Unsupervised anomaly detection in neuroimaging: contributions to representation learning and density support estimation in the latent space

Nicolas Pinon

N°d'ordre NNT : 2024ISAL0031

THESE de DOCTORAT DE L'INSA LYON, membre de l'Université de Lyon

Ecole Doctorale N° 160
Électronique, Électrotechnique, Automatique

Discipline de doctorat :
Traitement du signal et de l'image

Soutenue publiquement le 11/04/2024, par :
Nicolas Pinon

Unsupervised anomaly detection in neuroimaging: contributions to representation learning and density support estimation in the latent space

Devant le jury composé de :

Schnabel, Julia	*Professor* Technical University of Munich	Rapportrice
Bloch, Isabelle	*Professeure des universités* Sorbonne Université	Rapportrice
Meriaudeau, Fabrice	*Professeur des universités* Université de Bourgogne	Examinateur, Président du jury
Duchateau, Nicolas	*Maitre de conférences HDR* Université Lyon 1	Examinateur
Lartizien, Carole	*Directrice de recherche* CNRS	Directrice de thèse
Forbes, Florence	*Directrice de recherche* INRIA	Invitée

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	**CHIMIE DE LYON** https://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
ED 341 E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** http://e2m2.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX e2m2.codir@listes.univ-lyon1.fr
ED 205 EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** http://ediss.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
ED 34 EDML	**MATÉRIAUX DE LYON** http://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
ED 160 EEA	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** https://edeea.universite-lyon.fr Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
ED 512 INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 direction.infomaths@listes.univ-lyon1.fr
ED 162 MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** http://edmega.universite-lyon.fr Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Etienne PARIZET** INSA Lyon Laboratoire LVA Bâtiment St. Exupéry 25 bis av. Jean Capelle 69621 Villeurbanne CEDEX etienne.parizet@insa-lyon.fr
ED 483 ScSo	**ScSo**[1] https://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 bruno.milly@univ-lyon2.fr

[1] ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

"*Si on y réfléchit bien, et sans passion, une chambre comme le Sénat [...] c'est une* ***anomalie*** *parmi les démocraties.*"

– Lionel Jospin

Remerciements

Ça fait un peu mec qui se la raconte de faire des remerciements je trouve, mais bon il faut ce qu'il faut alors c'est parti. Il y a beauuuuuucoup de personnes à remercier et cela va être difficile de les mettre dans des cases, mais je vais essayer de faire une espèce de gradient des remerciements scientifiques aux remerciements plus personnels, même si plein de gens sont entre les deux, et surtout les deux en même temps, ainsi qu'un gradient temporel je crois, je ne sais pas. Lisez juste en fait.

Merci à Isabelle Bloch et Julia Schnabel d'avoir accepté de rapporter ma thèse, c'est un grand honneur d'avoir deux personnes que je respecte beaucoup scientifiquement en tant que rapportrices ! Merci à Fabrice Meriaudeau et Florence Forbes d'avoir complété ce super jury, pour les mêmes raisons. Nicolas, merci d'avoir accepté d'examiner, et pour tous les petits conseils et astuces de pro sur la 61 entre cafés et sciures de compost en salle de pause.

Merci à Guillaume et Robin d'avoir rendu cette thèse un peu plus fournie. Loin de ce qu'on pourrait penser, travailler avec des stagiaires ne fait pas gagner du temps mais ajoute à la montagne de choses à peut-être devoir ajouter à sa thèse. Robin je te souhaite la plus grande réussite scientifique, entre deux-trois lexus, mais j'ai pas trop de doutes en fait.

Enfin merci à la patronne, merci à Carole d'avoir encadré cette thèse, pendant un peu plus de 3 ans et 4 mois. Le dosage entre autonomie (good cop) et cadrage (bad cop) était parfaitement à ma convenance (c'est à dire 90% d'autonomie). Désolé d'avoir envoyé boulé le jeu de donnée épilepsie et de pas avoir trop cru aux cluster maps. Merci pour les conseils toujours pertinents, la grande flexibilité car j'étais souvent en retard (mais qui se ressemble s'assemble) et tout ce que j'ai oublié.

Une petite pensée pour le service admin (Anthony, Marie, Béatrice, Fabienne, Sophie Trift et Georges), Pierre Ferrier (le sang), et tout ceux qui font tourner ce labo plutôt chouette. Merci à Louise pour la charte couleur de la thèse, le template, les conseils esthétiques de manière générale, qui ont contribué à rendre cette thèse plus belle, même si je ne les ai pas tous suivi (oups). Sans oublier les discussions technico-techniques sur les barres du mouvement inter et une aptitude à tirer la gueule encore inégalée.

Enfin, que serait une thèse en traitement d'image sans une infrastructure nationale de calcul digne de ce nom ? Un véritable service public informatique, j'ai nommé : le supercalculateur Jean Zay. Véritable tête de pont de la puissance de calcul du CNRS, il se situe dans la droite ligne des grands équipements que financent le CNRS, tel qu'imaginé par Jean Perrin en 1939. Merci Jean Zay, vive la République, et comme disait Léon Blum : "*This work was granted access to the HPC resources of IDRIS under the allocations 2021-AD011012813, 2022-AD011012813R1, and 2023-AD011012813R2 made by GENCI.*".

Un peu d'histoire maintenant... Mes premiers contacts avec CREATIS, mine de rien, c'était en Avril 2020, en des temps un peu troublés de pandémie. Et qui de mieux, pour s'intégrer à un laboratoire en des temps agités que Ludmilla, Manu et Louise ? Tout d'abord au travers d'une fenêtre zoom, puis bien rapidement à une crémaillère (où j'ai pu expérimenté la plus grande

concentration de profs de maths de ma vie), vous avez été absolument parfait et je souhaite à tout le monde une intégration aussi rapide.

Rien à voir avec CREATIS, mais même thème : quoi de mieux quand le Covid et autre Castex ferment tous les bars, imposent des couvre-feux, confinement numéro 5, Vitemadose et tutti quanti, quoi de mieux qu'une super coloc ? Félix et Hugo, ainsi que la coloc extended universe : Clément, Joseph, Lucie, le voisin qui produit un album, RIBES, les composteurs très stricts, Dominique Pinon, Dutour, les techos de la fibre. Merci pour tous ces listo, riste, vaisseau, les flammes de l'enfer, la caution, la faille, le 2e canapé d'angle dont vous n'avez pas voulu, le sapin qu'on a jamais jetté, les $40m^3$ de verre, le faux gazon à $24€/m^2$, la chasse d'eau dragon, les duoQ (pour manger), daroneries, cachio i pepe, la monstera (RIP), "UN HP !!!", le colis de SCARA, et surtout ce superbe $76m^2$ au 16^e étage avec triple balcon et vue sur Mont Blanc. Merci Jerry, Alex et Clover, merci les Boys.

Après BP (ceux qui savent, savent), fut le temps du bureau 13, mais je ne vous parle pas du bureau 13 des jeunes, je vous parle des sexagénaires du bureau 13 : Charles, Nico Loiseau, Florian, Ludmilla (encore !), Zexian et Théo (l'expert). Merci les potes pour ce super bureau, de loin le meilleur bureau, sans débat. Plus généralement merci aux anciens pour leurs précieux conseils, pour l'ambiance, et des délires un peu chelous à base de big wipes et de goûter Prince : Max Dif, Antonio, Frank, Audrey (carte dino!), Mom, Nin, Louise (celle qui fait des belles figures encore oui), PJ, Anne-Lise, Séb, Daria, Benoît, Valentine, Suzanne, Noëlie.

Entre la pré-histoire que je viens de décrire et les temps modernes, s'est passé une espèce de période de transition, à base de plan 4 étoiles, flipper, Zinc-ries, blindtests et Chamas tacos extrêmement sombre, où mes acolytes furent Ludmilla (encore!!), Gaspard (récupère le flipper de ta grand-mère) et Pierre (without *réserve*). Pendant cette période de transition j'ai pu aussi partager le bureau 13 avec Julia (et oui tous les autres rédigeaient), merci pour tous ces moments et discussions 13 intéressantes sur l'Espagne, la Catalogne, la politique, Lille, Grenoble et tant d'autres.

Maintenant arrivent les temps modernes, et surtout mes compatriotes de thèse, la génération 2020, dont je suis un des 3 retardataires, avec mes deux bonhommes Sophie et Nathan. Nathan tu es l'enfant prodige du labo, le mariage parfait entre le Québec (solidaire) et la France, Charles de Gaulle n'aurait pas rêvé mieux, je vais essayer de rattraper tes 3 TMI mais je suis pas très bien parti. Sophie merci pour tous ces petits (nombreux, très nombreux) moments, à la fin on est toustes docteurs, j'espère que d'autres pourront glowup comme tu l'as fait en 4^e année, aussi bien sur le plan personnel que scientifique. Il faudra aussi penser à quitter tATER natale un de ces jours (pour aller à Nantes par exemple).

J'ai aussi pu profiter de deux comparses pour m'accompagner dans ce dur labeur que fut la rédaction, les deux plus gros charbonneurs du labo, la team 9to9 : Shusong et Cyril ! Merci pour les petites pauses de 21h (9pm break boss), grâce à vous je n'étais pas un travailleur isolé et j'ai pu respecter le réglement intérieur de CREATIS. Now it's your turn BIG BIG BOSS, grand patron MRC ! Bravo Shusong, maintenant t'es Docteur. Que de moments de joies partagés, une pure amitié franco-chinoise, je crois que j'ai jamais autant rigolé que quand t'as confondu Jésus avec Jean Zay, merci le sang ! T'inquiète boss.

Enfin pour clore cette génération dorée de CREATIS, mes deux iraniens préférés, Mehdi et Samaneh, vous êtes le sang de la veine, merci pour les discussions sur la laïcité et les mirza ghassemi.

On est enfin rendu à l'époque contemporaine, le CREATIS des jeunes, la relève, et notamment : le bureau 13 v2, golden edition. Quel plaisir d'avoir pu renouveler ce bureau avec une team de gens 100% cools, pas prise de tête, sympas, fin toutes les qualités possibles quoi. Merci à Enyi, Valentin, Romain, Émile et le petit dernier Thibaut. Ça va être dur de quitter ce bureau mais heureusement vous allez m'avoir dans les pattes encore un peu !

Et alors même hors bureau 13, quelle génération, une team ambiance absolument hors du commun, le CREATIS de la grande époque : Marta, Flora, Nolann, Hang-Jung, Antoine médecin, Célestine, Morgane (graine de tournesol), JB, Juliette (merci d'avoir été ma partenaire de trek !), Luis, Fernand, Gaël post-doc, Thomas (jte paie un moscow mule le 11 au soir), Serena (mystérieuse), Celia x2, Pierre-Eliott, Thierry (déjà sur le 2e TMI...), mini Thierry (n'a pas vu les buildings), Romain ingé. Maylis je compte sur toi pour instaurer la dictature du prolétariat au CU. À tous les stagiaires du CREATIS, à peine arrivés que vous partez et nous manquez déjà, dédicace en particulier à Slim, Pablo, Serge et Lucas Braz. Fabien, Odyssée, Sébastien, Duchat', Jeremy, Michaël, Fred, Hervé, Sorina, David, Eric vous êtes au top, des sortes de role-model de l'ESR.

Parlons un peu de la team L: on partage malheureusement tous les 4 une organisation travail/sommeil absolument catastrophique, svp on améliore ça au plus vite. Matthis, félicitations pour avoir obtenu le prix d'Oncologie le plus prestigieux de France, on est toustes fier de toi. Robin, tu es l'enfant prodige du labo à la française (c'est à dire sans les 3 TMI), en plus d'avoir bien aimé travaillé avec toi, il se trouve que t'es plutôt sympa donc continue d'ambiancer tout le monde, c'est un travail que peu font mais qui est absolument essentiel à la survie de l'ambiance absolument légendaire de CREATIS. Prends soin de ta santé des fois quand même.

Pierre, à la manière de la seconde République, tu es une espèce de courroie de transmission entre les temps modernes et l'époque contemporaine. Du Zinc à LFI, en passant par Mister Gucci, le personnage de Calimero cinéphile aigri sappé comme un étudiant en architecture te sied à ravir, François Bégaudeau, Usul, Edwy Plenel, Dany et Raz seraient fiers de toi. J'ai aucun doutes sur le fait qu'on se verra à IABM 2035. À bientôt, au coeur de la recherche française.

Et en parlant d'IABM, comment ne pas mentionner les deux plus gros ambianceurs du traitement d'image ? Rox et Rouky, Louis de Funès et Bourville, Bob l'éponge et Patrick, les deux strasbourgeois, les crackitos de la recherche FR : Alex et Luc. Franchement les gars j'étais pas prêt pour autant de dopamine en plein milieu de ma smart 3A, à base d'IABM, ISBI, de miradors, de jugo de lulo, des vendeurs de "mangue", plaza trinidad, Villani, Tamil annuel et autres grenobleries. Je vous souhaite le meilleur, tryhardez pas trop svp ça sert à rien, de toute façon la smart 3A n'est que le tremplin de la très smart 4A.

