outputs is also crucial for output interpretation, comparison and combination. For instance, in the context of medical image processing, the experts' decision usually depends on a multitude of tests

whose outputs should be combined to make the final decision. If these outputs correspond to well-calibrated probabilities, they can be much more easily used to compute the associated p-values and design a test that allows for a better control of the false discovery rate or that maximizes the detection rate. However, in many application domains such as pattern recognition, fraud, intrusion detection, and wireless sensors networks, the training datasets are often of small size, potentially noisy, and highly imbalanced and not enough labelled examples of outliers are available to construct a robust outlier detector. Distance and domain-based outlier detection methods are well suited for this type of datasets. They however fail at producing meaningful outputs. In Chap. 7, we proposed a framework for combining base OC-SVM classifiers for fusing data extracted from three MRI sequences. Classifiers' outputs were combined using majority voting. By using majority voting we avoided the need for meaningful outputs as it only requires having binary outputs. Nevertheless, for the CAD system interpretability, we proposed performing univariate statistical tests per feature to compute statistical z-scores that were further combined using Stouffer's combination rule.

Considering these key challenges (small dataset and the need for meaningful outputs), we propose an outlier detection framework that combines both a domain-based approach and a statistical method. To guarantee a good performance when only few training observations are available, we build on the support vector data description (SVDD) algorithm and propose a generalization of this algorithm. To get meaningful outputs, we convert the outputs of this generalized version of the SVDD algorithm into probability estimates by using a statistical framework.

Statistical approaches consist in learning the compact representation domain of the normal behaviour, in view of predicting whether or not a test observation belongs to this compact description. Defining the "expected behaviour" is however very difficult as the probability distribution of the normal data (N) is usually unknown or too complex to be modelled. More formally, let $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$ be a dataset consisting of n i.i.d observations drawn from the unknown probability distribution of the normal class N. An observation $\mathbf{x}$ is an outlier if the posterior probability of $\mathbf{x}$ to belong to the normal class is inferior to a given threshold η. The decision rule is then:

$$\begin{cases} \mathbb{P}(N|\mathbf{x}) < \eta, & \mathbf{x} \text{ is an outlier} \\ \\ \mathbb{P}(N|\mathbf{x}) > \eta, & \mathbf{x} \text{ is normal.} \end{cases}$$

Estimating the whole probability distribution from a finite number of n observations is usually very hard and even harder in the case of few and noisy observations or in high dimensional feature spaces. One way of going around this is by rewriting the decision rule as $-log(\mathbb{P}(\mathbb{N}|\mathbf{x})) - \eta' > 0$ or < 0, and then estimating a function $f : X \to \mathbb{R}$ instead of $\mathbb{P}$ such as: $sign(f(\mathbf{x})) = sign(-log(\mathbb{P}(\mathbb{N}|\mathbf{x})) - \eta')$ for a given threshold η'. This function can actually be thought of as defining a one dimensional projection of the data that coincides with the density level-set η' and the initial problem is reduced to estimating a single density

level-set. However, by using this function f, the obtained outlier scores can no longer be readily interpretable in terms of probability estimates.

Various attempts have been made to convert outlier scores into calibrated probabilities [Gao and Tan (2006), Nguyen *et al.* (2010), Kriegel *et al.* (2011)]. All these methods however make some assumptions on the distribution of the outlier scores. For instance, in [Gao and Tan (2006)], outlier scores are assumed to share the same score distribution with the normal samples or to be uniformly distributed. Depending on the assumed assumption the outlier scores are converted into posterior probabilities using either a sigmoid calibration function or a mixture of exponential and Gaussian distributions. The experiments showed an improved detection performance in the context of ensemble learning in comparison with single detectors, however the authors only tested for improvement when combining multiple kNN base detectors with different k values and not when combining different base detectors. In [Nguyen *et al.* (2010)], the main focus is on combining outlier scores obtained by different methods to improve the detection rate rather than focusing on deriving good estimates of the posterior probabilities. The authors therefore proposed different functions (weighted sum or weighted majority voting) to combine normalized outlier scores depending on the nature of the scores outputted by each base outlier detector. In [Kriegel *et al.* (2011)], a two-steps approach is adopted. First outlier scores are normalized and then scaled using a statistical model for the outlier score distribution by assuming either a Gaussian or a Gamma distribution. The authors however do not give a general rule as to which distribution is the most suited as this depends not only on the algorithm used to compute outlier scores (e.g. LOF [Breunig *et al.* (2000)] or kNN [Jin *et al.* (2001)]) but also on the considered datasets.

One-class classification methods, such as one-class support vector machine (OCSVM) [Schölkopf *et al.* (2001)] and its variants, have also been successfully applied in the context of outlier detection [Chandola *et al.* (2009)]. Such methods learn a boundary function f that encloses most of the training data from the normal class. In particular, support vector data description method (SVDD) [Tax and Duin (2004)] is a very similar approach to OCSVM, that searches for the enclosing hypersphere with minimum volume containing at least a fraction η' of the training data. A very interesting theoretical property of these methods was shown in [Vert and Vert (2006)] where the authors proved that asymptotically, the decision boundary computed by OCSVM converges to the minimum volume set (MV-set) with probability mass of at least $\eta' = 1 - \nu$ for a well-chosen Gaussian kernel width. These methods however allow only the computation of a single MV-set and the resulting scores cannot be easily interpretable in terms of posterior probabilities.

In the present study, our goal is to design an outlier detection method that:
1) allows estimating q MV-sets of given probability masses $\eta_j,\ j = 1 \dots q$,
2) converts the outputted outlier scores into probability estimates,
3) maximizes the detection rate.
To this end we proceed in two steps. First, we propose to build on the SVDD algorithm and reformulate the associated minimization problem to obtain q concentric hyper-spheres as-

sociated each with a probability mass η_j. Second, we propose two calibration functions for converting the obtained scores into posterior probabilities. Like in [Gao and Tan (2006)], the first calibration function is obtained by modelling outlier score distribution as a sigmoid function and taking advantage of the properties of the q concentric SVDD models as hierarchical MV-sets to derive maximum likelihood estimates of the calibration function parameters. The second calibration function is a Weibull distribution that is known to provide a better estimate of the tail of a given distribution.

To our knowledge, no method has been specifically developed to convert scores outputted either by OCSVM or SVDD into well calibrated posterior probabilities. The only approach that relates to our present study was recently proposed by Glazer et al. in 2013 [Glazer *et al.* (2013)] for the estimation of q hierarchical level-sets (q-OCSVM) by constructing q parallel separating hyperplanes. The focus in this study was quantile estimation for high dimensional distributions and no attempt was made to translate this work in the context of outlier detection or to convert q-OCSVM outlier scores into posterior probabilities. The study did show however the superiority of the q-OCSVM approach in comparison with other extensions of the OCSVM algorithm, namely the hierarchical minimum-volume estimator [Glazer *et al.* (2012)] and the nested one-class SVM [Lee and Scott (2010)].

8.2 SVDD generalization

In the following, we first introduce two generalizations of the SVDD algorithm for the estimation of q density level-sets.

8.2.1 Naive approach (iSVDD)

We first propose a straightforward generalization of the SVDD algorithm for the estimation of q MV-sets by constructing q independent SVDD models, each associated with a different probability mass $1-\nu_j$. For each $j = 1 \dots q$, the SVDD algorithm tries to find the enclosing hyper-sphere, of centre $\mathbf{a}_j$ and radius R_j, with minimum volume given ν_j. Let $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$ be a training dataset consisting of n i.i.d samples drawn from a normal class (N) with unknown probability distribution $\mathbb{P}$. The global ν-formulation for all q SVDD models is:

$$\begin{cases} \min\limits_{R_j, \mathbf{a_j}, \boldsymbol{\xi_j}} & \sum\limits_{j=1}^{q} R_j^2 + \sum\limits_{j=1}^{q} \dfrac{1}{\nu_j n} \sum\limits_{i=1}^{n} \xi_{ji} \\ \text{s.t} & (\mathbf{x}_i - \mathbf{a}_j)^\top (\mathbf{x}_i - \mathbf{a}_j) \leq R_j^2 + \xi_{ji}, \qquad i = 1, \dots, n \ , \ j = 1, \dots, q \\ \text{and} & \xi_{ji} \geq 0, \qquad\qquad\qquad\qquad\quad i = 1, \dots, n \ , \ j = 1, \dots, q, \end{cases} \tag{8.1}$$

where ξ_{ji} are slack variables that allow relaxing the inequality constraints and $\nu_j \in [0, 1]$ corresponds to an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors [Schölkopf *et al.* (2001)].

The outlier score associated to a given observation $\mathbf{x}$ by the jth independent SVDD model is then given by: $g_j(\mathbf{x}) = (\mathbf{x} - \mathbf{a}_j)^\top(\mathbf{x} - \mathbf{a}_j) - R_j^2$.

In [Vert and Vert (2006)], the consistency property of the SVDD algorithm was proven for estimating density level sets. An estimate of the level-set of probability mass at least $1 - \nu_j$ is given by:

$$MV_{1-\nu_j} = \{\mathbf{x} \mid g_j(\mathbf{x}) = 0\}.$$

It should be noted here that just thresholding the decision function of a single SVDD model at another offset, different from 0, does not necessarily give an MV-set. Therefore, to obtain q MV-sets, the SVDD optimization problem in equation 8.1 has to be solved for q different values of ν_j. The q SVDD models are derived independently from one another, thus not much can be said about the relation between these models. In particular, the q hyper-spheres do not necessarily share the same centre and no order relationship can be guaranteed between the different radii. To define a global outlier scores in this case, we first pick a reference SVDD model g_0, with centre $\mathbf{a}_0$ and radius R_0, to which the distance of a test sample is computed. The choice of a reference SVDD model among the q independent SVDD models is considered as an additional user-defined parameter of the method. In the remainder of this chapter, we will refer to this method as iSVDD.

8.2.2 Concentric SVDD models (cSVDD)

We extend the SVDD algorithm by constructing q hierarchical MV-sets with decreasing probability masses. This translates into having the same centre $\mathbf{a}$ for all q SVDD models and solving a single optimization problem.

Primal formulation Let R_j be the radius associated with the jth SVDD model, the primal problem formulation is:

$$\begin{cases} \min\limits_{R_j, \mathbf{a}, \boldsymbol{\xi_j}} & \sum\limits_{j=1}^{q} R_j^2 + \sum\limits_{j=1}^{q} \dfrac{1}{\nu_j n} \sum\limits_{i=1}^{n} \xi_{ji} \\ \text{s.t} & (\mathbf{x}_i - \mathbf{a})^\top(\mathbf{x}_i - \mathbf{a}) \leq R_j^2 + \xi_{ji}, \qquad i = 1, \ldots, n \ , \ j = 1, \ldots, q \\ \text{and} & \xi_{ji} \geq 0, \qquad\qquad\qquad\qquad\quad i = 1, \ldots, n \ , \ j = 1, \ldots, q. \end{cases} \tag{8.2}$$

Dual formulation Let $\alpha_{ji} \geq 0$ and $\beta_{ji} \geq 0$ be the Lagrange multipliers associated with the inequality constraints $(\mathbf{x}_i - \mathbf{a})^\top(\mathbf{x}_i - \mathbf{a}) \leq R_j^2 + \xi_{ji}$ and $\xi_{ji} \geq 0$ respectively.

The Lagrangian of problem 8.2 is:

$$\begin{aligned} L(\mathbf{a}, R_j, \boldsymbol{\xi_j}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \sum_{j=1}^{q} R_j^2 + \sum_{j=1}^{q} \frac{1}{\nu_j n} \sum_{i=1}^{n} \xi_{ji} \\ & - \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} [R_j^2 + \xi_{ji} - (\mathbf{x}_i - \mathbf{a})^\top (\mathbf{x}_i - \mathbf{a})] \\ & - \sum_{j=1}^{q} \sum_{i=1}^{n} \beta_{ji} \xi_{ji}. \end{aligned}$$

The derivatives of the Lagrangian with respect to the primal variables are:

- $\nabla_{\mathbf{a}} L(\mathbf{a}, R_j, \boldsymbol{\xi_j}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -2 \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} (\mathbf{x}_i - \mathbf{a})$

- $\nabla_{R_j} L(\mathbf{a}, R_j, \boldsymbol{\xi_j}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 2(1 - \sum_{i=1}^{n} \alpha_{ji})$.

- $\nabla_{\xi_j} L(\mathbf{a}, R_j, \boldsymbol{\xi_j}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{\nu_j n} \mathbf{e} - \boldsymbol{\alpha}_j - \boldsymbol{\beta}_j$,
  where $\mathbf{e} = [\underbrace{1 \ldots 1}_{\text{size n}}]^T$, $\boldsymbol{\alpha}_j = [\underbrace{\alpha_{j1} \ldots \alpha_{jn}}_{\text{size n}}]^T$ and $\boldsymbol{\beta}_j = [\underbrace{\beta_{j1} \ldots \beta_{jn}}_{\text{size n}}]^T$.

Satisfying the stationarity conditions of Karush-Kuhn-Tucker at the optimum gives:

- $\sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji}^* \mathbf{x}_i = \sum_{j=1}^{q} \underbrace{\sum_{i=1}^{n} \alpha_{ji}^*}_{=1} \mathbf{a} = q\mathbf{a} \ \leftarrow$ *the representer theorem.*

- $\sum_{i=1}^{n} \alpha_{ji}^* = 1, \ j = 1, \ldots, q.$

- $\frac{1}{\nu_j n} \mathbf{e} - \boldsymbol{\alpha}_j^* - \boldsymbol{\beta}_j^* = 0, \ j = 1, \ldots, q.$

Substituting the primal variables in the Lagrangian gives:

$$\begin{aligned} L(\mathbf{a}, R_j, \boldsymbol{\xi_j}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{j=1}^{q} R_j^2 + \sum_{j=1}^{q} \frac{1}{\nu_j n} \sum_{i=1}^{n} \xi_{ji} - \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} [R_j^2 + \xi_{ji}] \\ &\quad - \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} (\mathbf{x}_i - \frac{1}{q} \sum_{k=1}^{q} \sum_{l=1}^{n} \alpha_{kl} \mathbf{x}_l)^\top (\mathbf{x}_i - \frac{1}{q} \sum_{k=1}^{q} \sum_{l=1}^{n} \alpha_{kl} \mathbf{x}_l) \\ &\quad - \sum_{j=1}^{q} \sum_{i=1}^{n} (\frac{1}{\nu_j n} - \alpha_{ji}) \xi_{ji} \\ &= \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} \mathbf{x}_i^T \mathbf{x}_i - \frac{1}{q} \sum_{j,k=1}^{q} \sum_{i,l=1}^{n} \alpha_{ji} \alpha_{kl} \mathbf{x}_i^T \mathbf{x}_l. \end{aligned}$$

The dual formulation of problem 8.2 is then:

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \sum\limits_{j=1}^{q}\sum\limits_{i=1}^{n}\alpha_{ji}\mathbf{x}_i^T\mathbf{x}_i - \dfrac{1}{q}\sum\limits_{j,k=1}^{q}\sum\limits_{i,l=1}^{n}\alpha_{ji}\alpha_{kl}\mathbf{x}_i^T\mathbf{x}_l & \\ \text{s.t} & \sum\limits_{i=1}^{n}\alpha_{ji} = 1 & j = 1,\ldots,q \\ \text{and} & 0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n} & i = 1,\ldots,n\ ,\ j = 1,\ldots,q. \end{cases} \tag{8.3}$$

The outlier score assigned to a given observation $\mathbf{x}$ by the j^{th} SVDD model is then given by : $f_j(\mathbf{x}) = f_c(\mathbf{x}) - R_j^2$ where $f_c(\mathbf{x}) = (\mathbf{x}-\mathbf{a})^\top(\mathbf{x}-\mathbf{a})$. The center $\mathbf{a}$ is given by the representer theorem above. The different radii $R_j,\ j \in [1\ldots q]$ can be obtained either by considering the distance to the center $\mathbf{a}$ of the essential support vectors for the j^{th} model, or by deriving the bidual formulation of the dual problem given in equation 8.3 and identifying the radius. Note that this dual formulation is very similar to the dual formulation of the standard SVDD and that the kernelization of this algorithm is straight forward by using the kernel trick and associated representer theorem.