Et puis tant qu'on est à Strasbourg... Une petite dédicace aux francos (Pierre, Dimy, JDG, Myriam, MaxP, Arthur, Joseph, Josh, Camille, Julie, Maurus, Moritz & co), une sorte de famille d'accueil mais allemande, au compost (giga comeback dès le 12 Avril, Pierre tous les jours HOTS), à la team nouvel an (Élise, Théo, Thomas, Yannis & co) et aux frères du DUT MP: Félix, Hugo, Clément (Jesteeeeer), Joseph (finis ta thèse), Guillaume, Léo (RIP), Camille, Rhipsimée, Vicat, Trappeur, Sissé, Stéphane, Yin et Milad. Merci à la BNU de m'avoir emmené très loin, à Porret's de m'avoir réconcilié avec les maths, et aux profs iconiques de matières dank de type DENU MANU ELANA RPROT & co.

Après Strasbourg mais avant Lyon, j'ai pu aussi rencontrer des gens très chouettes, entre Cachan et Montréal : dédicace aux Van Hornies, à Guillaume Mestdagh et à mes PVTistes préfs

(Mariane, Fred, Floriane, Antoine et les autres, sans oublier Louis !), j'ai pas travaillé énormément mais j'ai beaucoup beaucoup rigolé pendant cette année. Ça c'était Montréal, maintenant à Cachan, ne tournons pas autour du pot, éééééénorme dédicace à ma fanfare d'amour : la F[ENS]foire ! Je vais sûrement en oublier mais gros big-up aux anciens (Alice, Miko, Émilie, Tim) aux moyens (Rémi, Léa, Karim, le Kbanon, Karlos, Émile, Martin, Victor, Soubaboy, Léo, Elio, Lucie!), aux jeunes (Sylvain, Vivien, Louise, Pilou, Nico jeune, Quentin, Raphi), à François Hollande, à Chameyras, aux Angles et à tous ceux que j'ai oublié. J'ai de si beaux souvenirs de toutes ces fêtes, manches, festivals, beuveries, répèts, solidays, peña et tout le reste ! Pull-up à mon compagnon d'EEA RBK et une très spéciale décicace à mon conteur d'histoire, musicien hors-pair, femme-biche, embrouille-man, huma-niste, entière-mesureur, mon compagnon de route, mon frère : Lulu.

Enfin, il est essentiel pour un doctorant d'être bien nourri pour être efficace, et pour cela j'aimerais chaleureusement remercier le Pied du Saule de m'avoir nourri pendant 3 ans. Souvent critiqué mais finalement jamais égalé, le PDS sait fournir une atmosphère *conviviale*, *chaleureuse* et *authentique*. Proposant systématiquement une option végétarienne, un repas de Noël hors du commun, des entrées chaudes à faire tourner les têtes, une solidarité avec les grévistes sans faille et depuis peu du yaourt de petits producteurs. Merci du fond du coeur à toute l'équipe du PDS. Enfin un petit mot pour le concurrent capitaliste du PDS : merci au Domus de m'avoir copieusement et qualitativement nourri en cette fin de thèse, malgré le manque flagrant d'ambiance de cet endroit et de gentillesse du personnel caissier.

Comme ceux qui lisent ces lignes le savent sûrement, je suis un peu écolo, j'aime un peu la politique, alors quoi de mieux pour concilier ces deux choses qu'on bon groupe de ZADistes au laboratoire ? Vous l'avez devinez je parle bien évidemment de LIFE : votre collectif préféré de chercheureuses radicalisés ! Merci à toute la team, le LIFE étendu comme le noyau dur : Laura, Axel A, Antoine, Axel B, JB, Maylis, Morgane, Max Dif, Madani, Pilar, Hervé, Fred et nos soutiens sans faille Sébastien, Magalie, Barbara, Patrick et plus récemment Olivier Beuf ! Ça parait pas grand chose mais je suis super fier de ce qu'on a accompli, c'est pas si facile d'aller au contact avec les boomers qui veulent rien changer, c'est pas si facile de prendre le temps de convaincre, de se positionner, de discuter, avec toutes et tous. Longue vie à LIFE.

Enfin, une pensée à ma famille, tout particulièrement à ma mère et ma soeur (et Umi !). Je vous aime. Merci à Lucie d'avoir brunché avec moi pendant ces 3 ans et plus (et...). Je t'aime.

Merci merci merci, la science c'est sympa mais il n'y a rien que j'aime plus au monde que de tous vous voir, vous parler, rigoler avec vous.

Nico/Raoul

Contents

Résumé étendu

Introduction

L'imagerie médicale joue un rôle prépondérant dans le diagnostic médical contemporain. L'émergence de nouvelles modalités d'imagerie, ainsi que des systèmes d'imagerie hybrides (TEP/TDM, TEP/IRM) permettant l'acquisition simultanée ou séquentielle de différentes informations anatomiques, fonctionnelles ou moléculaires, conduit à la génération de flux considérables de données. Cette masse de données multimodales et multiparamétriques constitue une source riche d'informations pour le diagnostic et la compréhension des pathologies, mais elles sont difficiles à exploiter par une simple analyse visuelle.

Les méthodes d'analyse statistique, en particulier celles impliquant l'apprentissage automatique, peuvent répondre à ce besoin. Le domaine de la recherche sur l'apprentissage automatique pour l'analyse des images médicales est très actif, en particulier pour l'apprentissage supervisé, qui nécessite l'accès à de grandes bases de données structurées. Cependant, l'accès pratique aux données médicales est limité car il implique un processus d'annotation intensif en main-d'œuvre, une tâche chronophage qui ne peut être effectuée que par un expert clinique.

Dans ce contexte, l'apprentissage automatique non supervisé, qui ne nécessite pas d'étiquettes, et en particulier la détection non supervisée d'anomalies (UAD : *unsupervised anomaly detection*), semble particulièrement adapté à la tâche de traiter de telles bases de données non structurées à grande échelle. La détection non supervisée d'anomalies, ou simplement la détection d'anomalies, vise à modéliser statistiquement la normalité observée dans les données, dans le but de trouver des anomalies déviant de la distribution normale[1] modélisée. Ainsi, les méthodes UAD ne nécessitent que des images de contrôle saines pour être mises en œuvre et permettent la détection de tout type de pathologie ou d'anomalie sans caractérisation préalable. Cela peut être particulièrement utile lorsque la pathologie étudiée est rare, ou lorsque les annotations sont difficiles à obtenir.

Ce travail tente de contribuer aux méthodes de détection d'anomalies utilisées en neuroimagerie, en particulier à l'estimation du support de densité dans l'espace latent des réseaux neuronaux profonds. Ce cadre général, introduit pour la première fois dans Alaverdyan et al. (2020), comprend une étape d'apprentissage de la représentation réalisée par modélisation profonde non supervisée, suivie de la détection d'anomalies avec estimation du support de densité. Cette étude novatrice s'est révélée puissante pour la détection des lésions épileptogènes, sur une base de données privée d'IRM multiparamétrique.

[1] Au sens des échantillons normaux, opposés aux anomalies, et non au sens de la distribution gaussienne, aussi appelée loi normale.

L'objectif principal de ce travail est de proposer des contributions méthodologiques pour améliorer soit l'apprentissage de la représentation, soit les étapes de détection des anomalies du modèle UAD, tout en évaluant les performances sur diverses bases de données publiques.

Notre première contribution consiste à évaluer ce modèle novateur sur trois bases de données publiques. Tout d'abord, nous utilisons une base de données d'images industrielle (MVTecAD Bergmann et al. 2021), qui présente une vérité terrain précise et divers types d'anomalies. Ensuite, nous exploitons une base de données d'imagerie médicale ouverte qui contient des petites lésions cérébrales hyperintenses en IRM T1 et FLAIR, connues sous le nom d'hyperintensités de la substance blanche (WMH), avec une vérité terrain précise Kuijf et al. (2019). Troisièmement, nous utilisons la base de données PPMI Marek et al. (2018), qui comprend des acquisitions IRM de controles et de patients Parkinson *de novo*, avec une vérité terrain au niveau de l'image. En évaluant le modèle sur ces bases de données ouvertes, nous facilitons les comparaisons avec la littérature existante et fournissons une évaluation plus précise des forces et des faiblesses de l'approche, grâce à la disponibilité des étiquettes.

Notre deuxième contribution est l'introduction d'une nouvelle stratégie pour l'entraînement des machines à vecteurs de support (SVM : *support vector machine*) uniclasse pour résoudre des problèmes liés à la dépendance au jeu de données d'entraînement (taille et caractéristiques extraites), à la sensibilité partielle au recalage spatial et au processus d'optimisation long du modèle novateur. Cela vise à améliorer la reproductibilité, la sensibilité et la spécificité. De plus, nous explorons des méthodes de conversion des scores d'anomalie (non bornés) en probabilités, permettant la construction de modèles d'ensemble et facilitant la calibration des cartes de scores. Ces méthodes contribuent à améliorer la capacité du modèle à combiner des données hétérogènes.

Notre troisième contribution est axée sur l'amélioration de la représentation de l'espace latent apprise par la modélisation profonde non supervisée, pour améliorer la sensibilité dans une tâche de détection plus difficile. Nous explorons des méthodes classiques pour structurer l'espace latent, telles que l'incorporation d'une régularisation variationnelle ou l'encodage de position. Nous introduisons ensuite un nouveau modèle bout à bout qui couple l'étape d'extraction des caractéristiques avec l'étape de détection des anomalies. Des expériences supplémentaires sont menées pour évaluer les performances des modèles et l'organisation de l'espace latent, en particulier dans le contexte de la détection de lésions subtiles.

Ce travail est divisé en cinq chapitres. Le premier chapitre introduit le concept de détection d'anomalies, y compris les métriques courantes et les bases de données utilisées pour cette tâche, en particulier dans l'imagerie médicale. Ensuite, nous passons en revue les méthodes de l'état de l'art utilisées dans la détection non supervisée d'anomalies pour l'imagerie médicale. Dans le deuxième chapitre, nous visons à formaliser le problème étudié et à établir les limites de l'étude d'Alaverdyan et al. (2020), afin d'introduire nos contributions. La thèse se conclut par un chapitre sur les conclusions et les perspectives.

Chapitre 1 : Détection d'anomalie non supervisée

Dans ce premier chapitre, nous commençons par donner une définition intuitive puis formelle de la tâche de détection d'anomalie (UAD : *unsupervised anomaly detection*. Ensuite, nous motivons le besoin de développer des méthodes de détection d'anomalie, en particulier pour l'imagerie médicale.

Nous reprenons ensuite la classification de Ruff et al. (2021) et présentons les trois grandes familles de méthodes d'UAD: les méthodes par reconstruction, estmation de densité et estimation de support. Ces deux dernières sont parfois appelées : méthode génératives et méthodes discriminatives.

Nous présetons figure 1 un schéma visuel des trois types de méthode.

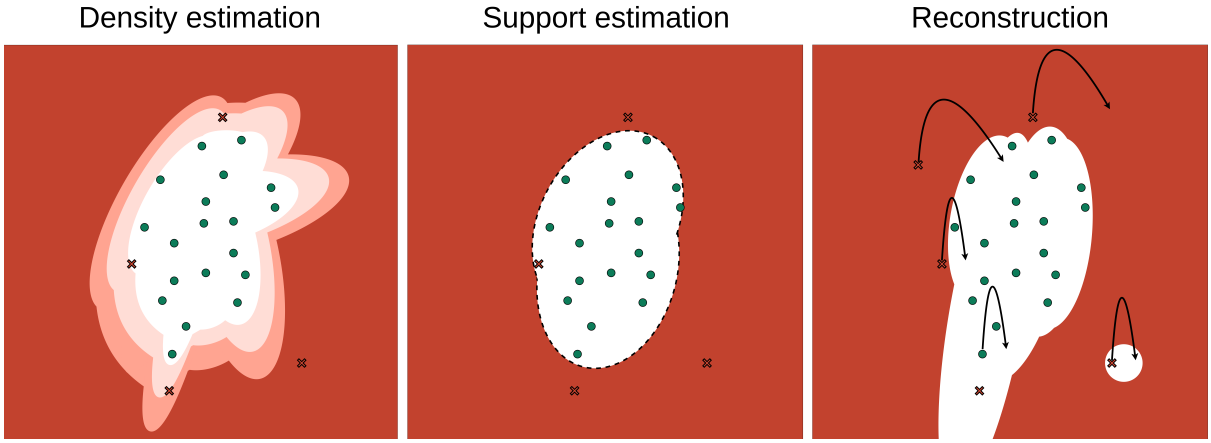

Figure 1: Représentation schématique des trois familles de méthodes de détection d'anomalie: l'estimation de densité (*density estimation*), l'estimation de support (*support estimation*) et la reconstruction (*reconstruction*). Inspiré de Ruff et al. (2021), les zones rouges indiques les zones détectés comme anormales. Les méthodes d'estimation de densité peuvent sur-apprendre ou sous-apprendre les extrémités de la distribution, ou simplifier excessivement la frontière. Les méthodes d'estimation de support peuvent créer des frontières soit trop lâches soit trop restreintes. Les méthodes de reconstruction peuvent générer des zones d'artefacts où les anomalies sont bien reconstruites.