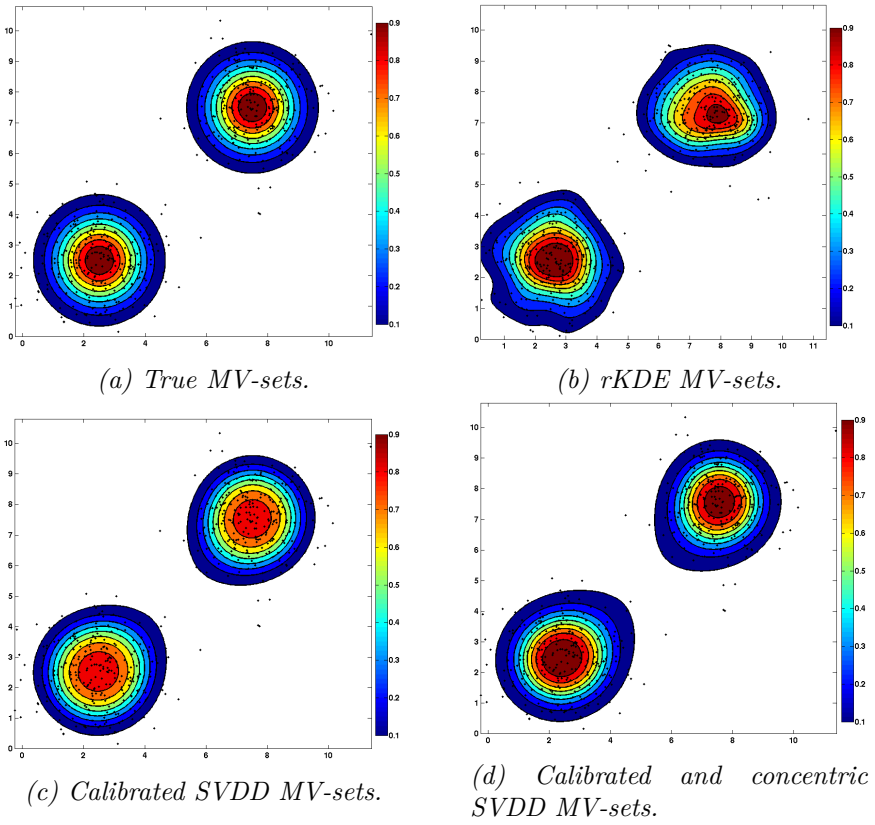
The nested property of cSVDD Let $f_1, f_2, \ldots, f_q$ be the q decision functions estimated using cSVDD, and let $R_1, R_2, \ldots, R_q$ be the associated radii. The resulting MV-sets are nested hyper-spheres. As the hyper-spheres share the same centre a by construction, to prove the nested property, we only need to prove that under some assumptions, the following order relation holds $R_q < \cdots < R_1$ for $\nu_q > \cdots > \nu_1$.

8.2.3 Method comparison

Fig. 8.1 illustrates the estimated MV-sets for a bimodal Gaussian distribution. The true MV-sets 8.1a were computed using Monte Carlo simulation. The MV-sets estimated by the two proposed generalization of the SVDD algorithm (iSVDD) and (cSVDD) are given in Fig. 8.1c and 8.1d. For comparison, we also show in Fig. 8.1b the MV-sets estimated using the robust kernel density estimator (rKDE), a non parametric statistical approach that was proposed by [Kim and D. Scott (2012)]. For a more detailed description of this algorithm see Sec. 8.4.1. For iSVDD, cSVDD and rKDE the Gaussian kernel width was optimized to obtain the best estimation of the true MV-sets. Compared with rKDE, iSVDD and cSVDD both captured better the nested and symmetrical nature of the MV-sets associated with the bimodal Gaussian distribution. cSVDD allowed a better estimate than iSVDD of the MV-set with the highest probability mass ($\eta = 0.9$).

8.3 Score conversion into probabilities

Converting outliers scores into well calibrated probabilities is a challenging task. Standard calibration methods such as in [Franc *et al.* (2011)] and included references can-



(a) True MV-sets. *(b) rKDE MV-sets.*

(c) Calibrated SVDD MV-sets. *(d) Calibrated and concentric SVDD MV-sets.*

Figure 8.1: Minimum volume sets obtained for a bimodal Gaussian distribution.

not be used as no labelled example is available for the outlier class. We propose two methods to model the distribution of outlier scores, by using either a sigmoid calibration function inspired by [Platt *et al.* (1999)] or by fitting a generalized extreme value distribution [Pickands III (1975)].

8.3.1 Calibration using sigmoid function

Let us consider n observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn from the normal class (N), and let f_i, $i \in [1, n]$ be outlier scores assigned to $\mathbf{x}_i$ by either cSVDD ($f_i = f_c(\mathbf{x_i})$) or iSVDD ($f_i = g_0(\mathbf{x}_i)$), with f_c and g_0 as defined in Sec. 8.2. We assume that the higher f_i is the more likely $\mathbf{x}_i$ is an outlier. We want to estimate the posterior probability $p(\mathbf{x}_i) = \mathbb{P}(O|f_i)$ that observation $\mathbf{x_i}$ is an outlier given its outlier score f_i.
Following [Platt *et al.* (1999), Gao and Tan (2006)], we propose to model this probability as a decreasing sigmoid function of f_i of the form:

$$p(\mathbf{x}_i) = \frac{A}{1 + exp(Bf_i + C)},$$

where A,B and C are the 3 model parameters that have to be estimated

In [Platt *et al.* (1999)], labelled examples from the outlier class are available and therefore estimates of (A, B and C) are derived using the expectation-maximization algorithm

(EM) via maximizing the likelihood of the observed labels. In [Gao and Tan (2006)], the authors proposed to adapt this method in the absence of labelled outlier samples. The labels are considered as hidden variables in the EM algorithm and are replaced in the first step of the EM algorithm by their expected value. The maximization step consists then on finding the parameters that minimize the negative log likelihood given the current fixed labels. The authors, showed improved detection performance especially for kNN ensemble learning, however the model fitting approach based on the EM algorithm was reported to be rather unstable [Kriegel *et al.* (2011)].
We propose to solve a much easier problem by taking advantage of the consistency property of each SVDD model to learn the parameters of the sigmoid model. The decision boundary of each SVDD model j is an estimate of the MV-set associated with a probability mass η_j that can be approximated by $1 - \nu_j$ [Vert and Vert (2006)]. This means that ν_j is an estimate of the probability $p(\mathbf{x}_i)$ of any observation $\mathbf{x}_i$ lying on the decision boundary of the j^{th} SVDD model (i.e. the essential support vectors). Let us denote by $\tilde{p}_j$ the estimate of the posterior probability associated with the essential support vectors of the j^{th} SVDD model whose score is f_{SVj}.

For cSVDD, we directly construct the training set $\{(\tilde{p}_j, f_{SVj}),\ j = 1 \dots q\}$ by associating the cSVDD score f_{SVj} for all support vectors lying on the decision boundary of the j^{th} SVDD model to the corresponding $\tilde{p}_j$ value.
For iSVDD, after selecting a reference SVDD model g_0, we can compute the distance to this model of all other SVDD models by considering the distance of their essential support vectors to the reference model g_0. Since the q independent SVDD models are not necessarily concentric, the value f_{SVk} will not be the same for all support vectors of model $g_k,\ k \neq 0$. We therefore, consider the mean score f_k averaged over all f_{SVk} to form a learning set of q coupled values $(\tilde{p}_k, f_k)$, where $\tilde{p}_k$ is given by ν_k.
For both cSVDD, and iSVDD, the proposed calibration approach is aimed at finding the sigmoid parameters (A, B and C) that maximize the likelihood of the observed empirical probabilities $\tilde{p}_j$ associated with the q SVDD models.

8.3.2 Calibration using extreme value distributions

In the context of outlier detection, our initial assumption was that, outliers correspond to rare observations with a very low probability and are located in the tail of the distribution. Therefore, to obtain good estimates of the probability of being an outlier, the estimator must allow a good estimation of the tail of the unknown distribution $\mathbb{P}$. Extreme value theory is a branch of statistics that deals with extreme deviations of a probability distribution [Pickands III (1975)]. The extreme value probability (EVP) of a random variable z is the probability of z being the largest value of the dataset. A key theorem in extreme value statistics, namely Fisher and Tippett theorem [Fisher and Tippett (1928)], states that the EVP can be expressed as the generalized extreme value (GEV) distribution.

The cumulative distribution function of a GEV distribution is given by:

$$F(z) = exp(-[1 + \zeta(\frac{z-\mu}{\sigma})]^{-\frac{1}{\zeta}}),$$

where the parameter ζ is the shape parameter that controls the tail behaviour of the distribution.
In our case, we are interested in estimating $p(\mathbf{x}_i)$ the probability of observation $\mathbf{x}_i$ of being an outlier given its outlier score f_i. To use EVP, we consider the univariate distribution of scores $z = f_i$ over the entire training dataset to fit the GEV distribution F. The resulting EVP value can then be used as a probabilistic measure of outlierness in the way that an observation is detected as an outlier if its EVP value exceeds a certain threshold.

8.4 Evaluation and parameter selection

We evaluate the performance of the proposed approaches iSVDD and cSVDD on both synthetic data and real datasets from the UCI repository [Lichman (2013)] . The performance of the proposed approaches are compared against that of a robust kernel density estimator (rKDE) [Kim and D. Scott (2012)], a non-parametric statistical approach.

8.4.1 Robust Kernel Density Estimator (RKDE)

Parametric and non parametric methods can be used to estimate an unknown probability density function from training data $\mathbf{x}_i$. Parametric methods, assume that the density function follows a particular model with parameters that can be estimated from the data using for instance the maximum likelihood approach. Non parametric methods, make no assumption on the underlying distribution and attempt to adapt the model to fit the training data whose size has to be large enough for a reliable estimation of all free parameters. These methods often involve some kind of smoothing and the choice of the smoothing bandwidth is always crucial.

One of the most popular non parametric density estimator is the kernel density estimator (KDE) [Silverman (1986)]. Recently in 2012, [Kim and D. Scott (2012)] proposed a generalization of the KDE method that is more robust to the presence of noise or contaminated training samples. This method uses the fact that the kernel density estimator can be interpreted as a sample mean in the reproducing kernel Hilbert space (RKHS) associated with the kernel. Therefore, the authors proposed to compute a robust estimation of the sample mean by using loss functions that are more robust than the original quadratic loss function used in KDE. The considered loss functions were: Hampel's [Hampel (1974)] and Huber's [Huber *et al.* (1964)] loss. The authors also characterized their method by a representer theorem by demonstrating that the resulting estimator, $\hat{f}_{RKDE}$, can be expressed as a weighted combination of kernels centred at the data points $\mathbf{x}_i$.

$$\hat{f}_{RKDE} = \arg\min_{\mathbf{g}\in\mathcal{H}} \sum_{i=1}^{n} \boldsymbol{\rho}(\|\Phi(\mathbf{x}_i) - g\|_{\mathcal{H}}) = \sum_{i=1}^{n} w_i K(\cdot, \mathbf{x}_i),$$

where $\mathcal{H}$ is an Hilbert space, $\Phi(\mathbf{x}) = K(\cdot, \mathbf{x})$ an element of $\mathcal{H}$, $\boldsymbol{\rho}$ is the robust loss function and w_i are positive weights that sum to one. The different weights are then computed by solving the minimization problem iteratively using the kernelized version of the iteratively re-weighted least squares (IRWLS) algorithm [Huber *et al.* (1964)].

In all our experiments, we used the implementation provided by the authors (www.eecs.umich.edu/∼cscott) and a Gaussian kernel. The rKDE method has two hyper-parameters that have to be optimized: the bandwidth of the kernel K and the choice of a loss function ρ. The Gaussian kernel width was set to the median distance of a training point to its nearest neighbour [Kim and D. Scott (2012)] and we tested both Huber and Hampel loss functions.

8.4.2 Parameter selection

As discussed in Chap. 4, parameter tuning for outlier detection method is very hard as no labelled examples from the outlier class are available during the training phase. Therefore, unlike in standard classification schemes, it is not possible to perform a cross validation on a grid to select the parameters that maximize a given detection performance measure (e.g. the AUC value).

We propose to take advantage of the probabilistic interpretation of the calibrated SVDD outlier scores to tune the model parameters by optimizing the quality of the probability estimates. Different metrics have been proposed to asses probability estimates in the context of MV-sets estimation [Scott and Nowak (2006)]. Computing these metrics requires generating either a large test set of uniform data or an artificial outlier class. However, the generation of uniform data is less suitable for high dimensional datasets and the generation of an artificial outlier class requires making assumptions on the distribution of observations in the outlier class.
An alternative metric to measure the quality of the estimated MV-sets is to consider the relative difference between the expected probability mass (η_j) of each estimated MV-set and the experimental probability mass ($\tilde{\eta_j}$) computed on a validation set. We define this measure as:

$$\text{Relative Error} = \frac{\tilde{\nu_j} - \nu_j}{\nu_j}$$

where, $\nu_j = 1 - \eta_j$ and $\tilde{\nu_j} = 1 - \tilde{\eta_j}$. For the SVDD generalizations, the kernel width parameter for the Gaussian kernel is varied on a log_2 scale ($2^{[-4:0.1:4]}$) and the optimal kernel width is selected by minimizing the average relative difference over all the considered MV-sets.

8.4.3 Performance measure

The evaluation of the probability estimates depends on the knowledge we have on the data. If the true probability distribution of the data is known, as it is mostly the case for synthetic datasets, than we can consider metrics such as the Kullback-Leibler divergence or the Hellinger distance to measure the reliability of probability estimates. In our experiment on synthetic data we used the Kullback-Leibler divergence measure defined as:

$$\text{DIV}_{kl} = \sum_{i=1}^{n} P_i ln(\frac{P_i}{\tilde{P}_i}),$$

where P is the true probability distribution and $\tilde{P}$ is the estimated probability distribution.

For real datasets, the true probability distribution of the samples is unknown and therefore the metrics mentioned above cannot be considered. For classification problems, the detection performance is usually evaluated by using the receiver operating characteristic (ROC) and the area under the curve (AUC). The ROC curve is constructed by thresholding scores outputted by each method at a level λ and computing the true positive rate and the false positive rate for each threshold. This metric is however not suitable for assessing the reliability of the probability estimates. Most calibration methods do not affect the ranking of outlier scores, therefore the benefits of using calibrated scores as opposed to just using the scores cannot be assessed using ROC curves.
[Kriegel *et al.* (2011)] introduced an Error cost metric that allows evaluating the impact of using calibrated scores on the detection performance while taking into account both the reliability of the probability estimates and each class cardinality to address imbalance between the target and the outlier class.

$$\text{Error Cost} = \frac{1}{2}(\frac{1}{|N|}\sum_{\mathbf{x}\in N}\mathbb{P}(O|\mathbf{x}) + \frac{1}{|O|}\sum_{\mathbf{x}\in O}\mathbb{P}(N|\mathbf{x})),$$

where, $|C|$ is the cardinal of class C, $C = N$ or O and $\mathbb{P}(N|\mathbf{x}) = 1 - \mathbb{P}(O|\mathbf{x})$.

8.5 Experimental Results

8.5.1 Experiments on synthetic data

The training data consists of $n = 500$ data points from a 2D Gaussian distribution $N(0, 1)$. To simulate the presence of noise, outliers drawn from a uniform distribution $U_{[-5,5]}$ were added to the training data. The noise ratio was varied between 1% and 10%. All experiments on synthetic data were repeated 100 times. For both SVDD generalizations we used a linear kernel. For this experiment, we used the Kullback-Leibler (KL) divergence metric between probability estimates ($\tilde{\mathbb{P}}$) and the true probability function ($\mathbb{P}$) to evaluate the performance.

For the considered unimodal Gaussian distribution, the true probability function P is given by the χ_2 cumulative distribution function with 2 degrees of freedom. The divergence

metric was computed between the calibrated SVDD scores for iSVDD and cSVDD and the true probability function ; and between the Gaussian probability density function and rKDE scores.

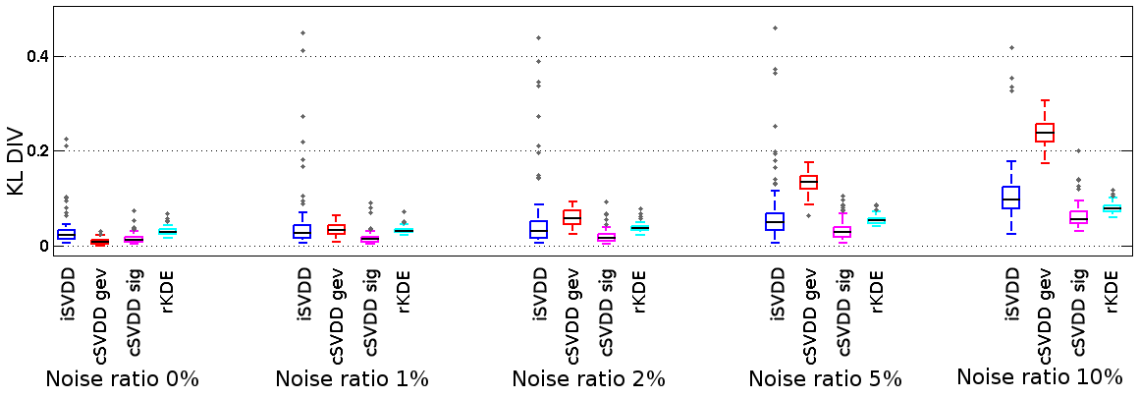


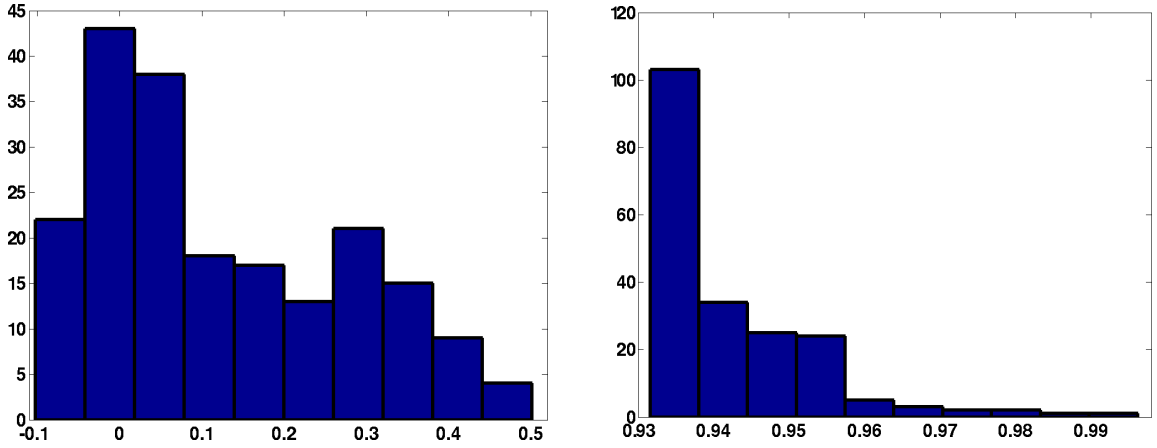
Figure 8.2: KL divergence computed using the true probability and scores obtained with iSVDD (blue), cSVDD gev (red), cSVDD sig (magenta) and rKDE (cyan) for varying noise levels.

Fig. 8.2 shows the results obtained for all four methods. For each method, a boxplot was constructed by considering the KL divergence values obtained over the 100 random samples. For rKDE, the best results were obtained when using Hampel loss and least square cross validation to select the kernel bandwidth.
For the noise free case, the cSVDD approach with the generalized extreme value distribution gave the best performance in terms of KL divergence and very few outlying values are observed. However, this approach is very sensitive to the presence of noise in the training data. For all other noise levels, the cSVDD approach combined with the sigmoid calibration performs best in terms of KL divergence measure and is also the least sensitive to the presence of noise in the training data. For all noise levels, the iSVDD approach seems to be less stable and generates more degenerate models than all other approaches.

8.5.2 The synthetic Banana dataset

The banana dataset is a synthetic two dimensional benchmark classification dataset from the UCI repository. The dataset contains 5300 samples balanced between two classes (2924 versus 2376).
In a first experiment, this dataset is used to illustrate the difference between the two SVDD generalizations and the two calibration schemes. We randomly selected 200 samples from the majority class to form the normal data (*i.e.* the target class) training examples. The parameters of iSVDD and cSVDD were optimized following the procedure in 8.4.2 and 9 MV-sets, linearly spaced between 0.1 and 0.9, were estimated.

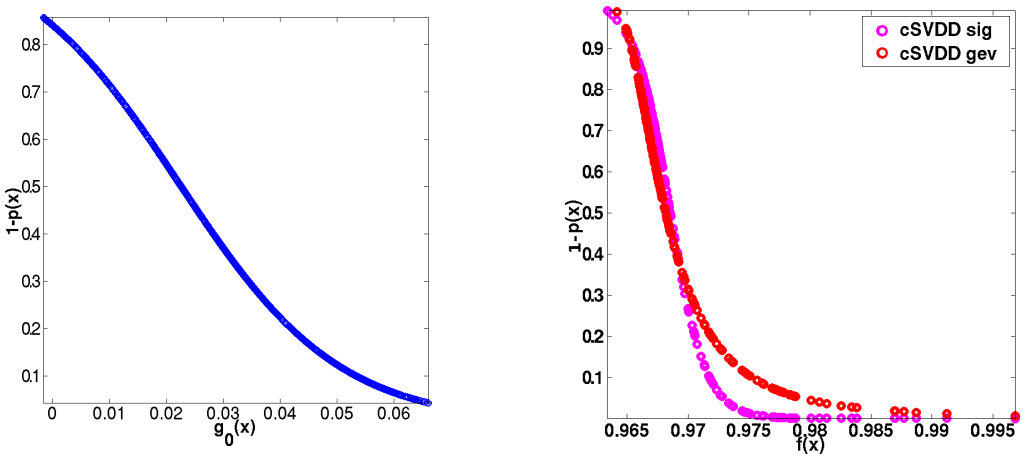
Fig. 8.3 shows score distribution for both iSVDD and cSVDD obtained for the normal training data. In both cases, the observed scores for the normal data tend to have an



(a) *(b)*

Figure 8.3: Score distribution histograms for iSVDD ($g_0 = 9^{th}$ SVDD model) 8.3a and cSVDD 8.3b.

exponential decay. This result justifies partly the suitability of using a sigmoid like calibration function or a generalized extreme value distribution.



(a) *(b)*

Figure 8.4: Calibration function for (a) iSVDD and (b) cSVDD.

Fig. 8.4a shows the calibration function obtained for iSVDD using the sigmoid model. Fig. 8.4b shows the calibration function obtained using the sigmoid model and using the general extreme value distribution model. Comparing the two curves for cSVDD, we see that the calibration function obtained by fitting a gev model presents a slower decay than with the sigmoid model. This is in accordance with the fact that the gev model is well-suited for fitting heavy tailed distributions [Markovich (2011)]. The calibration function obtained by fitting a sigmoid model to iSVDD scores presents a very slow exponential decay. For this dataset, the kernel width parameter selected by optimizing the relative error on the probability mass is larger for iSVDD than for both cSVDD approaches regardless of the considered calibration model. As the scores are proportional to the distance from the SVDD model centre, the scores corresponding to a large value of the kernel width will

have a slower decay.

The Banana dataset being a two dimensional dataset, we can visualize the estimated MV-sets for iSVDD Fig. 8.5a; for cSVDD sig Fig. 8.5c; for cSVDD gev Fig. 8.5d and rKDE Fig. 8.5e. The MV-sets estimated by iSVDD are clearly under-fitting the data and a smaller kernel width would give MV-sets estimates that model better the shape of the distribution of the target class. Fig. 8.5b shows the MV-sets estimated using a smaller kernel width. It should be noted however that the MV-sets on Fig. 8.5a were estimated by using the 9^{th} independent SVDD model. This explains why the estimate obtained for the 9^{th} level-set is more accurate than those obtained for the other level-sets. The MV-sets estimated with the cSVDD gev model successfully capture the shape of the normal distribution of target class and offer a smooth decay of the probability mass associated with the considered MV-sets. The MV-sets estimated with the cSVDD sig model succeed in capturing the shape of the normal distribution only for the MV-sets with high probability mass. The MV-sets estimated using rKDE also have right shape however, we see in this example that the estimated MV-sets overfit the target class training observations especially in regions with a lower density.

To evaluate the performance of the estimated models in the context of outlier detection we used the remaining samples from both classes as test data. As, the considered dataset is balanced, the ROC curve and the AUC value were used to measure performance. Fig. 8.5f shows the ROC curves obtained by using the optimal kernel width for each model. For means of comparison, we also evaluated the performance of the robust kernel density estimator. As no outliers were added to the training data, no notable differences were observed when Hampel or Huber loss function. The highest AUC value was achieved by the cSVDD-sig model, followed by the cSVDD-gev model and the rKDE model, and finally the iSVDD model. If a very low false positive rate is desired than the best model is given by the cSVDD-gev model.

8.5.3 Experiments on Real data : application to outlier detection

We used six benchmark datasets from the UCI database (Blood transfusion, Breast Cancer, SPECTF Heart, Balance, Banana and Pima Indian) [Lichman (2013)]. For each dataset, 200 examples were sampled from the majority class to form the target class training examples. The Balance dataset contains three classes, we therefore considered the two categories that allow having more than 200 instances in the target class and pooled the two remaining categories to form the outlier class.

To simulate the presence of noise in the training data, we randomly selected a portion of examples from the outlier class and added them to the training data. The noise ratio was varied between 0.01 and 0.15. The remaining examples from both classes were used to form the test dataset.

For the real data experiment, the true density (P) is unknown. We therefore compared the different models in terms of the Error costs defined above. For all SVDD generalizations,

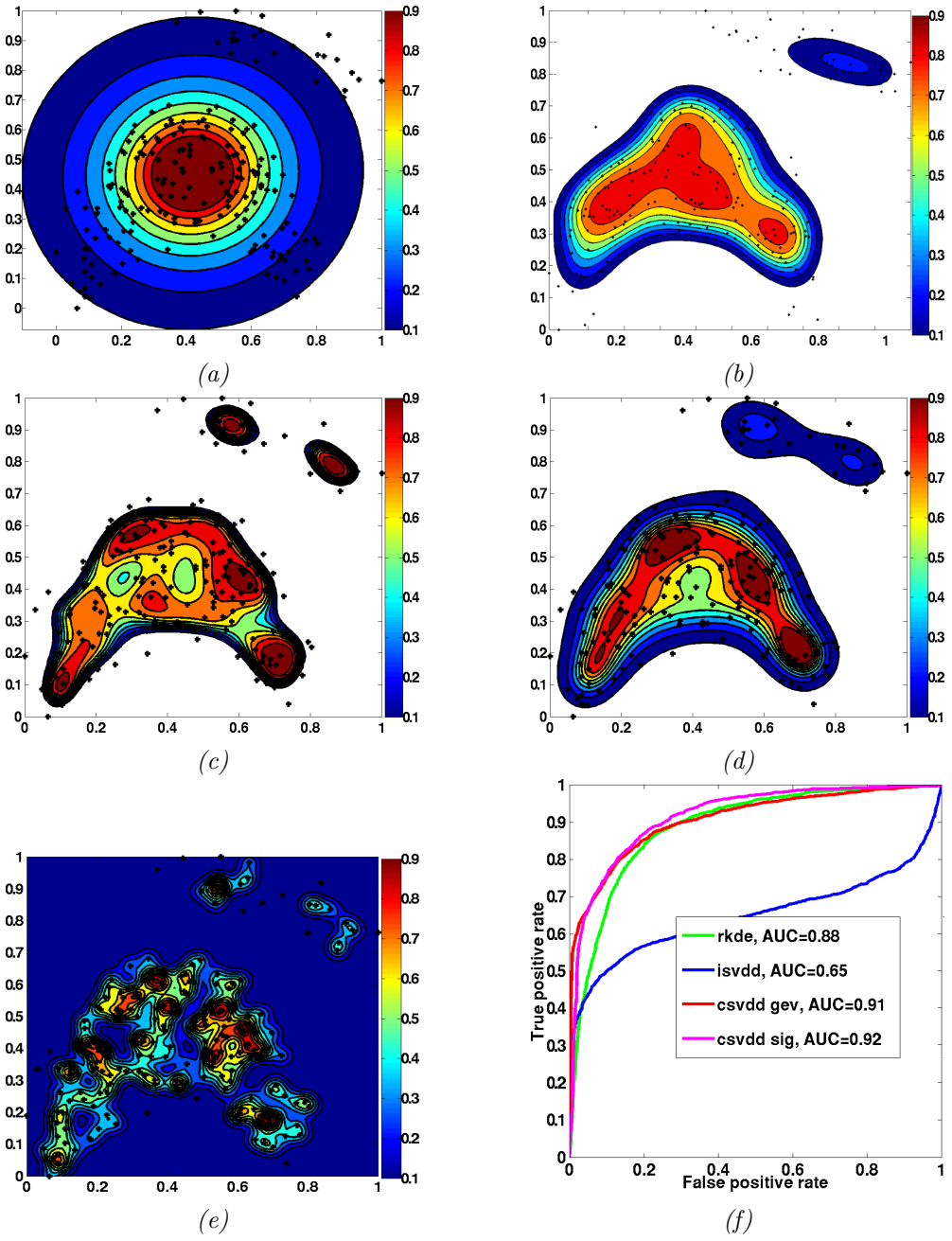


Figure 8.5: MV-sets estimated for the Banana dataset using (a) iSVDD ($g_0 = 9^{th}$ SVDD model), (b) iSVDD ($g_0 = 9^{th}$ SVDD model) with manually picked kernel width, (c) cSVDD-sig, (d) cSVDD-gev, and (e) rKDE. (f) ROC curve comparing all methods.

the Gaussian kernel width was selected following the procedure described in 8.4.2. All experiments using the UCI datasets were repeated 100 times.