Nous continuons ce chapitre en donnant une description précise des métriques classiquement utilisées en UAD, tel que l'*AU ROC*, *AU PR* ou encore l'*AU PRO*, et leurs versions à 30% de faux positifs.

Nous présentons ensuite les bases de données classiques utilisées en détection d'anomalie, notamment la base MVTecAD que nous utilisons au chapitre 3, et plusieurs bases de données spécifiques à l'imagerie médicale.

Ensuite, nous présentons les grandes méthodes de détection d'anomalie, que nous appelons "méthods fondamentales", qui serviront de base aux autres méthodes décrites ensuite. Encore une fois, la classification de Ruff et al. (2021) est utilisée. Ces méthodes incluent, pour les méthodes de reconstruction: les auto-encodeurs, U-net, auto-encodeurs variationels, quantifiés, les réseaux adversaires génératifs (GAN), les modèles de débruitages comme les modèles de diffusion. Pour les méthodes d'estimation de densité : les modèles de mélange, d'estimation de densité par noyau, ou encore les *normalizing flows*. Pour les méthodes d'estimation de support : SVM uniclasse, *support vector data description* (SVDD) et SVDD profond. un exemple de méthode hybride est l'utilisation de VAE quantifiés + modèles auto-regressifs.

Nous présentons ensuite un état de l'art partiel des méthodes les plus performantes pour l'UAD sur MVTecAD, notamment FastFlow (Yu et al. 2021) et PaDiM Defard et al. (2021), que nous réutilisons plus tard dans le chapitre 3.

Nous dressons ensuite un état de l'art quasi-exhaustif des méthodes d'UAD pour l'imagerie médicale. Le besoin de méthodes d'UAD spécifique à l'imagerie médicale est motivé dans le chapitre 2. Ces méthodes sont présentées en faisant appel aux méthodes fondamentales présentées plus tôt. Un bilan des méthodes est présentée tableau I.1, où chaque méthode est présentée succintement, ainsi que les caractéristiques utilisées (*features*), base de données, type de méthode, métriques reportées et type d'évaluation.

Chapitre 2 : Formulation du problème

Dans ce deuxième chapitre, nous exposons les défis relatifs à l'UAD en imagerie médicale. Premièrement, les défis liés aux bases de données et à l'évaluation, qui sont : la difficulté de la tâche, les détecteurs d'hyperintensité, l'absence de vérité terrain exacte, les anomalies à type unique, le *domain shift* et les capacités de généralisation. Deuxièmement, les cifficultés liées aux méthodes, chaque méthode ayant ses difficultés propres : interpretabilité pour les méthodes de reconstruction, complexité pour les méthodes d'estimation de densité, et faible étude des méthodes d'estimation de support.

Nous entrons ensuite dans le détail de la méthode proposée par Alaverdyan et al. (2020), notamment la première étape d'apprentissage de représentation par auto-encodeur siamois par patch, et la deuxième étape de détection d'anomalie à proprement parler, par SVM uniclasse. La figure 2 résumme cette méthode. Nous exposons ensuite quelques comparaisons avec la littérature.

Nous présentons ensuite les faiblesses du l'étude, en commençant par les limites liées à la méthode. Cette méthode utilise un grand nombre de machines à vecteurs de support (SVM) à classe unique, environ 1,5 million, chacune formée indépendamment sur une seule coordonnée de voxel en utilisant tous les contrôles disponibles. Cependant, cette approche présente plusieurs limitations. Tout d'abord, l'optimisation est très longue en raison du grand nombre de modèles. Deuxièmement, elle repose sur l'assumption de l'alignement parfait des sujets, ce qui rend la méthode sensible aux problèmes de registration non linéaire. Troisièmement, le nombre d'échantillons pour former un modèle SVM est limité au nombre de contrôles, contrairement à d'autres méthodes qui utilisent de nombreux échantillons de différents contrôles.

Étant donné que le modèle est formé pour détecter les déviations de la normalité à une échelle très locale, toute déviation, qu'elle soit pathologique ou bénigne (par exemple, la variabilité anatomique), est détectée comme une anomalie, rendant le modèle intrinsèquement peu spécifique. De plus, les expériences montrent une faible reproductibilité des résultats en raison de la stochasticité de l'optimisation du réseau siamois, ce qui peut influencer considérablement le reste du processus.

L'entrainement du modèle dépend de la base de données d'entraînement, sans *fine-tuning* sur le patient ni intégration de caractéristiques spécifiques au patient, ce qui peut entraîner une sensibilité au changement de domaine. Enfin, l'extraction des caractéristiques et la détection des anomalies sont effectuées en deux étapes distinctes, sans garantie que les caractéristiques extraites seront pertinentes pour la détection des anomalies.

Les faiblesses de la méthode sont résumées comme suit :

- Sensibilité à l'enregistrement en raison du modèle par voxel

- Sensibilité à la taille de l'ensemble de données d'entraînement (nombre de contrôles sains)
- Absence de caractéristiques spécifiques au patient
- Optimisation longue (1,5 million de SVM)
- Spécificité intrinsèquement faible
- Faible reproductibilité
- Extraction des caractéristiques dissociée de la détection des anomalies

Nous présentons ensuite les limites de cette étude liées à l'évaluation. Cette méthode utilise des données privées pour l'ajustement et l'évaluation des modèles, ce qui rend impossible la reproduction des résultats et la comparaison avec d'autres méthodes. La taille de la base de données, bien que raisonnable pour l'imagerie médicale, reste petite par rapport au nombre habituel d'échantillons utilisés en apprentissage automatique, ce qui rend difficile l'attribution d'une signification statistique solide aux résultats établis. Pour la base de données d'évaluation, il est important de noter que la constitution d'une grande base de données publique sur l'épilepsie est très difficile, principalement en raison de la difficulté d'obtenir une vérité terrain précise, car cette maladie et ses causes ne sont pas encore bien comprises. À notre connaissance, aucune base de données de ce type n'existe. Le faible nombre de contrôles sains, étant de même ordre de grandeur que la dimension de l'espace où le SVM est ajusté, remet en question la robustesse statistique des résultats. La taille de la base de données de contrôles remet également en question la capacité du modèle à capturer la variabilité anatomique de la population en bonne santé. De plus, l'âge moyen des populations de contrôle et de test a été soigneusement apparié, ce qui pourrait affecter la généralisabilité des performances sur des populations plus âgées/plus jeunes.

La vérité terrain utilisée était très approximative, donc le comptage d'une détection (vrai positif) ou d'une fausse alarme (faux positif) pour un cluster donné est discutable et a une définition vague. Malgré la sensibilité rapportée pour un taux de faux positifs donné, davantage de métriques pourraient être calculées pour donner une compréhension plus claire des cartes de score produites, et le post-traitement appliqué pour obtenir le cluster perd l'information qui pourrait être obtenue en examinant l'aire sous la courbe des métriques de détection, ce qui entrave la reproductibilité de l'analyse. Des expériences ultérieures ont également montré que les patients épileptiques étudiés, ayant subi une chirurgie dans la zone suspectée d'être responsable des crises d'épilepsie, avaient fait des rechutes, remettant ainsi en question la validité de la vérité terrain (zone suspectée). Nous résumons les faiblesses de l'évaluation dans la liste suivante :

- Base de données privée
- Vérité terrain incertaine
- Résultats incertains pour les patients
- Petite taille de la base de données des patients
- Peu de métriques évaluées
- Variabilité des résultats
- Petite taille de la base de données de contrôles
- Appariement des bases de données d'âge entre patients et contrôles

Pour finir, nous dressons la liste des contributions qui vont suivre dans les chapitres suivant, et proposons un schéma synthétique, figure 3, qui résume les contributions de cette thèse.

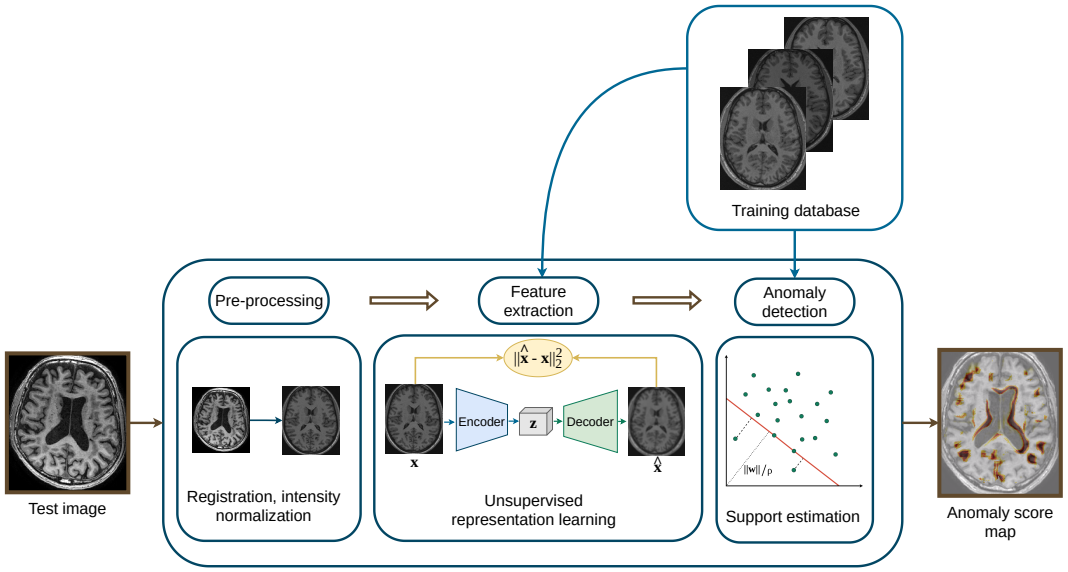

Figure 2: Pipeline complet de la méthode présentée par Alaverdyan et al. (2020).

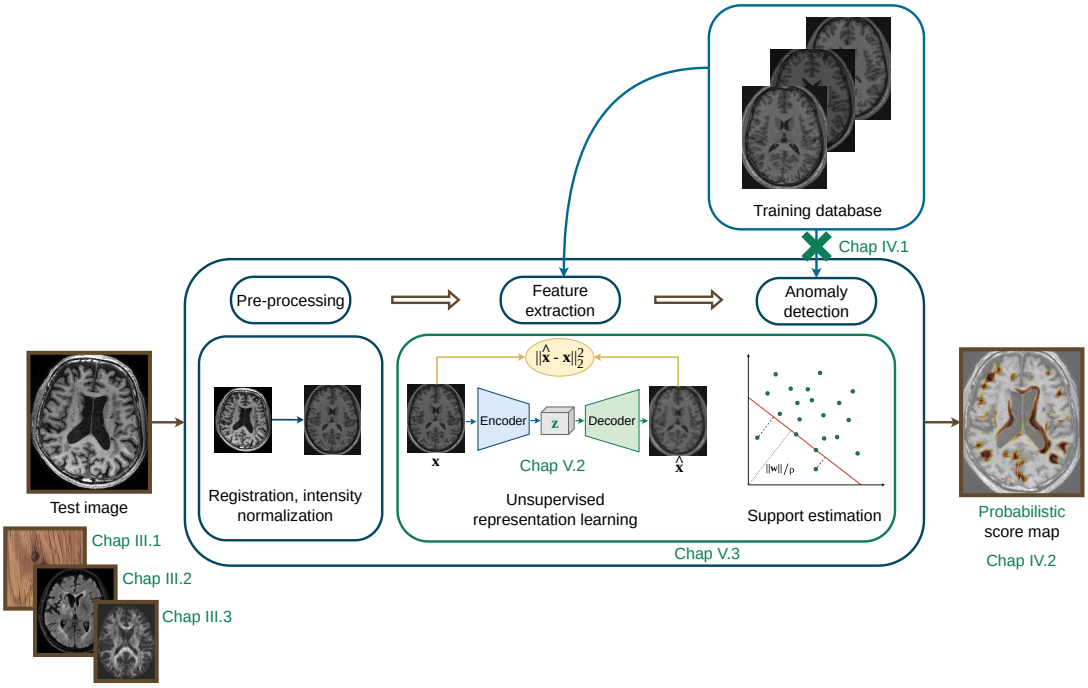

Figure 3: La plupart des contributions des différents chapitres et sections sont résumées dans cette modification de la figure 2, où la couleur verte indiques les contributions.