Fig. 8.6 shows the absolute value of the relative error on the MV-sets obtained using iSVDD for the Breast Cancer dataset and for varying values of the kernel width. Overall, the relative error is higher for low probability mass MV-sets. For iSVDD the SVDD models are estimated independently from one another and therefore the estimation of low probability mass MV-sets depends more heavily on the considered target training obser-

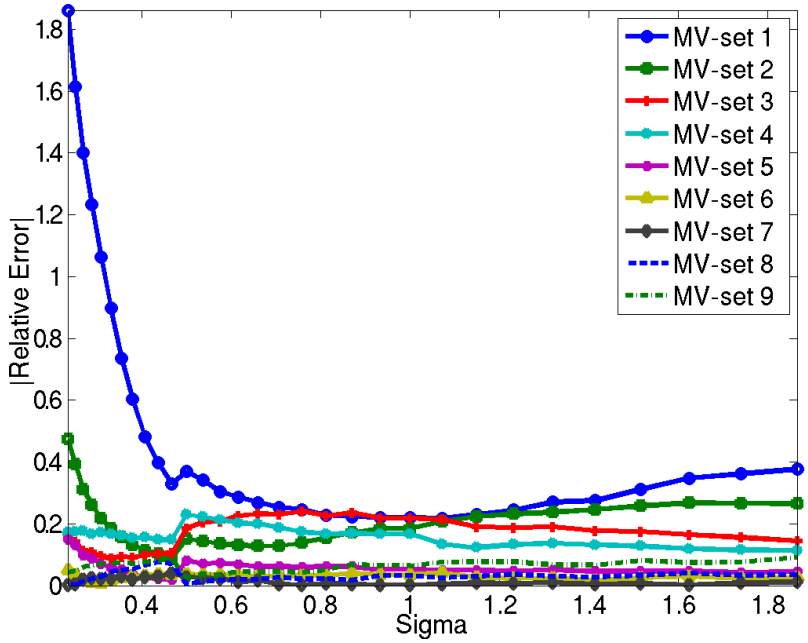


Figure 8.6: Breast Cancer: Absolute value of the relative error as a function of the kernel width sigma for the 9 MV-sets estimated using iSVDD. MV-set 1 corresponds to the MV-set with probability mass equal to 0.1 ($\nu = 0.9$) and MV-set 9 corresponds to the MV-set with probability mass equal to 0.9 ($\nu = 0.1$). The optimal sigma value corresponds to the value that minimizes the average relative error over all MV-sets.

vations in the sense that the empirical probability mass computed using an independent validation dataset would be different from the probability mass obtained using the training observations.

Dataset	Noise ratio 0%				Noise ratio 1%				Noise ratio 2%			
	iSVDD	*cSVDDsig*	*cSVDDgev*	*rKDE*	*iSVDD*	*cSVDDsig*	*cSVDDgev*	*rKDE*	*iSVDD*	*cSVDDsig*	*cSVDDgev*	*rKDE*
Pima Indian	39.4	38.8	38.1	39.1	39.9	39.2	38.3	39.2	39.2	39.0	38.3	39.3
Blood Transfusion	48.2	48.1	47.6	47.4	25.9	26.6	26.0	34.4	30.8	27.0	28.3	34.7
Breast Cancer	26.2	26.4	27.7	28.0	25.5	24.9	26.5	28.2	24.7	25.7	26.4	27.7
SPECTF Heart	60.3	63.0	66.8	61.9	60.2	60.4	63.3	61.2	59.7	60.5	62.8	62.4
Balance LB vs R	49.9	51.4	51.4	49.7	52.1	52.0	52.3	50.4	50.0	50.2	50.3	50.2
Balance RB vs L	29.9	30.6	31.0	33.0	30.9	31.7	30.2	32.6	29.7	29.6	30.4	32.8
Banana	32.6	29.7	30.2	39.1	32.2	28.9	29.9	38.7	30.5	29.5	30.0	39.1

Table 8.1: Median error cost for varying noise ratios. value indicates the best method and value indicates the second best method.

Tab. 8.1 and 8.2 show the obtained median error costs for all 7 test datasets and noise ratios. iSVDD gave good results for 4 out of the 7 considered datasets, and worked particularly well for the SPECTF Heart dataset. The SPECTF Heart dataset has a higher dimension compared with the other datasets (44 features against on average 5 features). Therefore, imposing a hierarchical structure (as in cSVDD) on MV-sets estimated using only 200 training observations may be not suitable for such dataset. Like for the experiments on synthetic data, iSVDD performance has a higher variability than cSVDD (see

Dataset	Noise ratio 5%				Noise ratio 10%				Noise ratio 15%			
	iSVDD	*cSVDDsig*	*cSVDDgev*	*rKDE*	*iSVDD*	*cSVDDsig*	*cSVDDgev*	*rKDE*	*iSVDD*	*cSVDDsig*	*cSVDDgev*	*rKDE*
Pima Indian	40.5	39.3	38.7	39.3	41.2	40.5	38.9	39.3	41.2	40.7	38.5	39.2
Blood Transfusion	29.2	24.7	25.8	34.3	28.3	27.2	27.4	36.0	32.4	29.7	29.8	36.5
Breast Cancer	24.7	25.2	26.5	27.5	26.5	23.6	25.4	27.1	30.5	23.4	25.1	27.3
SPECTF Heart	58.7	62.3	66.5	62.3	59.6	60.6	63.5	62.8	60.5	66.6	64.4	64.4
Balance LB vs R	47.6	49.8	48.3	50.8	49.5	49.1	49.5	51.1	52.0	52.2	52.0	52.0
Balance RB vs L	30.6	29.4	30.7	33.1	30.5	30.3	32.4	33.5	31.5	31.6	32.2	34.4
Banana	30.5	30.6	32.8	39.1	31.5	32.8	33.7	39.9	32.6	34.0	36.7	40.9

Table 8.2: Median error cost for varying noise ratios. value indicates the best method and value indicates the second best method.

Fig. 8.7, 8.8 and 8.9) Comparing the two calibration methods, it is not very clear which one performs best, as the performance depends on the considered dataset. For the Pima Indian dataset, the gev calibration gave the best results for all considered noise ratios. This may suggest that for this dataset, the score distribution is more heavy tailed than for the other datasets. Overall, SVDD generalizations performed better than rKDE for most datasets and noise ratios.

For the Banana dataset, the best performance, based on the error costs metric and without added noise, is obtained with cSVDD sig, followed by cSVDD gev. This results is very coherent with the results found in 8.5.2 when the AUC value was used to measure performance. For most datasets, both cSVDD approaches seem to be sensitive to the presence of noise in the training data. For some datasets such as the Breast cancer dataset (see Fig. 8.8), the performance of all SVDD generalization methods improves when adding more noisy observations. This results may seem surprising, however, the optimization of the kernel width is performed on a validation test that also contains noisy observations and therefore the optimal values of the kernel parameter are different for the different noise level. rKDE seems to be less sensitive to the presence of noise in the training dataset (see Fig. 8.7, 8.8 and 8.9). This may be explained by the use of the median to compute the value of the kernel width for rKDE. The median being less sensitive, than the mean for example, to the presence of noisy or outlier values.

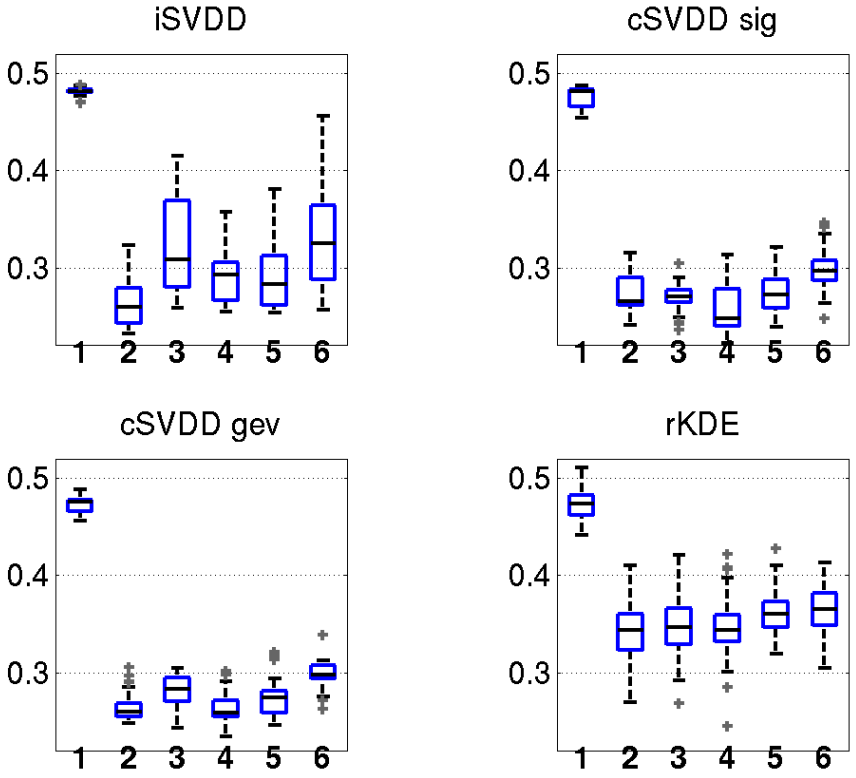


Figure 8.7: Blood Transfusion: Error Cost as a function of noise level. labels: 1, 2, 3, 4, 5, and 6 on the x axis refer to noise levels: 0%, 1%, 2%, 5%, 10% and 15%.

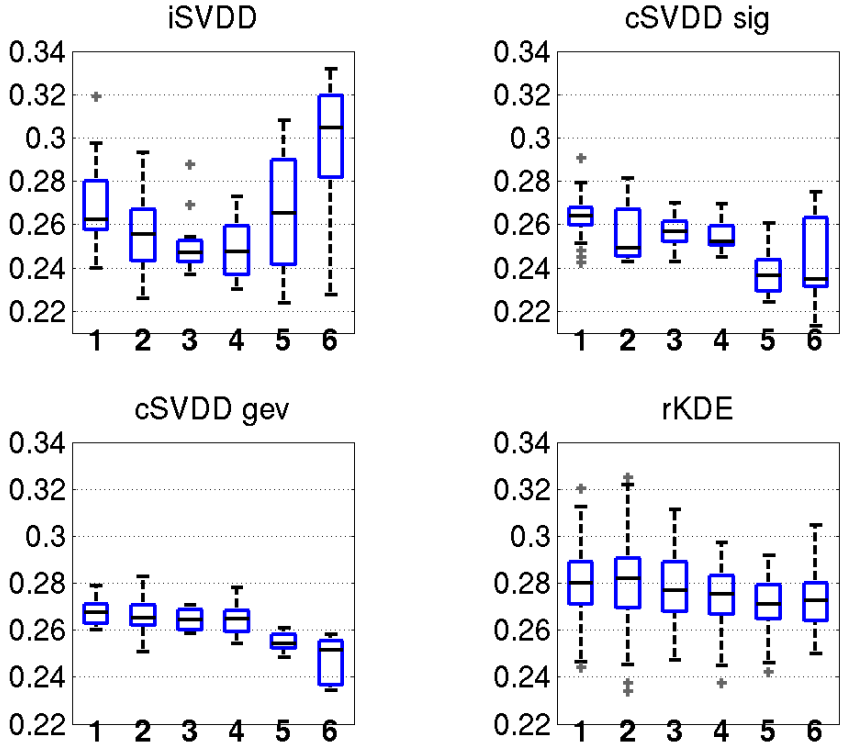


Figure 8.8: Breast Cancer: Error Cost as a function of noise level. labels: 1, 2, 3, 4, 5, and 6 on the x axis refer to noise levels: 0%, 1%, 2%, 5%, 10% and 15%.

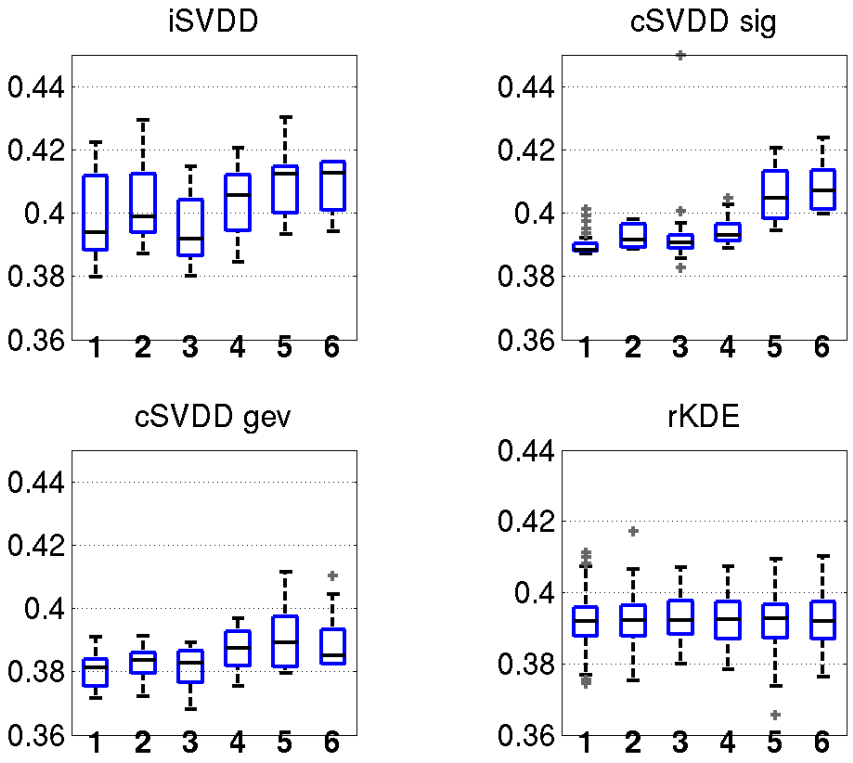


Figure 8.9: Pima Indian: Error Cost as a function of noise level. labels: 1, 2, 3, 4, 5, and 6 on the x axis refer to noise levels: 0%, 1%, 2%, 5%, 10% and 15%.

8.6 Conclusion

In this chapter, we introduced a generalization of the SVDD algorithm for the estimation of hierarchical MV-sets with specific probability masses. This generalization is obtained by transforming the initial SVDD optimization problem to construct q SVDD models that all share the same center. The experiments on synthetic and on real datasets suggest that, in most cases, this concentric model is more robust than the straightforward generalization of SVDD consisting of building q independent SVDD models associated with different probability masses. However, the computational complexity associated with the construction of q concentric SVDD models is way higher than in the case of independent models. Our experience showed that, the gain in performance observed in using concentric models rather then independent ones is only notable when the independent models are too far from being concentric. Consequently, one way of reducing computation time would be to first train q independent models and then train the concentric models only if the independent models are very far from being concentric. Recently, Thomas et al. [Thomas *et al.* (2015)] proposed a new extension of the OCSVM algorithm for constructing single OCSVM model that allows a better estimation of a single MV-set than the original version of the OCSVM algorithm. By using as a building block such a revised version after adapting it to the SVDD algorithm, the performance of our proposed generalization could also be improved.

To improve score interpretability of the proposed generalizations of the SVDD model, two calibration functions have been introduced. The key idea used in designing the calibration models is the consistency of the MV-sets estimated by each SVDD model. The estimated MV-sets were used to fit either a sigmoid like calibration function or a generalized extreme value distribution. For real data, the proposed models were evaluated using Error costs metric that allows evaluating both the detection performance and the calibration of the probability estimates.

With regards to the parameter selection method, we proposed to use the average relative error measured on the probability masses associated with each MV-set as an optimality criterion to choose the best parameters. Unlike for other performance metrics (e.g. accuracy, AUC) computing this measure does not require having labelled observations from the outlier class or making assumptions on the distribution of outliers (e.g. uniform).

Chapter 9

Conclusions and perspectives

The objective of this PhD work was to design a computer aided diagnosis system (CAD) capable of extracting the main discriminative information from neuroimaging data used during the pre-surgical evaluation of patients with intractable epilepsy.

Contributions

An important step in CAD system design is specification drafting and problem analysis. The data available in this project consisted of unlabelled intractable epilepsy patient neuroimaging data (mainly MR images) and databases of healthy controls subjects. The desired output of the CAD system was a labelled cluster map highlighting suspicious brain areas exhibiting abnormalities associated with intractable epilepsy. For a better interpretation of this output and further processing, cluster labels should represent well calibrated suspicion scores and ideally correspond to the probability of the cluster being an epilepsy related abnormality.

Given these specifications, we identified various challenges. Most of these challenges arise from to the nature of neuroimaging data. These included handling noisy (*e.g.* artefacts), high dimensional (approximately 1.5 million voxels in an MR scan), multi-modality images (*e.g.* PET and multiple MRI sequences) and small size of training datasets. Dealing with class imbalance was also a challenging task. Class imbalance is inherent to the detection task at hand as data from healthy control subjects is highly available compared to the very expensive annotated patient data. The imbalance is further accentuated by the small size of intractable epilepsy lesions and their heterogeneity. A last challenge was to provide meaningful outputs of the CAD system that can be combined with other

diagnostic tests in order to assist the experts in making their final decision.

Throughout this work we tried to base all our contributions on our analysis of the problem and to propose methods to deal with the identified challenges.

Our first contribution was to reformulate the problem in the context of outlier detection. Instead of building a classifier that allows discriminating between pathological and healthy observations, we proposed building a description of a typical healthy brain and then confront our patient data with this description. To this end, we proposed a first CAD system based on the one-class support vector (OC-SVM) algorithm. OC-SVM was trained in a voxel-wise basis using features extracted from MRI scans of healthy control subjects. The main motivation for this formulation was to avoid the dependence on labelled patient data and to deal with class imbalance. The outputs of the CAD system was transformed into calibrated scores by building a normative score distribution estimated using leave-one-out on the control subjects' score distributions. This also allowed controlling for the type I error by choosing a threshold value that corresponds to a user-defined p-value. The CAD system was evaluated using both realistic simulation data and patient data. We compared its performance against that of an optimized statistical parametric mapping (SPM) analysis and shown that the proposed system outperforms this analysis. It also compared favourably with recent state-of-the-art methods based on two step classification schemes and more complex features.

Our second contribution was to extend the support vector data description (SVDD) algorithm to the case of learning from uncertain observations. We proposed to substitute the Hinge loss in the original formulation of the SVDD algorithm by a 0-1 loss (l_0 penalty) that allows reducing the effect of noisy observations on the estimated target data description. An iterative procedure was adopted to solve the optimization problem after relaxing the l_0 penalty using a logarithmic approximation. The proposed L_0-SVDD was evaluated using realistic simulation based on clinical data and datasets from the UCI repository. Our results showed that, compared with the original formulation, the use of the l_0 penalty successfully reduces the effect of wrongly labelled training observations.

Our third contribution was to investigate an optimal fusion strategy to combine multi-sequence MRI data. We proposed two fusion strategies. The first one corresponded to an early fusion strategy consisting in a single global OC-SVM classifier built using features extracted from three MRI sequences (T1w, T2 FLAIR and DTI). The second fusion strategy corresponded to a late fusion strategy where three base OC-SVM models were trained separately, each using features extracted from a single sequence. The outputs of the base classifiers were then combined using majority voting. To obtain calibrated scores, the fusion approach was coupled with univariate statistical tests that were combined using Stouffer's method. The proposed fusion strategies were evaluated using clinical patient data and validated against the reference SEEG (depth electrodes) exam. A good correlation was found between the identified suspicious brain areas using the late fusion approach and the most activated SEEG electrodes.

Our last contribution was to propose a general framework for converting SVDD scores

into probability estimates. A two step strategy was proposed. First, we generalized the SVDD algorithm for the estimation of hierarchical minimum volume sets (MV-sets) with specific probability masses. To this end, we proposed estimating multiple SVDD models and controlling the probability masses of the associated MV-sets by taking advantage of the $\nu-$property of the SVDD algorithm. The hierarchical structure was imposed by forcing the different SVDD models to share the same hyper-sphere centre. The outputs of the SVDD generalization were then converted to posterior probabilities by considering two calibration functions: a sigmoid function or a generalized extreme value distribution. The proposed framework was tested using synthetic datasets and datasets from the UCI repository.

Perspectives

The work presented in the two first parts of this manuscript mainly focused on the design of a CAD system for intractable epilepsy lesion detection. The system proposed in Chap. 5 was based only on features extracted from T1 weighted MR images. In Chap. 7, we proposed an extension of this CAD system based on features extracted from multi-sequence MR images that included T1 weighted, T2 FLAIR and diffusion imaging. This was a first step toward a truly multi-modal CAD system for intractable epilepsy lesion detection. One interesting perspective of this work would be to test the proposed detection scheme using both multi-sequence MRI and PET. Most CAD systems proposed in the literature and reviewed in Chap. 2 focused on the use of MRI. Only a few studies tried to develop automated methods for the detection of epileptogenic lesions on PET images. The joint analysis of the two modalities was restricted to the visual inspection of co-registered T1w MRI and PET images. However, an interesting finding of these studies was that areas of PET hypo-metabolism associated with intractable epilepsy often co-localized with grey matter areas on structural T1w MRI. This finding helped greatly improve the visual detection performance of epileptogenic lesions and especially for MRI negative cases. Incorporating this type of prior in a learning framework can also help improving the detection performance of an automated system. The learning framework proposed in 7 could be easily adapted to take into account this prior. One way of doing so would be to modify the score combination rule. For instance, majority voting can be replaced by a weighted voting scheme that would give more weights to base detectors that provided an output that is coherent with the co-localisation prior. Of course, this prior could also be incorporated in an earlier step of the CAD system. For instance, one can also imagine modifying the similarity kernel matrix used by each base classifier to take into account this prior.

In Chap. 6 and Chap. 8, we proposed two interesting extensions of the SVDD learning algorithm. The development of both approaches was motivated by our clinical application. The L_0-SVDD formulation introduced in Chap. 6 was aimed at reducing the effect of the presence of wrongly labelled training observations on the decision boundary of the SVDD

algorithm. Through a motivation example, we showed in Chap. 6 that this type of noise is present in the datasets that we used in this project. We illustrated the benefits of using this new formulation on realistic simulation data. We did not however apply this method to our clinical data. This is mainly due to the fact that, unlike in the UCI datasets, in our clinical dataset we had no labelled pathological observations to help with model selection. As discussed at the end of Chap. 6, our attempts at optimizing the hyper-parameters of the L_0-SVDD algorithm using the leave-one-out estimate of the target error were not successful. One possibility here is either to consider gathering more patient cases and to ask expert neurologists to provide annotations for these patient cases. Another less tedious possibility could be to consider using simulations. Artificial simulations (*e.g.* a uniform distribution of outliers is assumed) can of course be used to provide examples from the outlier class and then standard cost sensitive performance measures can be used to optimize the hyper-parameters of the L_0-SVDD algorithm. Another alternative would be to use observations corresponding to the realistic simulations described in Chap. 5 in the validation dataset and then optimize the hyper-parameters via maximizing a standard cost sensitive performance measures.

The methodology proposed in Chap. 8 for converting outlier scores into probability estimates was also only evaluated using datasets from the UCI repository. An interesting perspective of this work would be to apply this method to our clinical data. The main challenge preventing the employment of this method in our clinical application is the small sample size. In all our healthy control subject datasets, the number of observations never exceeded 40. This sample size is way too small to allow accurately estimating multiple minimum volume sets with a hierarchical structure. For our UCI datasets, we used 200 observations for MV-sets estimation and our performance was degraded for high dimensional datasets (*e.g.* SPECTF Heart) suggesting that maybe not enough training observations were used. To deal with this sample size challenge, two future directions can be investigated. The first direction could consist in trying to make the best use of all available healthy control datasets. Through our collaborations, we already had access to three healthy control databases totalling 112 healthy control T1w images. However, the images in these three databases were not necessarily acquired using the same acquisition protocol and therefore do not share the same characteristics. Adaptation domain methods allow learning from different sources and can be investigated in this context to learn from all available datasets. The second direction could be to consider learning region specific models instead of voxel-wise models. This would increase the training sample size and in the same time allow taking into account the neighbourhood information. Extracting regular patchs randomly to form the regions and using a one-class algorithm (OC-SVM of SVDD) could result in estimating a too loose decision boundary that do not characterize well the target normal class distribution and would not allow detecting epilepsy related abnormalities that are usually located at the boundaries (or edges) of different brain structures. The goal of increasing the sample size is not really to add novel information that was not present in the training dataset but rather to give more statistical power to

the estimator. One possibility here is to consider a clustering approach where the goal is to put in the same cluster voxels that have similar feature vectors given a similarity measure. The output of this clustering step would correspond to homogeneous regions in terms of the features that can be used to estimate the MV-sets. Various clustering approaches exist and each of these methods has its own hyper-parameters that would require fine tuning which again would be difficult in this completely unsupervised setting. One possibility here is to use the metric introduced for evaluating probability estimates in Chap. 8 in order to find the best hyper-parameters of the clustering approach.

Publications

- M. E. Azami, C. Lartizien and S. Canu, "*Converting SVDD scores into probability estimates: Application to outlier detection*," Neurocomputing, 2017 [in preparation].
- M. E. Azami, A. Hammers, J. Jung, N. Costes, R. Bouet and C. Lartizien, "*Detection of lesions underlying intractable epilepsy on T1-weighted MRI as an outlier detection problem*," PloS ONE, 2016 [under revision].
- M. E. Azami, C. Lartizien and S. Canu, "*Converting SVDD scores into probability estimates*," 2016 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, 2016.
- M. E. Azami, R. Bouet, J. Jung, A. Hammers and C. Lartizien, "*Combining multi-parametric MR images for the detection of epileptogenic lesions*," 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), New York, NY, 2015, pp. 122-125.
- M. E. Azami, C. Lartizien and S. Canu, "*Robust outlier detection with L_0-SVDD*," 2014 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, 2014.
- M. E. Azami, A. Hammers, N. Costes and C. Lartizien, "*Computer Aided Diagnosis of Intractable Epilepsy with MRI Imaging Based on Textural Information*," Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on, Philadelphia, PA, 2013, pp. 90-93.

Appendix

Appendix A

Magnetic resonance imaging

Physical principle

What is a nucleus? Nuclei are composed of positively charged protons and uncharged neutrons bound together by nuclear forces. The nuclei are of interest in NMR because they possess a spin that can be thought of as a magnetic moment vector.

Nuclei response to an external magnetic field? When placed in an external static magnetic field ($\mathbf{B_0}$), the spin, similar to a small magnet with north and south poles, aligns either parallel or anti-parallel with the external field $\mathbf{B_0}$. The parallel configuration corresponds to a lower energy state and therefore, more spins align in this configuration. This imbalance produces a net longitudinal magnetisation $\mathbf{M}$ aligned with $\mathbf{B_0}$. The spins precess around the static field $\mathbf{B_0}$ at the Larmor frequency $\omega_L = \gamma B_0$, where γ is the gyromagnetic ratio characterizing the nuclei.

Nuclei excitation and relaxation? Applying an electromagnetic RF excitation field $\mathbf{B_1}$ oscillating at the Larmor frequency to the magnetisation net $\mathbf{M}$, removes it from equilibrium and tips it from longitudinal to transverse plane. When the RF signal is then switched off, the transverse component of the precessing magnetization produces a signal measurable by a receiver coil as the system returns to equilibrium. Let $\mathbf{M_z}$ and $\mathbf{M_{xy}}$ be the longitudinal and the transverse components of $\mathbf{M}$. $\mathbf{M_z}$ experiences exponential recovery with a time constant $T1$: $M_z(t) = M_0(1 - e^{-\frac{t}{T1}})$, where M_0 represents the longitudinal magnetisation at equilibrium. $\mathbf{M_{xy}}$ experiences exponential decay with a time constant $T2$: $M_{xy}(t) = M_{xy}(0)e^{-\frac{t}{T2}}$, where $M_{xy}(0)$ corresponds to the transverse magnetisation at the beginning of the relaxation. Time constants ($T1$, $T2$) are very important as different tissues have different relaxation parameters. Magnetic field inhomogeneities cause the transverse component $\mathbf{M_{xy}}$ to decay faster than can be explained by $T2$ relaxation alone. This effect is known as $T2^*$ relaxation.

Image encoding and formation

With only the homogeneous B_0 field present, the system does not contain any spatial information. The received signal after applying the B_1 oscillating field is a complex harmonic with a single frequency peak centred at the Larmor frequency. One way of addressing this problem is by making the magnetic field strength dependant of the spatial position.

Encoding gradients Spatial encoding is performed by successively applying magnetic field gradients. Three gradients are used: a slice selection gradient (SSG), a phase encoding gradient (PEG) and a frequency encoding gradient (FEG).

First, a SSG is applied along the z axis perpendicular to a slice plane. This gradient is added to the static magnetic field B_0 which results in a Larmor frequency varying linearly along the z axis. A RF pulse is simultaneously applied with the same Larmor frequency as that of the protons in the desired slice plane. This causes excitation of only the protons located in this plane and therefore no NMR signal is recorded from the protons located outside of the selected slice plane.

Then, a PEG is applied to modify the spin resonance frequencies inducing dephasing. This results in all the protons processing at the same frequency but in different phases. Protons belonging to the same row (y axis), perpendicular to the PEG direction, will have the same phase. The introduced phase differences last until the signal is recorded.