Chapitre 3 : Extension de l'évaluation sur des bases de données publiques

Nous avons vu dans le chapitre II qu'il est essentiel, si possible, d'effectuer l'évaluation des méthodes de détection d'anomalies sur des bases de données publiques avec une vérité terrain appropriée, et d'évaluer plusieurs métriques pour mettre en évidence les forces et faiblesses de chaque méthode. Dans ce chapitre III, nous le faisons avec trois bases de données publiques : tout d'abord, un ensemble de données d'images industrielles avec plusieurs anomalies et une vérité terrain précise ; deuxièmement, avec un ensemble de données d'imagerie cérébrale par IRM de patients présentant des lésions de la substance blanche et une vérité terrain précise ; enfin, avec un ensemble de données d'IRM cérébrale comprenant des sujets témoins et des patients parkinsoniens dé novo sans annotation sémantique mais avec une annotation au niveau du patient concernant l'état de progression de la maladie.

Nous espérons que ce chapitre démontre que les méthodes d'estimation de support, en particulier celles qui détectent les anomalies dans l'espace latent des autoencodeurs, sont des alternatives viables aux méthodes de reconstruction, sous la surveillance de bases de données diverses et hétérogènes, ce qui prouvera la grande utilité de ces méthodes. Nous espérons également que l'évaluation claire de plusieurs métriques contribuera à identifier les forces et faiblesses des méthodes proposées.

Une partie significative des technologies de pointe que nous avons abordées dans le chapitre I a été évaluée sur l'ensemble de données public MVTecAD que nous présentons dans la section III.1.1. Nous souhaitons investiguer les performances de la méthode d'estimation de support présentée dans la section II.3 sur cet ensemble de données de vision par ordinateur. Cela permet tout d'abord d'optimiser l'architecture de l'autoencodeur dans la section III.1.2. Ceci est rendu possible grâce à la vérité terrain exacte fournie avec cet ensemble de données, permettant ainsi de mesurer les performances avec plusieurs métriques. Enfin, dans la section III.1.3, nous évaluons plusieurs méthodes en utilisant cet autoencodeur : la reconstruction, l'estimation de support et la restauration, et nous élargissons la comparaison avec deux méthodes de pointe.

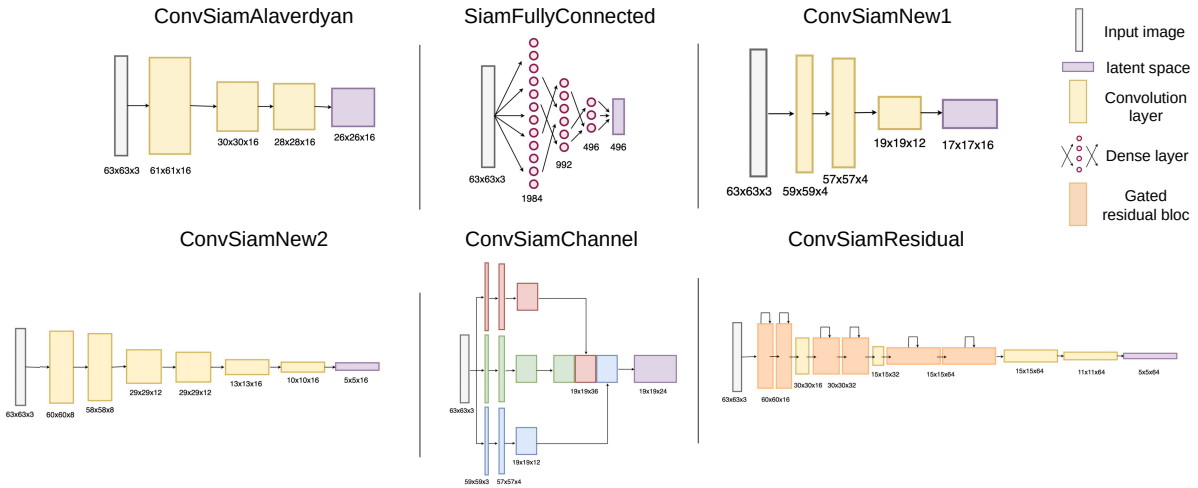

Figure 4: L'encodeur originel proposé par Alaverdyan et al. (2020), nommé 'ConvSiamAlaverdyan', ainsi que 5 architectures alternatives proposées. Les décodeurs sont construits pour être symétriques des encodeurs. Chaque auto-encodeur est entrainé par patch et avec la contrainte siamoise.

La figure 4 présente les différentes architectures d'auto-encodeurs par patch étudiées dans ce chapitre. Finalement, 'architecture ConvSiamNew1 semble être la plus performante.

Nous comparons ensuite plusieurs méthodes, sur le sous ensemble *wood* et *carpet* de MVTecAD, plusieurs basées sur l'auto-encodeur siamois par patch (SAE) : erreur de reconstruction, erreur de reconstruction quantifiée, encodeur + SVM uniclasse, encodeur quantifié + SVM uniclasse, restoration sur auto-encodeur quantifié. Nous ajoutons à cette comparaison un SVM uniclasse entrainé sur les caractéristiques (*features*) d'un ResNet50 pré-entrainé sur imagenet, ainsi que deux méthodes de l'état de l'art présentées précédemment : FastFlow et PaDiM. Des exemples de résultats sont présentées figure 5.

Nous nous attaquons ensuite à une tâche d'UAD sur imagerie médicale, sur le jeu de données WMH : *White Matter Hyperintensities* challenge (Kuijf et al. 2019). Ce jeu de données comporte 60 patients atteints majoritairement de lesions de la matière blanche, apparaissants comme hyperintenses. Nous utilisons le jeu de données de Mérida et al. (2021) comme jeu de données controles pour l'entrainement. Le tableau 1 montre les performances de 2 méthodes de l'état de l'art : Pinaya et al. (2022b) et Baur et al. (2021b), ainsi que la méthode d'Alaverdyan et al. (2020), que nous appelons SAE+*loc*OC-SVM.

WMH (T1+FLAIR) 3 hopitaux	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM
AU ROC	0.69 ± 0.13	0.53 ± 0.09	0.52 ± 0.19
AU ROC 30	0.40 ± 0.20	0.20 ± 0.12	0.19 ± 0.16
AU PR	0.065 ± 0.079	0.028 ± 0.030	0.023 ± 0.031
AU PRO	0.55 ± 0.10	0.50 ± 0.08	0.43 ± 0.17
AU PRO 30	0.19 ± 0.13	0.15 ± 0.07	0.09 ± 0.13
$\lceil Dice \rceil$	0.11 ± 0.10	0.06 ± 0.05	0.05 ± 0.05

Table 1: Moyenne (± écart-type) de chaque métrique sur tous les patients des 3 hopitaux. L'$AU\ PR$ d'un classifieur aléatoire serait de 0.007±0.006.

Dans un troisième temps nous étudions l'applicabilité de la méthode SAE+*loc*OC-SVM à la détection d'anomalies potentiellement caractéristiques de la maladie de Parkinson, sur la base de données PPMI (Marek et al. 2018). Nous comparons cette méthodes à l'erreur de reconstruction obtenue avec le même auto-encodeur par patch, ainsi qu'à l'erreur de reconstruction obtenue avec un auto-encodeur image entière. Après plusieurs étapes de post-traitement, nous concluons que les performances des méthodes par erreur de reconstruction demeurent supérieures à la méthode SAE+*loc*OC-SVM. Nous présentons les résultats quantitatifs figure 6 et quelques résultats qualitatifs figure 7.

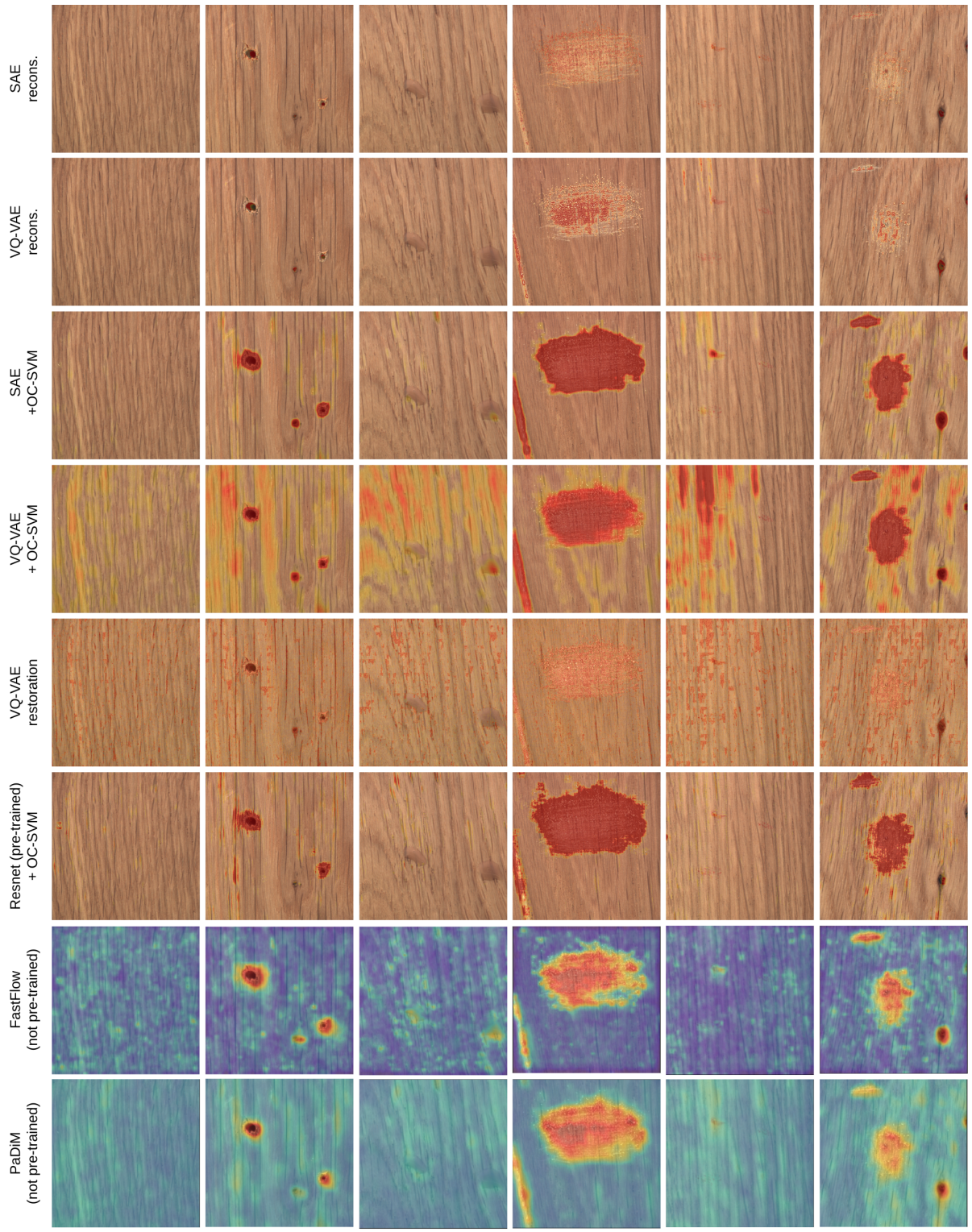

Figure 5: Comparaison visuelle des différentes cartes de score d'anomalie des méthodes étudiées. Les images sont une superposition de la carte de score (de transparent à rouge, rouge signifiant plus anormal) et de l'image d'entrée. Les deux dernières lignes ont des échelles de couleur différentes (de bleu à rouge, rouge signifiant plus anormal). Chaque colonne représente un type de défaut (ou normal), la vérité terrain peut être visualisée figure III.4.