Finally, a FEG in the last direction, in our example x axis, is applied when the NMR signal is received. This results in having columns of protons that have the same Larmor frequency. A proton within the selected slice is than unequivocally characterized by a specific phase which depends upon its position in y, and a specific frequency which depend upon its position in x.

k-space For each selected slice plane, the recorded MR signal is stored in k-space (k_x,k_y) which is equivalent to a Fourier plane. The frequency encoding gradient is mapped to k_x axis and the phase encoding gradient is mapped to k_y axis. To reconstruct the image slice corresponding to the selected slice plane, the 2D inverse Fourier transform is applied to the k-space.

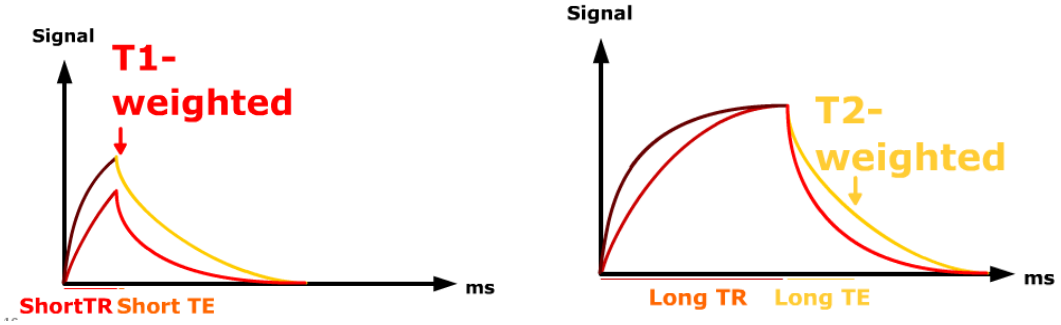
The easiest way to fill the k-space is to use a line-by-line rectilinear trajectory. One line of k-space is fully acquired at each excitation, containing low and high-horizontal spatial frequency information. Between each repetition, there is a change in the phase encoding gradient PEG strength (from negative to positive intensity), corresponding to a change in k_y coordinate from top to bottom. While Cartesian trajectories are by far the most popular ones, many other trajectories are in use, including sampling along radial lines and sampling along spiral trajectories. These alternative sampling trajectories are aimed at reducing the acquisition time and/or improving the image quality.

MRI sequences

The RF pulse that tips $\mathbf{M}$ into the transverse plane is usually referred to as the 90° RF pulse whereas the one that flips the net magnetisation M from $+z$ to $-z$ without producing any transverse magnetisation is called the 180° RF pulse. After a 90° RF pulse, spins dephase and applying the 180° RF pulse brings back the spins in phase.

Spin Echo sequence is a commonly used sequence in MR imaging. This sequence is based on the repetition of 90° and 180° RF pulses. It has two parameters: the echo time (TE), the time between the 90° RF pulse and MR signal sampling and the repetition time (TR), the time between two 90° RF pulses. The 180° RF pulse is applied at time TE/2.

Setting the values of TE and TR allows signal weighting and adjusting tissue contrast. Fig. A.1 shows the recorded NMR signals from two tissues A and B with different time constants T1 and T2. Depending on the choice of TE and TR, differences from either relaxation times T1 or T2 can be exploited to adjust tissue contrast. By setting TR to short values (see Fig. A.1a), tissue contrast will depend on differences in longitudinal magnetisation recovery (T1). This sequence is usually called T1-weighted. By setting TR to long values (see Fig. A.1b), the T1 effect on tissue contrast will be reduced. If TE is long enough differences in transverse relaxation will alter tissue contrast. This sequence is usually called T2-weighted.



(a) *(b)*

Figure A.1

Inversion recovery sequence is a pulse sequence that allows suppressing the signal of a given tissue. This sequence starts with a 180° RF pulse that is followed by a standard spin echo sequence. After the first inversion, and due to longitudinal relaxation, the longitudinal magnetisation $\mathbf{M_z}$ will increase to return to its initial value, passing through null value. The delay between the 180° FR pulse and the 90° RF pulse from the spin echo sequence is referred to as the inversion time TI. In the recorded signal, depending on their time constant T1, some tissues will have negative signal values while others will have positive ones. The inversion time TI can be adapted to a specific T1 value to suppress (render null) the signal from the associated tissue.

The fluid-attenuated inversion recovery (FLAIR) sequence is a special case of the inversion recovery where the fluid signal is suppressed. In brain imaging it is used to suppress cerebrospinal fluid (CSF) effects on the image so as to bring out periventricular hyperintense lesions.

Diffusion sequence The aim of these diffusion-weighted sequences is to obtain images whose contrast is influenced by the differences in water molecule mobility. The water molecules encounter different obstacles in the body (cell membranes, proteins, macromolecules, fibers...), which vary according to the tissues and certain pathological modifications (intracellular edema, abscess, tumors...). Diffusion data therefore provides indirect information about the structure surrounding these water molecules.

A Diffusion-weighted sequence is obtained by adding diffusion gradients during the

preparatory phase of a spin echo sequence. The diffusion gradients are strong and symmetrical in relation to the 180° rephasing pulse. Since precession is proportional to the magnetic field strength, during the first gradient, the protons begin to precess at different rates, resulting in dispersion of the phase and signal loss. The second gradient pulse is applied in the same magnitude but with opposite direction to rephase the spins. The protons that have not moved during the time interval between the gradient pulses will be rephased and the NMR signal is recovered. However, the rephasing will not be perfect for protons that have moved and the recorded NMR signal is reduced.

Appendix B

Neuroanatomy

The human brain is one of the most complex organs in the body. It is made up of two cerebral hemispheres, each of which can be divided into four main lobes. Fig. B.1 shows the different brain lobes.

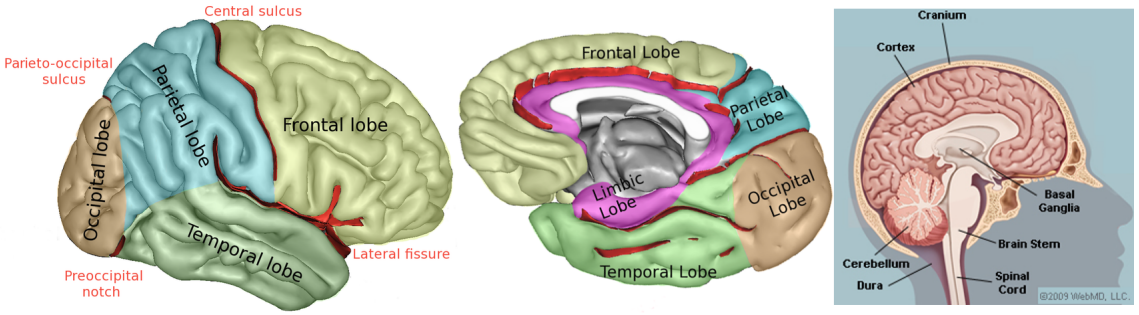


Figure B.1: Brain lobes.

Each hemisphere includes a cortex composed of grey matter that consists mainly of neuronal cell bodies. It contrasts with the underlying white matter consisting mainly of neuronal axons connecting the cell bodies. The cerebral cortex is folded. The peak of a fold is called a gyrus, and its trough is called a sulcus. Thanks to its folded structure, a larger cortical surface is allowed within the cranium. The brain and the spinal cord are enveloped by three layers of tissue called meninges: the dura, the arachnoid and the pia. The cerebrospinal fluid (CSF) is located in the subarachnoid space between the arachnoid and the pia.

The limbic lobe is an arc shaped region of the cortex on the medial surface of each hemisphere. The hippocampus and the amygdala are one of the main structures that compose the limbic system.

Appendix C

Computation of the SVDD radius R and OC-SVM bias ρ

SVDD primal formulation

Let $(\mathbf{x}_i)_{i=1\ldots n}$, $\mathbf{x}_i \in \mathcal{R}^p$ be n training observations from a target class. The decision function associated with the SVDD algorithm is of the form $f(\mathbf{x}) = \text{sign}((\mathbf{x}-\mathbf{a})^\top(\mathbf{x}-\mathbf{a}) - R^2)$ where R is the radius of the enclosing hypersphere with minimum volume and $\mathbf{a}$ its centre. The centre a and the radius R are obtained by solving the following constrained minimization problem:

$$\begin{cases} \min\limits_{R,\mathbf{a},\boldsymbol{\xi}} & R^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{s.t} & (\mathbf{x}_i-\mathbf{a})^\top(\mathbf{x}_i-\mathbf{a}) \leq R^2+\xi_i, \qquad i=1,\ldots,n \\ \text{and} & \xi_i \geq 0, \qquad\qquad\qquad\qquad\quad i=1,\ldots,n, \end{cases} \tag{C.1}$$

where ξ_i are slack variables that allow relaxing the inequality constraints.

SVDD dual Formulation

To derive the dual formulation of problem C.1, we look for a saddle point of the Lagrangian $\max\limits_{\boldsymbol{\alpha},\beta}\min\limits_{\mathbf{a},R,\boldsymbol{\xi}} L(\mathbf{a},R,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$ with $\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0$.

$$L(\mathbf{a},R,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[(R^2+\xi_i) - (\mathbf{x}_i-\mathbf{a})^\top(\mathbf{x}_i-\mathbf{a}))] - \sum_{i=1}^{n}\beta_i\xi_i$$

Setting to 0 the derivatives with respect to the primal variables gives:

- $\nabla_{\mathbf{a}} L(\mathbf{a},R,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = -2\sum\limits_{i=1}^{n}\alpha_i(\mathbf{x}_i-\mathbf{a})$

- $\frac{\partial L}{\partial R}(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 2R(1 - \sum_{i=1}^{n} \alpha_i)$
- $\nabla_{\boldsymbol{\xi}} L(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = C\mathbf{e} - \boldsymbol{\alpha} - \boldsymbol{\beta}$ where $\mathbf{e} = [\underbrace{1 \dots 1}_{\text{size n}}]^T$

The optimality conditions are then:

$$\begin{cases} \mathbf{a} & = & \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \\ \sum_{i=1}^{n} \alpha_i & = & 1 \\ 0 \leq & \alpha_i & \leq C \end{cases} \tag{C.2}$$

Substituting the primal variables in the Lagrangian gives:

$L(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) =$

$$= \quad R^2 + C\sum_{i=1}^{n} \xi_i - [\sum_{i=1}^{n} \alpha_i (R^2 + \xi_i)] - \sum_{i=1}^{n} \alpha_i [(\mathbf{x}_i - \sum_{j=1}^{n} \alpha_j \mathbf{x}_j)^\top (\mathbf{x}_i - \sum_{j=1}^{n} \alpha_j \mathbf{x}_j)] - \sum_{i=1}^{n} (C - \alpha_i)\xi_i$$

$$= \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^{n} \alpha_i \mathbf{x}_i^\top \mathbf{x}_i$$

Noting that: $\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j = \alpha^\top G \alpha$ and $\sum_{i=1}^{n} \alpha_i \mathbf{x}_i^\top \mathbf{x}_i = \alpha^\top diag(G)$ with G being the Gram matrix defined as: $G_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$, the following dual formulation of the SVDD algorithm is obtained:

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top diag(G) \\ \text{s.t} & \mathbf{e}^\top \boldsymbol{\alpha} = 0 \\ \text{and} & 0 \leq \alpha_i \leq C, \qquad\qquad i = 1 \dots n \end{cases} \tag{C.3}$$

SVDD radius computation using the bi-dual

To derive the expression of the primal variable R, we consider the bi-dual formulation of problem C.1. We look for a saddle point of the Lagrangian associated with problem C.3: $\max\limits_{eq, \boldsymbol{p}_0^0, \boldsymbol{p}_0^1} \min\limits_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, eq, \boldsymbol{p}_0^0, \boldsymbol{p}_0^1)$, where eq is the Lagrange multiplier associated with the equality constraint on $\boldsymbol{\alpha}$ and $p_0^0, p_0^1 \geq 0$ are the Lagrange multipliers associated with the box constraints.

Le Lagrangian of problem C.3 is:

$$L(\boldsymbol{\alpha}, eq, \boldsymbol{p}_0^0, \boldsymbol{p}_0^1) = \boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top diag(G) + eq(e^\top \boldsymbol{\alpha} - 1) + \boldsymbol{p}_0^{1\top}(\boldsymbol{\alpha} - C\mathbf{e}) - \boldsymbol{p}_0^{0\top}\boldsymbol{\alpha} \quad \text{(C.4)}$$

with: $\boldsymbol{p}_0^0, \boldsymbol{p}_0^1 \geq 0$.

The derivative of the Lagrangian with respect to $\boldsymbol{\alpha}$ is:

$$\frac{\partial L}{\partial \boldsymbol{\alpha}} = 2\boldsymbol{\alpha}^\top G - diag(G) + eq\mathbf{e}^\top + \boldsymbol{p}_0^{1\top} - \boldsymbol{p}_0^{0\top}.$$

To derive a simple bi-dual formulation, the trick is to consider the following implication:

$$\frac{\partial L}{\partial \boldsymbol{\alpha}} = 0 \Rightarrow \frac{\partial L}{\partial \boldsymbol{\alpha}}\boldsymbol{\alpha} = 0.$$

We have:

$$\frac{\partial L}{\partial \boldsymbol{\alpha}}\boldsymbol{\alpha} = 2\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top diag(G) + eq\ \mathbf{e}^\top\boldsymbol{\alpha} + \boldsymbol{p}_0^{1\top}\boldsymbol{\alpha} - \boldsymbol{p}_0^{0\top}\boldsymbol{\alpha} \quad \text{(C.5)}$$

By identifying the Lagrange expression (Eq. C.4) in (Eq. C.5), we obtain that:

$$\frac{\partial L}{\partial \boldsymbol{\alpha}}\boldsymbol{\alpha} = 0 \Leftrightarrow L = -\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{p}_0^{1\top}C\mathbf{e} - eq$$

The resulting bi-dual optimization problem is:

$$\begin{cases} \min\limits_{eq, \boldsymbol{p}_0^0, \boldsymbol{p}_0^1} & -L = \boldsymbol{\alpha}^\top G\boldsymbol{\alpha} + eq + C\boldsymbol{p}_0^{1\top}\mathbf{e} \\ \text{s.t} & 2G\boldsymbol{\alpha} - diag(G) + eq\mathbf{e}^\top + \boldsymbol{p}_0^1 - \boldsymbol{p}_0^0 = 0 \\ \text{and} & 0 \leq \boldsymbol{p}_0^0 \\ \text{and} & 0 \leq \boldsymbol{p}_0^1 \end{cases} \quad \text{(C.6)}$$

From the first dual formulation, we already had that:

$$\mathbf{a} = x^\top \boldsymbol{\alpha}$$

$$\mathbf{a}^\top \mathbf{a} = \boldsymbol{\alpha}^\top G\boldsymbol{\alpha}$$

$$\forall i \in [1, n],\ (\mathbf{a}^\top \mathbf{x}_i) = \sum_{j=1}^{n} \alpha_j \mathbf{x}_j^\top \mathbf{x}_i = G_{i,\cdot}\boldsymbol{\alpha},$$

where $G_{i,\cdot}$ corresponds to the i^{th} line of matrix G.

We also note that the equality constraint and the inequality constraints on the Lagrange multipliers $\boldsymbol{p}_0^0$ give:

$$2G\boldsymbol{\alpha} - diag(G) + eq\mathbf{e}^T + \boldsymbol{p}_0^1 = \boldsymbol{p}_0^0 \geq 0.$$

By writing:

$$\boldsymbol{p}_0^1 = \boldsymbol{\xi}$$

$$eq + \boldsymbol{\alpha}^\top G\boldsymbol{\alpha} = R^2$$

Eq. C.6 becomes:

$$\begin{cases} \min\limits_{eq,\boldsymbol{\xi}} & R^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{s.t} & \boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - 2G\boldsymbol{\alpha} + diag(G) \leq R^2 + \xi_i \quad i = 1\dots n \\ \text{and} & 0 \leq \xi_i \qquad\qquad\qquad\qquad\qquad\quad i = 1\dots n \end{cases} \tag{C.7}$$

But,

$$(\mathbf{x}_i - \mathbf{a})^\top(\mathbf{x}_i - \mathbf{a}) = \mathbf{x}_i^\top\mathbf{x}_i - 2\mathbf{a}^\top\mathbf{x}_i + \mathbf{a}^\top\mathbf{a}$$
$$(\mathbf{x}_i - \mathbf{a})^\top(\mathbf{x}_i - \mathbf{a}) = diag(G) - 2G\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top G\boldsymbol{\alpha}$$

Put all together, we obtain the original SVDD primal problem in Eq. C.1.

$$\begin{cases} \min\limits_{R,\boldsymbol{\xi}} & R^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{s.t} & (\mathbf{x}_i - \mathbf{a})^\top(\mathbf{x}_i - \mathbf{a}) \leq R^2 + \xi_i \quad i = 1\dots n \\ \text{and} & 0 \leq \xi_i \qquad\qquad\qquad\qquad\quad i = 1\dots n \end{cases} \tag{C.8}$$

And the radius is given by:

$$R^2 = eq + \mathbf{a}^\top\mathbf{a}.$$

What about the bias term ρ for OC-SVM

The same computation can be carried to deduce the expression of the bias term ρ for OC-SVM. In this case, we obtain that:

$$\rho = eq,$$

where, eq is the Lagrange multiplier associated with the equality constraint on the dual variable $\boldsymbol{\alpha}$ like for the SVDD case.

Appendix D

Equivalence between OC-SVM and SVDD

The following proof, written by Pr. Stéphane Canu, proves the equivalence between OC-SVM and SVDD when a kernel with constant diagonal is used. The relation between OC-SVM and SVDD primal variables is also given.

ABOUT THE EQUIVALENCE BETWEEN SVDD AND OCSVM (THE DEVIL IS IN THE DETAILS)

STÉPHANE CANU, LITIS, INSA DE ROUEN

Abstract. SVDD and OCSVM are equivalent when $\forall \mathrm{x} \in \Omega,\ k(\mathrm{x},\mathrm{x}) = c$ where c is some constant. But be careful with the Lagrange multipliers and the way you compute R.

1. Equivalence between SVDD and OCSVM for diagonal constant kernels

Lemma 1.1. *Let $\mathcal{H}$ be a RKHS on some domain Ω endowed with kernel k. If there exists some constant c such that $\forall \mathrm{x} \in \Omega,\ k(\mathrm{x},\mathrm{x}) = c$, then the two following problems are equivalent,*

$$\left\{\begin{array}{cl} \min\limits_{f\in\mathcal{H},R\in\mathbb{R},\xi\in\mathbb{R}^n} & R + C\sum\limits_{i=1}^{n}\xi_i \\ \textit{with} & \|k(\mathrm{x}_i,.) - f(.)\|^2_{\mathcal{H}} \leq R+\xi_i \\ & \xi_i \geq 0 \qquad i = 1,n \end{array}\right. \qquad \left\{\begin{array}{cl} \min\limits_{f\in\mathcal{H},\rho\in\mathbb{R},\xi\in\mathbb{R}^n} & \frac{1}{2}\|f\|^2_{\mathcal{H}} - \rho + C\sum\limits_{i=1}^{n}\varepsilon_i \\ \textit{with} & f(\mathrm{x}_i) \geq \rho - \varepsilon_i \\ & \varepsilon_i \geq 0 \qquad i = 1,n \end{array}\right.$$

with $\rho = \frac{1}{2}(c + \|f\|^2_{\mathcal{H}} - R)$ and $\varepsilon_i = \frac{1}{2}\xi_i$.

The SVDD in $\mathcal{H}$ is the solution of the following quadratic program (QP):

$$\left\{\begin{array}{cl} \min\limits_{f\in\mathcal{H},R\in\mathbb{R},\xi\in\mathbb{R}^n} & R + C\sum\limits_{i=1}^{n}\xi_i \\ \text{with} & \|k(\mathrm{x}_i,.) - f(.)\|^2_{\mathcal{H}} \leq R+\xi_i, \qquad \xi_i \geq 0 \qquad i = 1,n \end{array}\right.$$

since $\|k(\mathrm{x}_i,.) - f(.)\|^2_{\mathcal{H}} = k(\mathrm{x}_i,\mathrm{x}_i) + \|f\|^2_{\mathcal{H}} - 2f(\mathrm{x}_i)$ this QP can be written also

$$\left\{\begin{array}{cl} \min\limits_{f\in\mathcal{H},R\in\mathbb{R},\xi\in\mathbb{R}^n} & R + C\sum\limits_{i=1}^{n}\xi_i \\ \text{with} & 2f(\mathrm{x}_i) \geq k(\mathrm{x}_i,\mathrm{x}_i) + \|f\|^2_{\mathcal{H}} - R-\xi_i, \qquad \xi_i \geq 0 \qquad i = 1,n. \end{array}\right.$$

Introducing $\rho = \frac{1}{2}(c+\|f\|^2_{\mathcal{H}}-R)$ that is $R = c+\|f\|^2_{\mathcal{H}}-2\rho$, and since $k(\mathrm{x}_i,\mathrm{x}_i)$ is constant and equals to c the SVDD problem becomes

$$\left\{\begin{array}{cl} \min\limits_{f\in\mathcal{H},\rho\in\mathbb{R},\xi\in\mathbb{R}^n} & \frac{1}{2}\|f\|^2_{\mathcal{H}} - \rho + \frac{C}{2}\sum\limits_{i=1}^{n}\xi_i \\ \text{with} & f(\mathrm{x}_i) \geq \rho-\frac{1}{2}\xi_i, \qquad \xi_i \geq 0 \qquad i = 1,n \end{array}\right.$$

Date: May 2013.

leading to the classical one class SVM formulation (OCSVM)

$$\left\{\begin{array}{ll} \min\limits_{f\in\mathcal{H},\rho\in\mathbb{R},\xi\in\mathbb{R}^n} & \frac{1}{2}\|f\|^2_{\mathcal{H}} - \rho + C\sum\limits_{i=1}^{n}\varepsilon_i \\ \text{with} & f(\mathrm{x}_i) \geq \rho - \varepsilon_i, \qquad \varepsilon_i \geq 0 \qquad i=1,n \end{array}\right.$$

with $\varepsilon_i = \frac{1}{2}\xi_i$. Note that by putting $\nu = \frac{1}{nC}$ we can get the so called ν formulation of the OCSVM

$$\left\{\begin{array}{ll} \min\limits_{f'\in\mathcal{H},\rho'\in\mathbb{R},\xi'\in\mathbb{R}^n} & \frac{1}{2}\|f'\|^2_{\mathcal{H}} - n\nu\rho' + \sum\limits_{i=1}^{n}\xi'_i \\ \text{with} & f'(\mathrm{x}_i) \geq \rho' - \xi'_i, \qquad \xi'_i \geq 0 \qquad i=1,n \end{array}\right.$$

with $f' = Cf$, $\rho' = C\rho$, and $\xi' = C\xi$.

2. A remark about the duality and the way R is computed

Note that the dual of the SVDD is

$$\left\{\begin{array}{ll} \min\limits_{\alpha\in\mathbb{R}^n} & \alpha^\top G\alpha - \alpha^\top g \\ \text{with} & \sum_{i=1}^{n}\alpha_i = 1 \quad 0 \leq \alpha_i \leq C \qquad i=1,n \end{array}\right.$$

where G is the kernel matrix of general term $G_{i,j} = k(\mathrm{x}_i, \mathrm{x}_j)$ and g the diagonal vector such that $g_i = k(\mathrm{x}_i, \mathrm{x}_i) = c$. The dual of the OCSVM is the following equivalent QP

$$\left\{\begin{array}{ll} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha \\ \text{with} & \sum_{i=1}^{n}\alpha_i = 1 \quad 0 \leq \alpha_i \leq C \qquad i=1,n \end{array}\right.$$

Both dual forms provide the same solution α, but not the same Lagrange multipliers. ρ is the Lagrange multiplier of the equality constraint of the dual of the OCSVM and $R = c + \alpha^\top G\alpha - 2\rho$. Using the SVDD dual, it turns out that $R = \lambda_{eq} + \alpha^\top G\alpha$ where λ_{eq} is the Lagrange multiplier of the equality constraint of the SVDD dual form.

Bibliography

[A.F. Pimentel *et al.* (2014)] A.F. Pimentel, M., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

[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., and Thesen, T. (2015). Cortical feature analysis and machine learning improves detection of "MRI-negative" focal cortical dysplasia. *Epilepsy & behavior : E&B*, 48:21–8.

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

[An and Liang (2013)] An, W. and Liang, M. (2013). Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises. *Neurocomputing*, 110:101–110.

[An and Tao (2005)] An, L. and Tao, P. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46.

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

[Antoniadis *et al.* (2011)] Antoniadis, A., Gijbels, I., and Nikolova, M. (2011). Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63(3):585–615.

[Ashburner (2007)] Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113.

[Ashburner and Friston (2000)] Ashburner, J. and Friston, K. (2000). Voxel-Based Morphometry – The Methods. *NeuroImage*, 11:805–821.

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

[Atrey *et al.* (2010)] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). *Multimodal fusion for multimedia analysis: a survey*, volume 16.

[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. and Kuzniecky, R. (1996). Neuroimaging of focal malformations of cortical development [Review]. *J. Clin. Neurophysiol*, 13:481–494.

[Bartolomei *et al.* (2008)] Bartolomei, F., Chauvel, P., and Wendling, F. (2008). Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG. *Brain : a journal of neurology*, 131:1818–30.

[Bengio *et al.* (2013)] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and ...*, 35(1993):1–30.

[Bernasconi *et al.* (2001)] Bernasconi, A., Antel, S. B., Collins, D. L., Bernasconi, N., *et al.* (2001). Texture analysis and morphological processing of magnetic resonance imaging assist detection of focal cortical dysplasia in extra-temporal partial epilepsy. *Annals of Neurology*, 49(6):770–5.

[Bernasconi and Bernasconi (2015)] Bernasconi, N. and Bernasconi, A. (2015). Clinical and advanced techniques for optimizing MRI in refractory focal epilepsy. In So, E. L. and Ryvlin, P., editors, *MRI-negative epilepsy: evaluation and surgical management*, pages 16–27. Cambridge University Press.

[Besson *et al.* (2008)] Besson, P., Andermann, F., Dubeau, F., and Bernasconi, A. (2008). Small focal cortical dysplasia lesions are located at the bottom of a deep sulcus. *Brain : a journal of neurology*, 131(Pt 12):3246–55.

[Bishop (1995)] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

[Blümcke *et al.* (2011)] Blümcke, I., Thom, M., Aronica, E., , *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. *Epilepsia*, 52(1):158–74.

[Bowman and Azzalini (1997)] Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series)*. Oxford University Press, USA.

[Breiman (2001)] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Breunig *et al.* (2000)] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *Procedings of the ACM SIGMOD International Conference on Management of Data*, volume 29, pages 93–104. ACM.

[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.

[Calhoun and Adali (2009)] Calhoun, V. D. and Adali, T. (2009). Feature-based fusion of medical imaging data. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):711–720.

[Candès *et al.* (2008)] Candès, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted L1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905.

[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.

[Canu *et al.* (2005)a] Canu, S., Grandvalet, Y., Guigue, V., and Rakotomamonjy, A. (2005a). SVM and Kernel Methods Matlab Toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France.

[Canu *et al.* (2005)b] Canu, S., Grandvalet, Y., Guigue, V., and Rakotomamonjy, A. (2005b). Svm and kernel methods matlab toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France.

[Caruana and Niculescu-Mizil (2006)] Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23th International Conference on Machine Learning*, pages 161–168.

[Chandola *et al.* (2009)] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, 41(3):1–58.

[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., Daumas-Duport, C., Merlet, P., Dulac, O., and Chiron, C. (2010). FDG-PET improves surgical outcome in negative MRI Taylor-type focal cortical dysplasias. *Neurology*, 75(24):2168–2175.

[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):83–9.

[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.

[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. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

[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.

[David *et al.* (2011)] David, O., Blauwblomme, T., Job, A.-S., Chabardès, S., Hoffmann, D., Minotti, L., and Kahane, P. (2011). Imaging the seizure onset zone with stereo-electroencephalography. *Brain*, 134(10):2898–2911.

[De Carvalho Fonseca *et al.* (2012)] De Carvalho Fonseca, V., 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. *Front Neurol*, 3.

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

[Duda *et al.* (2012)] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

[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 (1994)] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[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.

[Fauser *et al.* (2004)] Fauser, S., Schulze-Bonhage, A., Honegger, J., Carmona, H., *et al.* (2004). Focal cortical dysplasias: surgical outcome in 67 patients in relation to histological subtypes and dual pathology. *Brain : a journal of neurology*, 127(Pt 11):2406–18.

[Fisher (1936)] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[Fisher and Tippett (1928)] Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press.

[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):786–93.

[Focke *et al.* (2012)] Focke, N. K., Yogarajah, M., Symms, M. R., Gruber, O., *et al.* (2012). Automated MR image classification in temporal lobe epilepsy. *Neuroimage*, 59(1):356–62.

[Forero *et al.* (2012)] Forero, P. A., Kekatos, V., and Giannakis, G. B. (2012). Robust clustering using outlier-sparsity regularization. *IEEE Trans. Signal Process.*, 60(8):4163–4177.

[Franc *et al.* (2011)] Franc, V., Zien, A., and Schölkopf, B. (2011). Support vector machines as probabilistic models. In *ICML 2011*, pages 665–672.

[Friston *et al.* (1999)] Friston, K., Holmes, A., Price, C., BÃijchel, C., and Worsley, K. (1999). Multisubject fMRI Studies and Conjunction Analyses. *Neuroimage*, 10(4):385 – 396.

[Friston *et al.* (2005)] Friston, K. J., Penny, W. D., and Glaser, D. E. (2005). Conjunction revisited. *Neuroimage*, 25(3):661–7.

[Frénay and Verleysen (2014)] Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–69.

[Fukunaga and Hayes (1989)] Fukunaga, K. and Hayes, R. R. (1989). Estimation of classifier performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(10):1087–1101.

[Gao and Tan (2006)] Gao, J. and Tan, P. N. (2006). Converting output scores from outlier detection algorithms into probability estimates. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 212–221.

[Gasso *et al.* (2009)] Gasso, G., Rakotomamonjy, A., and Canu, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Trans. Signal Process.*, 57(12):4686–4698.

[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.