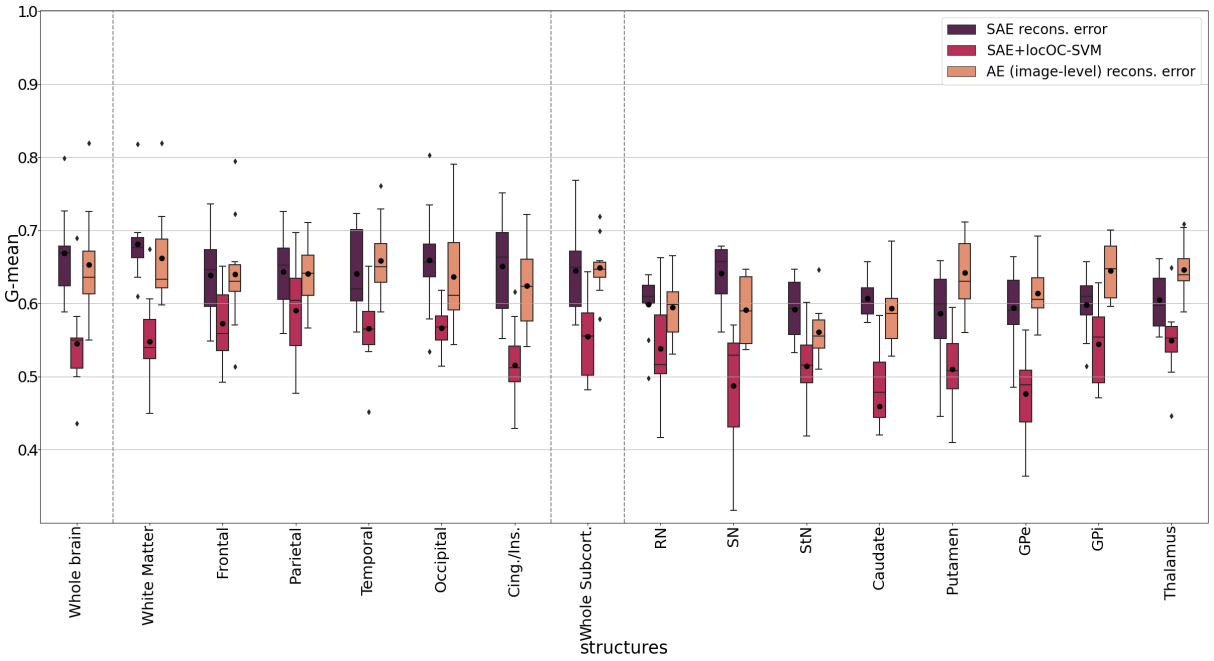

Figure 6: Moyenne géométrique (*g-mean*) de la sensibilité et de la spécificité des méthodes étudiées sur le cerveau entier et pour différentes structures anatomiques. La ligne en pointillée verticale sépare les macro des micro structures.

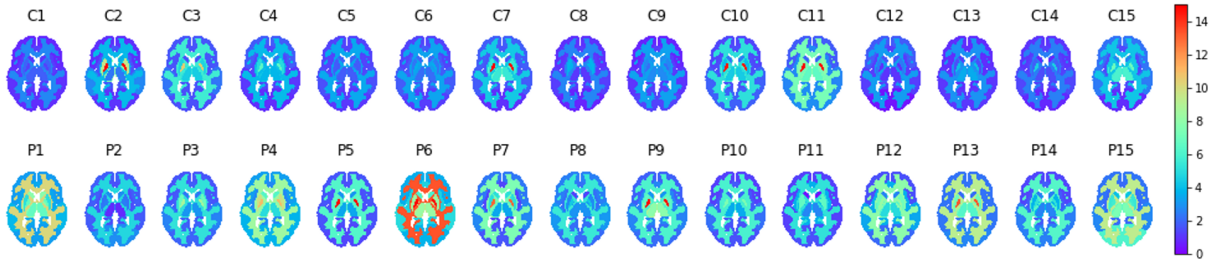

Figure 7: Le pourcentage de voxels anormaux détectés par l'erreur de reconstruction du SAE dans chaque macro-structure anatomique. Haut : les controles test du pli 0. Bas : 15 patients parkinsoniens sélectionnés aléatoirement.

Dans l'ensemble, dans ce chapitre, nous avons principalement abordé les lacunes des études précédentes sur l'évaluation. Nous avons testé le modèle sur trois bases de données différentes, publiques et difficiles, avec une large gamme de métriques. Dans les chapitres suivants, nous proposons des contributions méthodologiques supplémentaires pour renforcer les performances des méthodes d'estimation de support, que nous considérons comme pertinentes pour la détection non supervisée d'anomalies en imagerie médicale.

Chapitre 4 : Détection d'anomalie adaptée au patient et robuste

Dans le chapitre III, nous avons constaté que la méthode proposée par (SAE+*loc*OC-SVM, Alaverdyan et al. 2020), c'est-à-dire l'apprentissage de la représentation via un autoencodeur basé sur des patchs et l'estimation du support avec une SVM à classe unique, semblait être surpassée par des méthodes de pointe basées sur l'estimation de densité (pour MVTecAD, section

III.1), la restauration (pour WMH, section III.2) et la reconstruction (pour PPMI, section III.3). Nous avons toutefois constaté que les performances obtenues sur MVTecAD étaient proches de l'état de l'art, et le modèle proposé présente plusieurs autres avantages, tels que sa légèreté et son caractère basé sur le support.

Nous proposons donc dans ce chapitre d'améliorer les méthodes d'estimation de support existantes en les rendant spécifiques au patient et plus robustes. Dans la section IV.1, nous proposons d'abord une stratégie d'apprentissage différente pour la SVM à classe unique, permettant d'apprendre une frontière spécifique à chaque patient et éliminant la dépendance à la taille limitée de l'ensemble d'entraînement. Nous montrons que cela permet d'atteindre des performances de pointe sur la base de données WMH. Dans la section IV.2, nous proposons ensuite d'étudier des méthodes permettant de convertir la distance en probabilité, dans le but de réaliser un apprentissage en ensemble ou une uniformisation, permettant une détection d'anomalies plus robuste.

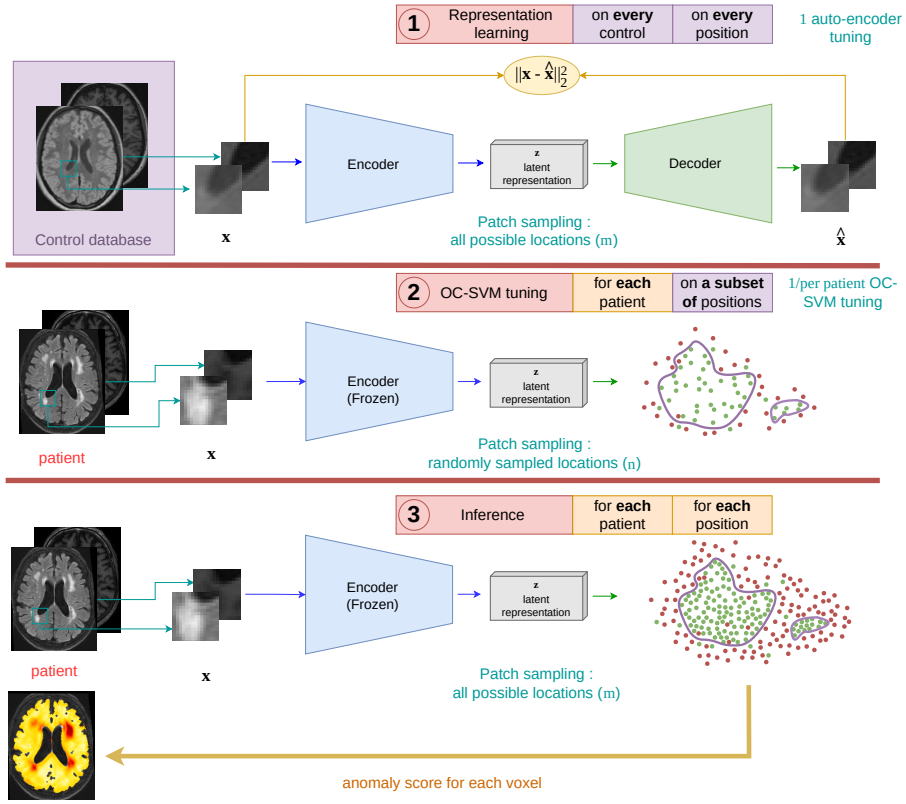

Figure 8: Diagrame synthétique du SAE+*ps*OC-SVM, comprenant l'entrainement de l'auto-encodeur siamois par patch et du SVM uniclasse spécifique au patient.

Comme nous pouvons le voir sur la figure IV.2, la première amélioration que nous proposons concerne l'étape de détection d'anomalie en tant que telle. Nous proposons cette méthode dans le but de retirer la dépendance à la taille du jeu de données d'entrainement, de rendre la méthode plus patient-spécifique, et de s'affranchir partiellement des contraintes de recalage sur un atlas commun.

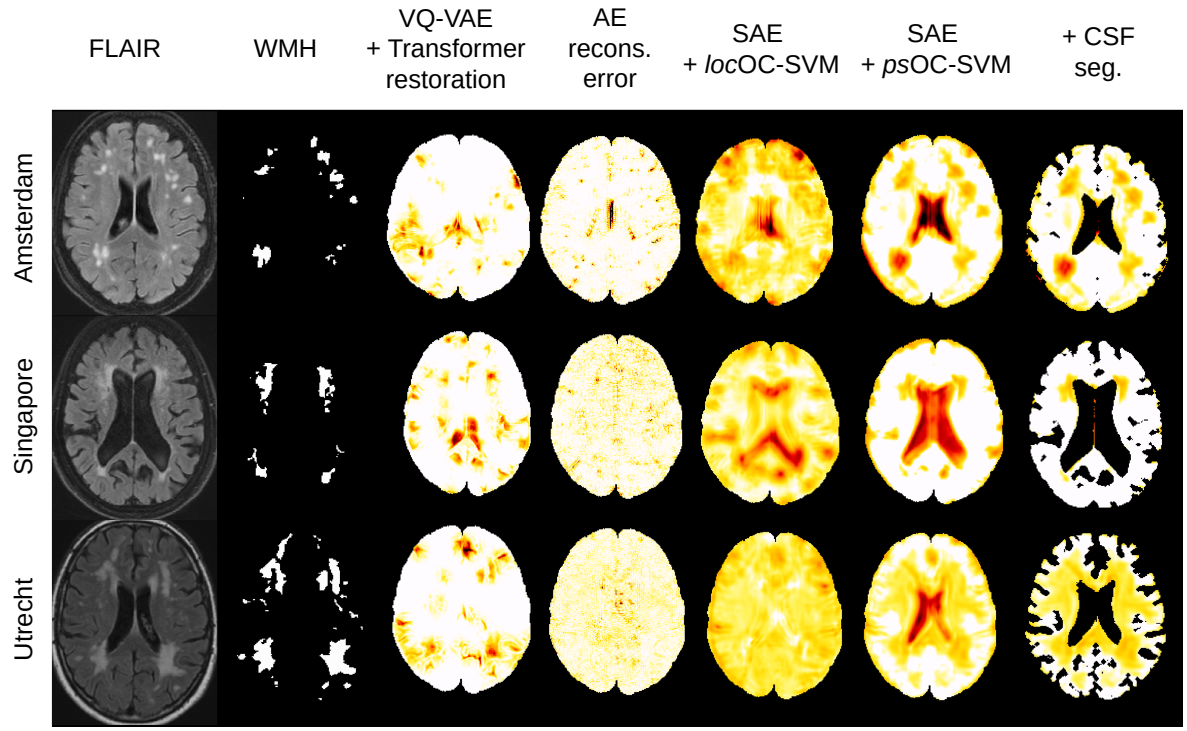

Figure 9: Démonstration des cartes de scores obtenues pour les trois méthodes étudiées sur trois coupes de trois patients (AM114, SIN63 et UT37) de chaque hôpital (les mêmes que figure IV.3), Baur et al. (2021b), Pinaya et al. (2022b) et notre méthode. Rouge signifie plus anormal. T1 et FLAIR sont utilisés comme entrées des modèles mais seulement FLAIR est montré ici.

WMH (T1+FLAIR) 3 hopitaux	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	SAE +*ps*OC-SVM	SAE +*ps*OC-SVM + CSF seg
AU ROC	0.69±0.13	0.53±0.09	0.52±0.19	**0.80**±0.09	**0.81**±0.10
AU ROC 30	0.40±0.20	0.20±0.12	0.19±0.16	**0.48**±0.20	**0.59**±0.17
AU PR	0.065±0.079	0.028±0.030	0.023±0.031	**0.084**±0.099	**0.165**±0.168
AU PRO	0.55±0.10	0.50±0.08	0.43±0.17	0.71±0.11	**0.80**±0.07
AU PRO 30	0.19±0.13	0.15±0.07	0.09±0.13	0.33±0.18	**0.48**±0.13
$\lceil Dice \rceil$	0.11±0.10	0.06±0.05	0.05±0.05	**0.14**±0.13	**0.22**±0.17

Table 2: Moyenne (± écart-type) de chaque métrique sur tous les patients des 3 hopitaux. L'*AU PR* d'un classifieur aléatoire serait de 0.007±0.006. Les modèles mis en gras sont ceux qui ne sont pas inférieurs statistiquement (p-valeur ≥ 0.01) au meilleur modèle pour la métrique considérée.

Nous évaluons cette nouvelle méthode sur le jeu de données WMH présenté au préalable, le tableau 2 présente ces résultats, où l'on constate la supériorité de la méthode présenté sur plusieurs métriques, de manière statistiquement significative. La figure 9 présente quelques résultats qualitatifs.

Nous continuons ensuite l'évaluation de cette nouvelle méthode sur la base de données PPMI présentée au préalable. Nous élargissons la comparaison à une méthode d'estimation de densité

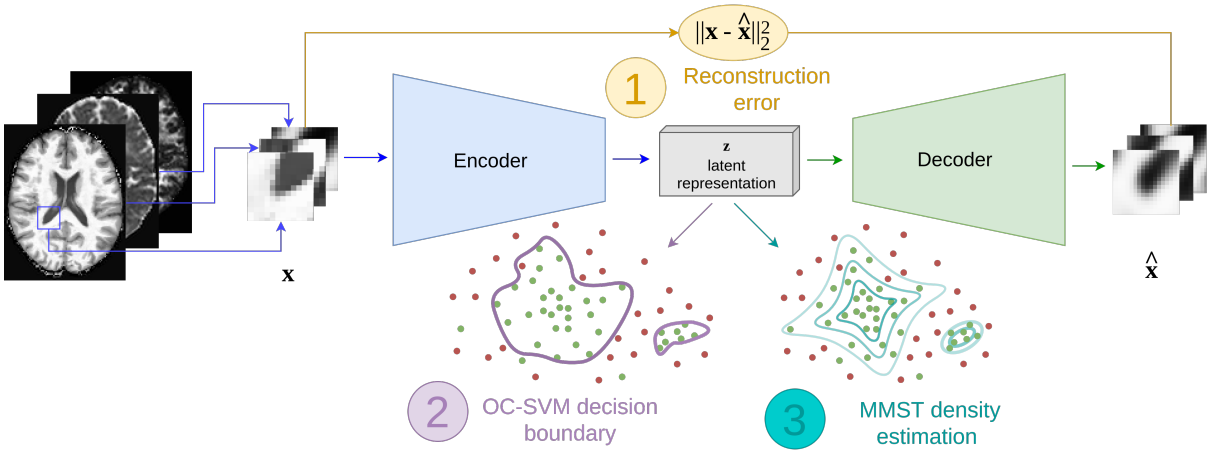

Figure 10: Schéma des trois méthodes d'UAD comparées dans cette section. L'auto-encodeur par patch est utilisé pour calculer l'erreur de reconstruction and pour extraire les représentations latente utilisées pour calculer la frontière de décision du SVM et la densité de probabilité du modèle de mixture.

développée en collaboration avec une équipe de Grenoble (LJK/GIN). La figure 10 présente une vision schématique des trois méthodes étudiées pour cette comparaison. Nous étendons la comparaison à 2 méthodes totalement supervisées.

La figure 11 présente les résultats obtenus après post-traitement. Les trois méthodes obtiennent des résultats comparables, en particulier sur les macro-structures cérébrales. Les méthodes supervisés obtiennent de moins bonnes performances que les méthodes non supervisées.

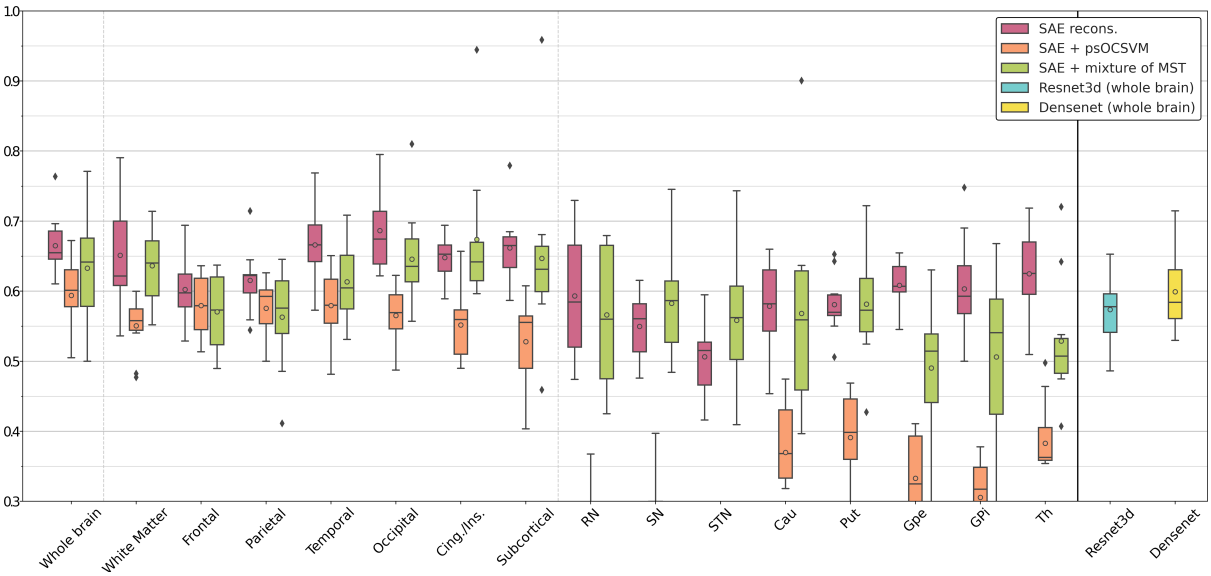

Figure 11: Moyenne géométrique (*g-mean*) de la sensibilité et de la spécificité des méthodes étudiées sur le cerveau entier et pour différentes structures anatomiques. La ligne en pointillée verticale sépare les macro des micro structures. Le Resnet3D et le Densenet sont calculés sur cerveau entier.

Dans la deuxième partie du chapitre 4, nous proposons l'étude de méthode de calibration de probabilité. Nous étudions 2 méthodes, une calibration par bacs (*binning*), ainsi qu'une

méthode plus complexe mettant en oeuvre des SVDD concentriques (cSVDD). Nous utilisons les deux méthodes présentées pour deux tâches : la fusion de plusieurs cartes de score pour de l'apprentissage d'ensemble (présenté figure 12, résultats quantitatifs tableau 3) et la calibration dans le but d'uniformiser des cartes de scores issus de modèles indépendants (présenté figure 13, résultats quantitatifs tableau 4).

WMH (T1+FLAIR) 3 hopitaux	SAE +*ps*OC-SVM	SAE +5 *ps*OC-SVM calibré par bacs et moyenné	SAE +*psc*SVDD	SAE +5 *psc*SVDD calibré avec sigmoïde et moyenné
AU ROC	0.80±0.09	0.75±0.10	0.75±0.11	0.75±0.11
AU ROC 30	0.48±0.20	0.44±0.15	0.45±0.16	0.45±0.17
AU PR	0.084±0.099	0.071±0.078	0.081±0.083	0.081±0.083
AU PRO	0.71±0.11	0.53±0.09	0.55±0.10	0.55±0.11
AU PRO 30	0.33±0.18	0.14±0.09	0.15±0.10	0.15±0.10
$\lceil Dice \rceil$	0.14±0.13	0.13±0.10	0.14±0.11	0.14±0.11

Table 3: Moyenne (± écart-type) de chaque métrique sur tous les patients des 3 hopitaux. L'*AU PR* d'un classifieur aléatoire serait de 0.007±0.006.

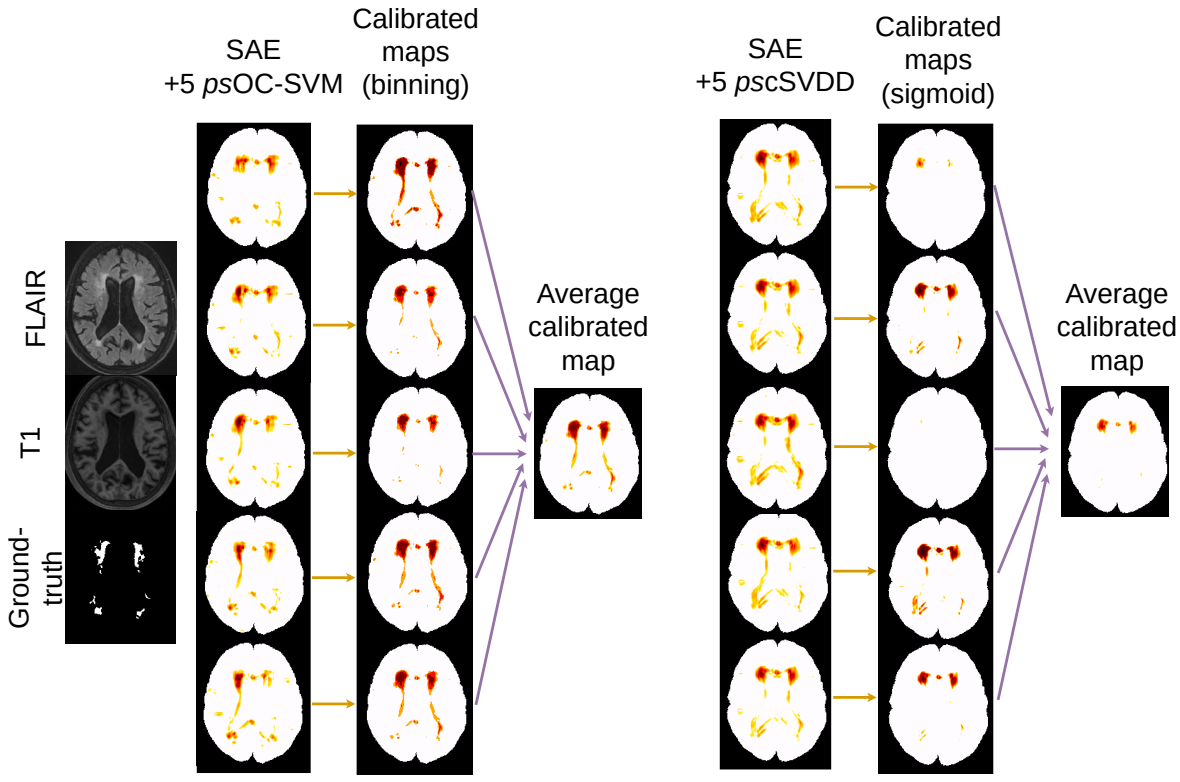

Figure 12: Démonstration des deux méthodes de calibration étudiées sur un patient aléatoire de l'hôpital *Singapore*. Sur la gauche, calibration par bacs, sur la droite, calibration avec cSVDD + sigmoïde. Les flèches oranges représentent la calibration, les flèches violettes représentent le moyennage.

WMH (T1+FLAIR) 3 hopitaux	SAE +*loc*OC-SVM	SAE +*locc*SVDD	SAE +*locc*SVDD calibré
AU ROC	0.52±0.19	0.59±0.18	0.60±0.16
AU ROC 30	0.19±0.16	0.29±0.21	0.28±0.20
AU PR	0.023±0.031	0.045±0.061	0.032±0.036
AU PRO	0.43±0.17	0.49±0.11	0.49±0.11
AU PRO 30	0.09±0.13	0.13±0.08	0.14±0.08
⌈*Dice*⌉	0.05±0.05	0.09±0.10	0.07±0.07

Table 4: Moyenne (± écart-type) de chaque métrique sur tous les patients des 3 hopitaux. L'*AU PR* d'un classifieur aléatoire serait de 0.007±0.006.

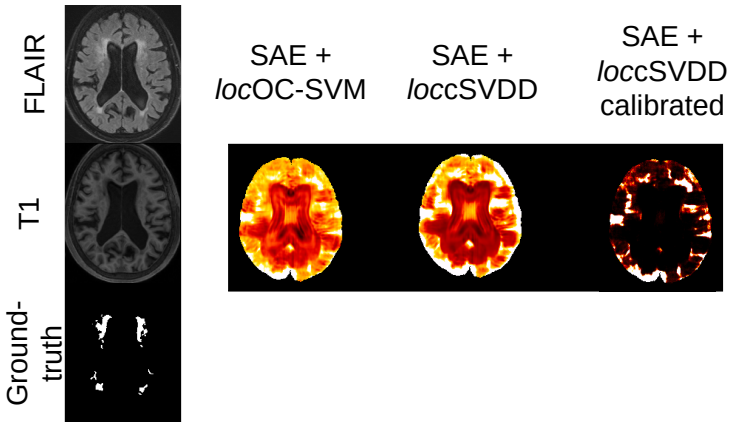

Figure 13: Démonstration de la carte de score uniformisée obtenue avec la calibration de probabilités, pour un patient aléatoire de *Singapore* (le même que sur la ligne du milieu de la figure IV.4 : SIN63).

Dans ce deuxième chapitre de contribution, nous avons proposé un nouveau cadre d'apprentissage pour les méthodes d'estimation de support, que nous avons appliqué à la détection de lésions hyperintenses et à la classification de la maladie de Parkinson. Nous avons constaté des performances supérieures à l'état de l'art sur l'ensemble de données du défi WMH et avons surpassé les méthodes supervisées pour la classification de la maladie de Parkinson.

Ce cadre proposé a permis de relâcher la contrainte sur le nombre de contrôles d'entraînement, ce qui nous a autorisés à utiliser des méthodes de calibration pour fusionner différentes cartes de score issues de sous-échantillons différents de patches. À ce stade, les différents modèles manquent de variabilité pour rendre cette fusion intéressante en termes de performances. La méthode de calibration étudiée a également été utilisée pour uniformiser les cartes de score d'anomalies par voxel, ce qui semble améliorer la détection de lésions sur cet ensemble de données, bien que ce résultat doive être interprété avec prudence.