[Glazer *et al.* (2012)] Glazer, A., Lindenbaum, M., and Markovitch, S. (2012). Learning high-density regions for a generalized Kolmogorov-Smirnov test in high-dimensional data. In *Advances in Neural Information Processing Systems*, pages 728–736.

[Glazer *et al.* (2013)] Glazer, A., Lindenbaum, M., and Markovitch, S. (2013). q-OCSVM: A q-quantile estimator for high-dimensional distributions. In *Advances in Neural Information Processing Systems*, pages 503–511.

[Gray *et al.* (2013)] Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., *et al.* (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage*, 65:167–175.

[Gönen and Alpaydin (2011)] Gönen, M. and Alpaydin, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268.

[Hammers (2015)] Hammers, A. (2015). PET in MRI-negative refractory focal epilepsy. In So, E. L. and Ryvlin, P., editors, *MRI-negative epilepsy: evaluation and surgical management*, pages 28–37. Cambridge University Press.

[Hammers *et al.* (2003)] Hammers, A., Allom, R., Koepp, M. J., Free, S. L., *et al.* (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Hum. Brain Mapp.*, 19(4):224–247.

[Hampel (1974)] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

[Hand (2009)] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123.

[Hanley and McNeil (1982)] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

[Haralick *et al.* (1973)] Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6):610–621.

[He and Garcia (2009)] He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

[Hoffmann (2007)] Hoffmann, H. (2007). Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863–874.

[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.

[Huber *et al.* (1964)] Huber, P. J. *et al.* (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

[Huppertz *et al.* (2005)] Huppertz, H.-J., Grimm, C., Fauser, S., Kassubek, J., *et al.* (2005). Enhanced visualization of blurred gray-white matter junctions in focal cortical dysplasia by voxel-based 3D MRI analysis. *Epilepsy Research*, 67(1âĂŞ2):35 – 50.

[Huppertz *et al.* (2009)] Huppertz, H.-J., Kurthen, M., and Kassubek, J. (2009). Voxel-based 3D MRI analysis for the detection of epileptogenic lesions at single subject level. *Epilepsia*, 50(1):155–6.

[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):146–156.

[Jiang *et al.* (1996)] Jiang, Y., Metz, C. E., and Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201(3):745–750.

[Jin *et al.* (2001)] Jin, W., Tung, A. K., and Han, J. (2001). Mining top-n local outliers in large databases. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 293–298. ACM.

[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.

[Keihaninejad *et al.* (2012)] Keihaninejad, S., Heckemann, R. A., Gousias, I. S., Hajnal, J. V., *et al.* (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.

[Kelly and Chung (2011)] Kelly, K. M. and Chung, S. S. (2011). Surgical treatment for refractory epilepsy: review of patient evaluation and surgical options. *Epilepsy research and treatment*, 2011:303624.

[Kim and D. Scott (2012)] Kim, J. and D. Scott, C. (2012). Robust kernel density estimation. *Journal of Machine Learning Research*, pages 2529–2565.

[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*.

[Klöppel *et al.* (2008)] Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., *et al.* (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain : a journal of neurology*, 131(Pt 3):681–9.

[Klein *et al.* (2009)] Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., *et al.* (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802.

[Kotsiantis *et al.* (2007)] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[Kriegel *et al.* (2011)] Kriegel, H., Kröger, P., Schubert, E., and Zimek, A. (2011). Interpreting and Unifying Outlier Scores. *SDM*, pages 13–24.

[Krsek *et al.* (2009)] Krsek, P., Maton, B., Jayakar, P., Dean, P., Korman, B., Rey, G., Dunoyer, C., Pacheco-Jacome, E., Morrison, G., Ragheb, J., Vinters, H. V., Resnick, T., and Duchowny, M. (2009). Incomplete resection of focal cortical dysplasia is the main predictor of poor postsurgical outcome. *Neurology*, 72(3):217–223.

[Lahat *et al.* (2015)] Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal Data Fusion : An Overview of Methods , Challenges , and Prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

[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.

[Lee *et al.* (2007)] Lee, K., Kim, D.-W., Lee, K., and Lee, D. (2007). Density-induced support vector data description. *IEEE Trans. on Neural Netw.*, 18(1):284–289.

[Lee and Scott (2010)] Lee, G. and Scott, C. (2010). Nested support vector machines. *IEEE Transactions on Signal Processing*, 58(3):1648–1660.

[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, J., Fried, I., Cepeda, C., Andre, V. M., Levine, M. S., Miyata, H., Yong, W. H., Vinters, H. V., and Mathern, G. W. (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):1310–35.

[Lichman (2013)] Lichman, M. (2013). UCI Machine Learning Repository.

[Lin *et al.* (2004)] Lin, C.-f. *et al.* (2004). Training algorithms for fuzzy support vector machines with noisy data. *Pattern recognition letters*, 25(14):1647–1656.

[Liu *et al.* (2013)] Liu, B., Xiao, Y., Cao, L., Hao, Z., and Deng, F. (2013). SVDD-based outlier detection on uncertain data. *Knowledge and information systems*, 34(3):597–618.

[Loosli and Canu (2007)] Loosli, G. and Canu, S. (2007). Comments on the core vector machines: Fast SVM training on very large data sets. *J. Mach. Learn. Res.*, 8:291–301.

[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.

[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.

[Mahendran and Vedaldi (2015)] Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5188–5196. IEEE.

[Mairal *et al.* (2009)] Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., and Bach, F. R. (2009). Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040.

[Manly (2006)] Manly, B. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition*. Chapman & Hall texts in statistical science series. Taylor & Francis.

[Markovich (2011)] Markovich, N. (2011). Nonparametric estimation of a heavy-tailed density Specific features of the analysis of heavy-tailed distributions Combined parametric-nonparametric methods. In *58th World Statistical Congress2011, Dublin*, pages 2951–2960.

[Mazziotta *et al.* (2001)] Mazziotta, J., Toga, A., Evans, A., Fox, *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. Series B, Biological sciences*, 356(1412):1293–322.

[McCullagh and Nelder (1989)] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

[McLachlan (2004)] McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons.

[McLachlan and Krishnan (2007)] McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

[Meagher (1980)] Meagher, D. (1980). *Octree Encoding: a New Technique for the Representation, Manipulation and Display of Arbitrary 3-D Objects by Computer*.

[Mourão Miranda *et al.* (2011)] Mourão Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., *et al.* (2011). Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. *Neuroimage*, 58(3):793–804.

[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.

[Mwangi *et al.* (2014)] Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–44.

[Nagae *et al.* (2016)] Nagae, L. M., Lall, N., Dahmoush, H., Nyberg, E., Mirsky, D., Drees, C., and Honce, J. M. (2016). Diagnostic, Treatment and Surgical Imaging in Epilepsy. *Clinical Imaging*.

[Nguyen *et al.* (2010)] Nguyen, H. V., Ang, H. H., and Gopalkrishnan, V. (2010). Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *DASFAA*, volume 5981, pages 368–383, Berlin, Heidelberg.

[Niaf *et al.* (2011)] Niaf, E., Flamary, R., Lartizien, C., and Canu, S. (2011). Handling uncertainties in SVM classification. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, pages 757–760. IEEE.

[Niaf *et al.* (2014)] Niaf, E., Flamary, R., Rouviere, O., Lartizien, C., and Canu, S. (2014). Kernel-based learning from both qualitative and quantitative labels: application to prostate cancer diagnosis based on multiparametric MR imaging. *Image Processing, IEEE Transactions on*, 23(3):979–991.

[Noble (2004)] Noble, W. S. (2004). Support vector machine applications in computational biology. *Kernel Methods in Computational Biology*, pages 71–92.

[Norman *et al.* (2006)] Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–30.

[Orrù *et al.* (2012)] Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., *et al.* (2012). Using Support Vector Machine to identify imaging biomarkers of Neurological and Psychiatric disease: a critical review. *Neuroscience and biobehavioral reviews*, 36(4):1140–52.

[Parzen (1962)] Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

[Petrick *et al.* (2013)] Petrick, N., Sahiner, B., Armato, S. G., Bert, A., *et al.* (2013). Evaluation of computer-aided detection and diagnosis systems. *Medical physics*, 40(8):087001.

[Pickands III (1975)] Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

[Platt *et al.* (1999)] Platt, J. *et al.* (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

[Polikar (2012)] Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer.

[Rakotomamonjy *et al.* (2007)] Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. (2007). More Efficiency in Multiple Kernel Learning. *Proceedings of the 24th international conference on Machine learning*, pages 775–782.

[Rakotomamonjy *et al.* (2016)] Rakotomamonjy, A., Flamary, R., and Gasso, G. (2016). DC Proximal Newton for Nonconvex Optimization Problems. *IEEE Transactions on Neural Networks and Learning Systems*, 27(3):636–647.

[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.

[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.

[Rosenblatt (1962)] Rosenblatt, F. (1962). Principles of neurodynamics.

[Rugg-Gunn *et al.* (2001)] Rugg-Gunn, F. J., Eriksson, S. H., Symms, M. R., and et al. (2001). Diffusion tensor imaging of cryptogenic and acquired partial epilepsies. *Brain*, 124(3):627–636.

[Rumelhart *et al.* (1985)] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, DTIC Document.

[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.

[Sato *et al.* (2012)] Sato, J. R., Rondina, J. M., and Mourão Miranda, J. (2012). Measuring abnormal brains: building normative rules in neuroimaging using one-class support vector machines. *Frontiers in neuroscience*, 6(December):178.

[Schölkopf *et al.* (2001)] Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., *et al.* (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, 13(7):1443–1471.

[Schölkopf *et al.* (1997)] Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In *Artificial Neural NetworksâĂŤICANN'97*, pages 583–588. Springer.

[Scott and Nowak (2006)] Scott, C. D. and Nowak, R. D. (2006). Learning minimum volume sets. *The Journal of Machine Learning Research*, 7:665–704.

[Shawe-Taylor and Cristianini (2004)] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

[Silverman (1986)] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. Chapman and Hall Boca Raton, London, Glasgow, Weinheim. Titre sur le dos du livre : Density estimation.

[Sonnenburg *et al.* (2006)] Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565.

[Sotiris *et al.* (2006)] Sotiris, K., Dimitris, K., and Panayiotis, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36.

[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.

[Stouffer *et al.* (1949)] Stouffer, S., Suchman, E., DeVinney, L., Star, S., and Williams, R. J. (1949). Adjustment during Army Life. *The American Soldier Vol 1, Princeton University Press*.

[Tax and Duin (2004)] Tax, D. M. J. and Duin, R. P. (2004). Support Vector Data Description. *Mach. Learn.*, 54(1):45–66.

[Taylor *et al.* (1971)] Taylor, D., Falconer, M., Bruton, C., and Corsellis, J. (1971). Focal dysplasia of the cerebral cortex in epilepsy. *Journal of Neurol Neurosurg Psychiatry*, 34:369–387.

[Thesen *et al.* (2011)] Thesen, T., Quinn, B. T., Carlson, C., Devinsky, O., *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.

[Thomas *et al.* (2015)] Thomas, A., Feuillard, V., and Gramfort, A. (2015). Calibration of One-Class SVM for MV set estimation.

[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.

[Vapnik (1998)] Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.

[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.

[Vert and Vert (2006)] Vert, R. and Vert, J.-P. (2006). Consistency and convergence rates of one-class SVMs and related algorithms. *The Journal of Machine Learning Research*, 7:817–854.

[Von Oertzen *et al.* (2002)] Von Oertzen, J., Urbach, H., Jungbluth, S., Kurthen, M., Reuber, M., Fernandez, G., and Elger, C. (2002). Standard magnetic resonance imaging is inadequate for patients with refractory focal epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 73(6):643–647.

[Wagner *et al.* (2011)] Wagner, J., Weber, B., Urbach, H., Elger, C. E., *et al.* (2011). Morphometric MRI analysis improves detection of focal cortical dysplasia type II. *Brain*, 134(Pt 10):2844–54.

[Weston *et al.* (2003)] Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461.

[Xiao *et al.* (2014)] Xiao, Y., Wang, H., Zhang, L., and Xu, W. (2014). Two methods of selecting Gaussian kernel parameters for one-class SVM and their application to fault detection. *Knowledge-Based Systems*, 59:75–84.

[Xu and Wunsch (2005)] Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

[Zeiler and Fergus (2014)] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer.

[Gaëlle, Loosli and Aboubacar (2014)] Gaëlle, Loosli and Aboubacar, H. (2014). Using SVDD in SimpleMKL for 3D-Shapes Filtering. In *CAP*, number 2, pages 1–9.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : EL AZAMI
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 23/09/2016

Prénoms : Meriem

TITRE : Computer aided diagnosis of epilepsy lesions based on multivariate and multimodality data analysis

NATURE : Doctorat

Numéro d'ordre :

Ecole doctorale : Electronique, Electrotechnique, Automatique (EEA)

Spécialité : Traitement du signal et de l'image

RESUME :

One third of patients suffering from epilepsy are resistant to medication. For these patients, surgical removal of the epileptogenic zone offers the possibility of a cure. Surgery success relies heavily on the accurate localization of the epileptogenic zone. The analysis of neuroimaging data such as magnetic resonance imaging (MRI) and positron emission tomography (PET) is increasingly used in the pre-surgical work-up of patients and may offer an alternative to the invasive reference of Stereo-electro-encephalo -graphy (SEEG) monitoring. To assist clinicians in screening these lesions, we developed a computer aided diagnosis system (CAD) based on a multivariate data analysis approach. Our first contribution was to formulate the problem of epileptogenic lesion detection as an outlier detection problem. The main motivation for this formulation was to avoid the dependence on labelled data and the class imbalance inherent to this detection task. The proposed system builds upon the one class support vector machines (OC-SVM) classifier. OC-SVM was trained using features extracted from MRI scans of healthy control subjects, allowing a voxelwise assessment of the deviation of a test subject pattern from the learned patterns. System performance was evaluated using realistic simulations of challenging detection tasks as well as clinical data of patients with intractable epilepsy.
The outlier detection framework was further extended to take into account the specificities of neuroimaging data and the detection task at hand. We first proposed a reformulation of the support vector data description (SVDD) method to deal with the presence of uncertain observations in the training data. Second, to handle the multi-parametric nature of neuroimaging data, we proposed an optimal fusion approach for combining multiple base one-class classifiers. Finally, to help with score interpretation, threshold selection and score combination, we proposed to transform the score outputs of the outlier detection algorithm into well calibrated probabilities.

MOTS-CLÉS :

Outlier detection, Pattern recognition, OC-SVM, SVDD, Computer aided diagnosis, Neuroimaging, MRI, Epilepsy

Laboratoire (s) de recherche : Centre de recherche en acquisition et traitement de l'image pour la santé

Directeur de thèse : Denis Friboulet

Président de jury : Stéphane Canu

Composition du jury :

Rakotomamonjy, Alain	Professeur des universités	Université de Rouen	Rapporteur
Rueckert, Daniel	Professeur des universités	Imperial College London	Rapporteur
Canu, Stéphane	Professeur des universités	INSA-ROUEN	Examinateur
Hammers, Alexander	Professeur des universités	King's College London	Examinateur
Friboulet, Denis	Professeur des universités	INSA-LYON	Directeur de thèse
Lartizien, Carole	Chargé de recherche	CNRS	Co-directrice de thèse