Le cadre spécifique au patient que nous avons proposé nous a permis de relâcher la contrainte sur le nombre de contrôles d'entraînement, mais aussi supposément sur la qualité de l'enregistrement, car pour l'étape de détection des anomalies, il n'est pas nécessaire d'avoir un modèle par voxel. Une extension naturelle des expériences proposées serait de former un autoencodeur basé sur des patchs (non-siamois, qui nécessite un appariement entre les sujets enregistrés) sur des sujets non enregistrés, puis d'utiliser le SAE+psOC-SVM. Cela permettrait de ne pas avoir d'étape d'enregistrement dans le pipeline complet. C'est bénéfique car l'enregistrement est un problème et un créneau littéraire à part entière, et il est particulièrement

difficile d'avoir un enregistrement non linéaire précis lorsqu'on est confronté à une grande variété de modalités, de fournisseurs d'IRM, d'âges des patients, etc. Nous avons également constaté que nos modèles ont tendance à générer des anomalies là où les erreurs d'enregistrement sont censées être les plus élevées (près de la frontière corticale), ce qui soutiendrait la nécessité de modèles sans enregistrement pour l'imagerie médicale.

Dans ce chapitre, nous avons également constaté, tout au long de la validation approfondie dans trois hôpitaux différents, que les résultats, tant qualitatifs que quantitatifs, pouvaient varier significativement d'un centre à un autre. Cela renforce la nécessité de techniques d'adaptation de domaine, surtout en imagerie médicale, où la variabilité inter-patients et inter-centres est si élevée. Nous avons également constaté que parfois les résultats quantitatifs ne correspondaient pas aux résultats qualitatifs. Nous souhaitons, dans le chapitre suivant, approfondir l'analyse, car nous pensons que même en examinant six métriques quantitatives différentes, certains points aveugles peuvent encore apparaître.

Chapitre 5 : Espaces latents structurés pour la détection d'anomalie

Nous avons étendu l'évaluation du modèle sur trois bases de données publiques dans le chapitre III et proposé des contributions méthodologiques dans le chapitre IV pour améliorer l'étape de détection des anomalies. Nous souhaitons maintenant proposer des méthodes qui améliorent l'étape d'apprentissage de la représentation, notamment en couplant l'apprentissage de la représentation et l'étape de détection des anomalies dans un cadre unifié.

Tout d'abord, dans la section V.1, nous proposons de prendre du recul et d'examiner certaines lacunes d'évaluation que nous devons traiter lors de l'évaluation des détections de lésions hyperintenses, tout en motivant la nécessité d'une évaluation sur une tâche plus difficile. Nous explorons ensuite les méthodes existantes qui structurent l'espace latent de l'autoencodeur dans la section V.2, afin d'améliorer les performances des méthodes proposées.

Nous proposons ensuite un nouveau cadre pour coupler l'autoencodeur et la SVM à classe unique dans la section V.3, permettant un apprentissage de bout en bout et, par conséquent, une représentation latente adaptée à la tâche ultérieure d'estimation de support. Dans la section V.4, nous concluons ce chapitre en approfondissant l'analyse des cartes de score d'anomalies obtenues et des espaces latents, afin de renforcer nos conclusions.

WMH (T1+FLAIR) 3 hopitaux **hyperintensités**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	AE Siamois +*ps*OC-SVM	- FLAIR seuillage
AU ROC	0.69±0.14	0.54±0.10	0.55±0.20	0.81±0.09	**0.96**±0.03
AU ROC 30	0.40±0.20	0.21±0.12	0.25±0.22	0.54±0.17	**0.90**±0.08
AU PR	0.061±0.078	0.027±0.030	0.028±0.035	0.091±0.081	**0.494**±0.276
AU PRO	0.56±0.10	0.50±0.08	0.42±0.12	0.64±0.12	**0.92**±0.04
AU PRO 30	0.21±0.13	0.14±0.06	0.11±0.08	0.26±0.14	**0.77**±0.10
$\lceil Dice \rceil$	0.11±0.10	0.06±0.05	0.07±0.07	0.16±0.10	**0.53**±0.22

Table 5: Moyenne (± écart-type) de chaque métrique sur tous les patients des 3 hopitaux. L'*AU PR* d'un classifieur aléatoire serait de 0.007±0.006. En gras le meilleur modèle.

Le tableau 5 montre la supériorité d'une méthode simple de seuillage sur les autres méthodes présentées, justifiant la nécessité d'évaluer les méthodes proposées sur une tâche plus difficile: la détection de lesions subtiles en T1.

Dans une première partie, nous évaluons la pertinence d'un ajout de contrainte variationelle et/ou d'encodage de position, et concluons que ces pistes ne sont pas prometteuses pour la structuration de l'espace latent.

Dans une deuxième partie, nous proposons un modèle original (équation ci-dessous), qui consiste à fusionner l'apprentissage de représentation avec l'étape de détection d'anomalie. Nous proposons de séparer la fonction de coût en deux termes : un étendeur (*expander*) et un compacteur (*compactor*). La figure 14 présente une vue intuitive de ces deux termes.

[H]

$$L_{JZAD}(\mathbf{x}) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 + \lambda\beta_1 \overbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k(\mathbf{z^{SVM}}_j, \text{sg}[\mathbf{z^L}_i]) - \rho^*)}^{\text{Gradient flow only through the } \mathbf{z^{SVM}}_i}$$
$$\underbrace{+ \lambda\beta_2 \sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \text{sg}[\alpha_j^*] k(\text{sg}[\mathbf{z^{SVM}}_j], \mathbf{z^L}_i) - \text{sg}[\rho^*])}_{\text{Gradient flow only through the } \mathbf{z^L}_i}$$

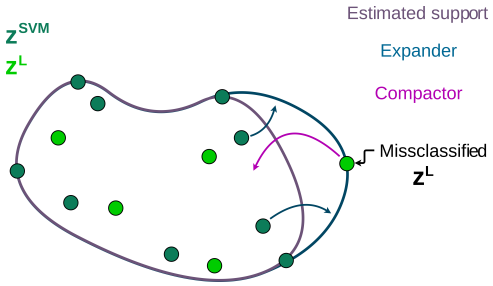

Figure 14: À l'itération N, un certain support (violet) est estimé en utilisant les $\mathbf{z^{SVM}}$. Le terme compacteur (*compactor*) agit sur les $\mathbf{z^L}$ mal classifiés et les pousse à l'intérieur du support estimé. Le terme étendeur (*expander*) agit sur les $\mathbf{z^{SVM}}$ et les pousse en dehors du support estimé pour inclure les $\mathbf{z^L}$ mal classifiés.

Figure 6, nous présentons les résultats de cette méthode versus les autres méthodes étudiées dans cette thèse. Nous concluons à une légère supériorité de notre méthode. Figure 15 présente quelques résultats qualitatifs.

WMH (T1 only) 3 hopitaux **'hyperintensités'**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	AE Siamois +*ps*OC-SVM	JZAD *expander* (*locOC-SVM*)
AU ROC	0.57±0.09	0.48±0.04	0.41±0.16	0.53±0.13	0.64±0.12
AU ROC 30	0.22±0.09	0.10±0.03	0.13±0.13	0.17±0.11	0.23±0.20
AU PR	0.022±0.023	0.013±0.013	0.017±0.018	0.015±0.012	0.040±0.054
AU PRO	0.47±0.05	0.50±0.06	0.51±0.20	0.52±0.15	0.57±0.12
AU PRO 30	0.10±0.04	0.11±0.05	0.18±0.17	0.13±0.10	0.14±0.14
⌈*Dice*⌉	0.05±0.05	0.03±0.03	0.04±0.04	0.04±0.03	0.08±0.09
WMH (T1 only) 3 hopitaux **autres pathologies**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	AE Siamois +*ps*OC-SVM	JZAD *expander* (*locOC-SVM*)
AU ROC	0.71±0.10	0.50±0.05	0.60±0.15	0.50±0.15	0.75±0.09
AU ROC 30	0.40±0.18	0.14±0.05	0.26±0.22	0.17±0.17	0.41±0.18
AU PR	0.018±0.031	0.004±0.007	0.017±0.025	0.005±0.008	0.032±0.071
AU PRO	0.25±0.34	0.18±0.24	0.22±0.30	0.08±0.17	0.15±0.30
AU PRO 30	0.14±0.21	0.05±0.07	0.09±0.17	0.02±0.07	0.09±0.19
⌈*Dice*⌉	0.04±0.06	0.01±0.02	0.04±0.06	0.01±0.02	0.06±0.10

Table 6: Moyenne (± écart-type) de chaque métrique sur tous les patients des 3 hopitaux. L'*AU PR* d'un classifieur aléatoire serait de 0.006±0.006 pour les "hyperintensités" et de 0.001±0.002 pour les autres pathologies.

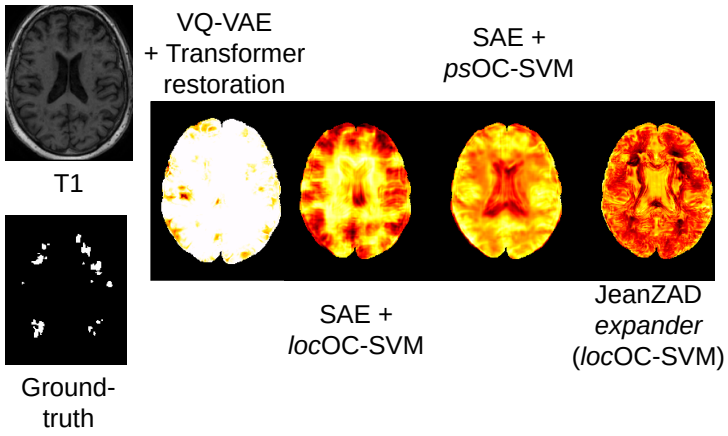

Figure 15: Démonstration d'une carte de score obtenue avec l'agorithme JeanZAD, comparé avec SAE+*loc*OC-SVM, SAE+*ps*OC-SVM et Pinaya et al. (2022b) pour le patient AM114 (même patient que la ligne du haut de la figure IV.4)

Dans la dernière partie, nous évaluons plus en détails les performances de notre méthode proposée, en fonction de l'intensité ou de la taille des lésions détectées, ainsi quà l'aide de projections UMAP (McInnes et al. 2018).

En résume, dans ce chapitre, nous avons tout d'abord démontré qu'il était nécessaire de s'attaquer à des tâches plus complexes, telles que la détection subtile de lésions sur les images par résonance magnétique (IRM) de type T1. Cependant, évaluer les performances de différents modèles de détection non supervisée pour cette tâche s'est révélé difficile en raison de sa complexité.

Ensuite, nous avons proposé d'ajouter davantage de structure à l'espace latent de l'autoencodeur basé sur des patchs pour améliorer les performances de détection. Nos premières tentatives pour y parvenir en ajoutant une contrainte variationnelle ou un encodage positionnel n'ont pas donné de résultats positifs. Cependant, des expériences supplémentaires doivent être réalisées pour déterminer si la contrainte siamoise interfère avec les améliorations proposées.

Par la suite, nous avons introduit une méthode novatrice qui fusionne l'étape d'apprentissage de la représentation et l'étape de détection des anomalies. Nous avons dérivé la fonction de perte pour ce modèle et proposé une manière intuitive de penser aux deux termes de perte (expander / compactor). Nous avons démontré que ce modèle obtenait des résultats supérieurs dans la détection de lésions subtiles, surpassant à la fois les méthodes de pointe et les méthodes précédemment proposées. Cependant, ces conclusions doivent être nuancées, car des évaluations supplémentaires sur d'autres tâches seraient nécessaires pour confirmer les performances prometteuses d'un tel modèle. De plus, l'influence des termes expander / compactor sur les performances reste à étudier, tout comme leur capacité à structurer l'espace latent.

Dans la dernière section, nous avons présenté des graphiques et des visualisations supplémentaires pour aider à comprendre l'impact du cadre de bout en bout sur l'espace latent. Nous avons également évalué les performances en termes d'intensité et de taille, et réalisé une analyse de cluster préliminaire. Bien que ces graphiques nécessitent une interprétation prudente, ils ont montré que nous avons atteint un certain niveau de structuration de l'espace latent et une certaine pertinence clinique en détectant des lésions qui n'étaient ni hyperintenses ni trop grandes. Nous avons également fourni des cartes préliminaires de lésions d'intérêt clinique pour le diagnostic assisté par ordinateur.

Conclusion

Conclusion générale

Ce travail visait à apporter des contributions méthodologiques dans le domaine de la détection d'anomalies non supervisée (UAD) en neuroimagerie. Plus précisément, nous avons étudié des méthodes qui estiment le support de densité de probabilité de la distribution normative dans un espace de représentation latent.

Après avoir introduit les concepts fondamentaux de l'UAD et les métriques classiques ainsi que les bases de données utilisées, nous avons réalisé une revue bibliographique des méthodes actuelles de l'état de l'art pour l'UAD en imagerie médicale. Nous avons également identifié certaines lacunes dans la littérature actuelle et observé les limites de l'étude menée par Alaverdyan et al. (2020).

Notre première contribution a consisté à étendre l'évaluation de la méthode proposée par Alaverdyan et al. (2020) sur plusieurs bases de données publiques et à comparer ses performances à celles des méthodes de l'état de l'art. Nous avons effectué cette comparaison sur trois ensembles de données : un ensemble de données populaire de détection d'anomalies industrielles (MVTecAD), un ensemble de données public pour le défi d'analyse de segmentation des lésions de la substance blanche cérébrale en IRM (WMH), tous deux fournissant un masque de segmentation de référence, et enfin, un ensemble de données IRM multiparamétrique (PPMI) où la détection d'anomalies a été utilisée comme tache prétexte pour effectuer une tâche de classification entre les patients Parkinson et les sujets controles.

Notre deuxième contribution se composait de deux parties visant à améliorer la robustesse du modèle proposé de détection d'anomalies non supervisé. Tout d'abord, nous avons proposé un nouveau cadre pour le SVM uniclasse, permettant l'entraînement d'un modèle unique spécifique à chaque patient. Ce modèle apprend la frontière du support basée sur des patchs extraits du patient uniquement, éliminant ainsi la dépendance à la taille limitée de l'ensemble d'entraînement des controles et moins sensible aux erreurs de recalage spatial. Cette nouvelle stratégie a été appliquée avec succès à des tâches de segmentation et de classification, sur les ensembles de données WMH et PPMI, respectivement, démontrant des performances supérieures par rapport aux méthodes de l'état de l'art établies sur WMH. Deuxièmement, nous avons abordé le problème de conversion des scores d'anomalie non bornés en probabilités. Cela a notamment permis l'apprentissage de modèles d'ensemble ou l'uniformisation des cartes de scores.

Notre dernière contribution visait à apporter davantage de structure à l'espace latent des auto-encodeurs pour l'UAD. Nous avons proposé d'atteindre cet objectif grâce à une régularisation variationnelle ou un encodage de position. Ensuite, nous avons introduit un nouveau modèle UAD permettant de coupler l'entraînement de l'auto-encodeur et du SVM uniclasse de manière intégrée (bout à bout). Les performances de cette nouvelle architecture ont été évaluées sur l'ensemble de données WMH, mais en considérant la tâche de détection bien plus difficile des lésions cérébrales en IRM T1, dont le signal est très subtil et non hyperintense comme dans les images FLAIR. Nous avons également fourni une analyse plus approfondie des succès et des échecs de ces modèles.

Limites et perspectives

Nous avons constaté que les performances obtenues sur les différents sous-ensembles de la base de données WMH étaient influencées par les caractéristiques de cette base de données, notamment les statistiques démographiques de la population et les conditions d'acquisition des images par IRM. Par exemple, la population de la base de données WMH était en moyenne beaucoup plus âgée que la population témoin utilisée pour apprendre la représentation cérébrale normative. Étant donné que le rétrécissement du cerveau est un effet reconnu du vieillissement, notre modèle était sensible à un tel effet, ce qui pourrait entraîner un rétrécissement des circonvolutions de la matière grise et ainsi potentiellement induire un grand nombre de fausses détections. L'autre caractéristique principale qui a affecté les performances de détection était la différence de scanner et/ou de paramètres d'acquisition entre les différentes bases de données, c'est-à-dire entre la base de données témoin et les trois bases de données de patients acquises dans différents hôpitaux (Utrecht, Amsterdam et Singapour). La différence de motif du signal induite par ces différentes conditions d'acquisition a affecté les performances de détection. Pour atténuer ce problème, des techniques d'adaptation de domaine pourraient être mises en œuvre.

La dépendance partielle du pipeline développé à la qualité du recalage sur un atlas souligne l'importance de ce processus. Des recherches ultérieures pourraient se concentrer sur le développement d'un algorithme dispensant du recalage, car les petites lésions risquent d'être effacées lors de cette opération, impactant ainsi la sensibilité de l'algorithme. De plus, des faux positifs semblent émerger au niveau des erreurs de recalage, notamment au niveau des circonvolutions de la matière grise et des ventricules.

Les données que nous avons étudiées dans cette thèse étaient toutes volumétriques, c'est-à-dire en 3D, tandis que les approches présentées ne considéraient que des coupes (2D) des volumes. Aborder le problème en 2D a du sens en raison du processus d'acquisition des données (généralement coupe par coupe en IRM), de la vue de référence des cliniciens pour certaines images et de la charge de calcul moindre. Cependant, nous avons constaté des pertes de contexte spatial, en particulier avec la détection anormale du rétrécissement cérébral chez les personnes âgées. Une extension naturelle de ce travail serait d'utiliser des cubes 3D au lieu de patchs 2D. Nous avons également motivé l'utilisation de patchs au lieu d'images entières, mais une autre comparaison significative consisterait à étudier la mise en œuvre de la méthode proposée sur des coupes entières, même si cela nécessiterait des adaptations supplémentaires pour obtenir une représentation latente par voxel. Enfin, la quantité de données controles disponibles a augmenté depuis le début de cette thèse. Il serait bénéfique d'inclure ces controles supplémentaires, car la base de données témoin utilisée dans ce travail était limitée en taille (75 sujets).

Étant donné que nous avons introduit des méthodes pour convertir les scores d'anomalie en sorties probabilistes, cela ouvre une application clinique naturelle : la fusion tardive de données d'image multimodales, telles que la TEP/IRM. La fusion précoce peut être réalisée en considérant les données multimodales comme des canaux, mais la fusion tardive (de cartes de scores) ne peut être réalisée qu'avec des sorties calibrées. Cette fusion pourrait également améliorer la sensibilité et la spécificité des méthodes proposées. Une combinaison des différentes méthodes proposées pourrait également être réalisée. Une autre approche consisterait à étudier la fusion de données multimodales hétérogènes, telles que des images et du texte. Cela pourrait être réalisé en incorporant les données non-image dans l'espace latent. Des approches faiblement supervisées, qui n'ont pas été étudiées dans cette thèse, pourraient également être utilisées pour cette fusion. Bien que la collecte d'un grand nombre d'anomalies puisse sembler irréaliste, la collecte d'une petite portion est une option viable dans un contexte clinique.

Dans cette thèse, nous avons motivé l'étude des méthodes d'estimation du support et proposé une nouvelle architecture combinant l'apprentissage bout à bout de la représentation et du modèle d'estimation du support. Une extension méthodologique naturelle de ce travail serait la comparaison avec les méthodes d'estimation de densité bout à bout, ainsi qu'une preuve de concept plus fondamentale sur des ensembles de données plus simples. D'autres méthodes, en dehors du SVM uniclasse, pourraient être étudiées et couplées à l'extracteur de caractéristiques qu'est l'auto-encodeur. De plus, il y a eu de nombreuses avancées scientifiques concernant l'apprentissage profond au cours des dernières années : une analyse plus avancée de l'entrainement des modèles, le choix des algorithmes d'optimisation, les régularisations et l'ajustement des hyperparamètres pourraient être entrepris pour optimiser davantage les capacités du modèle.

Une autre perspective de cette thèse est d'examiner comment les caractéristiques cliniques telles que l'âge, le sexe et les facteurs de risque affectent les représentations latentes. Une idée prometteuse serait de structurer explicitement l'espace latent en fonction de ces critères, avec l'utilisation de contraintes supplémentaires sur les auto-encodeurs. De plus, les interactions entre la contrainte siamoise et les autres contraintes ajoutées (variationnelle, de localisation, bout à bout) n'ont pas été étudiées au cours de cette thèse et pourraient être d'un grand intérêt.

Des faux positifs ont parfois été générés en raison d'une définition peu claire du concept d'« anomalie », comme le rétrécissement cérébral chez les personnes âgées produisant des anomalies sur les bords du cortex. Atténuer cette difficulté pourrait être réalisé en caractérisant l'anomalie trouvée avec des approches faiblement supervisées ou entièrement supervisées. De plus, de nombreuses étapes de post-traitement pourraient être ajoutées pour améliorer les cartes de scores, en éliminant les faux positifs attendus ou courants. Enfin, obtenir les résultats les plus cliniquement pertinents nécessite encore du travail, car l'un des objectifs de cette recherche est l'utilisation d'algorithmes UAD en pratique clinique.

Introduction

Medical imaging plays a predominant role in today's medical diagnosis. The emergence of new imaging modalities, as well as hybrid imaging systems (PET/CT, PET/MRI) enabling the simultaneous or sequential acquisition of various types of anatomical, functional, or molecular information leads to the generation of considerable data streams. This mass of multimodal and multiparametric data serves as a rich source of information for diagnosing and advancing the understanding of pathologies, but they are challenging to exploit through simple visual analysis.

Statistical analysis methods, especially those involving machine learning, can address this need. The field of research on machine learning for the analysis of medical images is highly active, particularly in supervised learning, which requires access to large structured databases. However, practical access to medical data is limited because it involves a labor-intensive annotation process, a time-consuming task that can only be performed by a clinical expert.

In this context, unsupervised machine learning, which does not require labels, and especially unsupervised anomaly detection (UAD), appears particularly suited for the task of addressing such large unstructured databases. UAD, or simply anomaly detection, aims at statistically modeling the normality observed in the data, with the goal of finding anomalies as deviating from the modeled normal distribution. Thus, UAD methods only require healthy control images to be put into practice and allow the detection of any type of pathology or abnormality without prior characterization. This can be particularly useful when the studied pathology is rare, or when the annotations are challenging to acquire.

This work attempts to contribute to the anomaly detection methods used in neuroimaging, especially density support estimation in the latent space of deep neural networks. This general framework, first introduced in Alaverdyan et al. (2020), consists of a representation learning step done through unsupervised deep modeling, followed by anomaly detection with density support estimation. This seminal study proved powerful for epileptogenic lesion detection, on a on a private multiparametric MRI database.

The primary objective of this work is to propose methodological contributions to enhance either the representation learning or outlier detection steps of the UAD model while evaluating the performances on diverse public databases.

Our first contribution is to evaluate this seminal model on three public databases. First, we utilize an industrial image database (MVTecAD Bergmann et al. 2021), which features precise ground truth and various anomaly types. Second, we leverage an open medical imaging database that contains hyperintense small brain lesions in MRI T1 and FLAIR images, known as white matter hyperintensities (WMH), with accurate ground truth Kuijf et al. (2019). Third, we utilize the PPMI database Marek et al. (2018), which includes MRI acquisitions of controls and *de novo* Parkinson patients, with ground truth at the image level. By evaluating the model

on these open databases, we facilitate comparisons with existing literature and provide a more precise assessment of the strengths and weaknesses of the approach, thanks to the availability of labels.

Our second contribution is the introduction of a new strategy for training one-class Support Vector Machines (SVMs) to address issues related to dependency on the training set (both size and extracted features), partial sensitivity to spatial registration, and the lengthy optimization process of the seminal model. This aims to enhance reproducibility, sensitivity, and specificity. Additionally, we explore methods for converting unbounded anomaly scores to probabilities., enabling the construction of ensemble models, and facilitating the calibration of score maps. These methods contribute to improving the model's ability to combine heterogeneous data.

Our third contribution is focused on enhancing the latent space representation learned through unsupervised deep modeling, to improve sensitivity in a more challenging detection task. We explore classical methods for structuring the latent space, such as incorporating variational regularization or positional encoding. We then introduce a novel end-to-end model that couples the feature extraction step with the outlier detection step. Further experiments are conducted to assess the performance of the models and the organization of the latent space, particularly in the context of subtle lesion detection.

This work is divided into five chapters. The first chapter introduces the concept of anomaly detection, including common metrics and datasets used for this task, particularly in medical imaging. Then, we review the state-of-the-art methods used in UAD for medical imaging. In the second chapter, we aim to formalize the problem studied and establish the boundaries of the seminal study. This provides an introduction to our contribution outline. Chapters three to five introduce the aforementioned contributions. The thesis concludes with a chapter on conclusions and perspectives.

I | Unsupervised Anomaly detection

I.1 Principles of Unsupervised Anomaly Detection

I.1.1 Introduction

This chapter introduces the basis of unsupervised anomaly detection (UAD). We characterize what is an anomaly, and how it can be detected in an unsupervised context.

An anomaly is defined as an event, an observation, or a sample that deviates considerably from some kind of 'normality'. In any anomaly detection task, we must first precise this concept of 'normality' to then be able to consider events that deviate from such normality. We will first clarify what we mean by anomaly detection and normality, then give a formal definition of anomaly detection, and finally describe the different families of methods for unsupervised anomaly detection.

I.1.2 Semantic clarifications on the concept of anomaly detection

Anomalies are sometimes described as outliers or novelty, depending on the context. Some studies use these terms interchangeably or make some subtle difference. As pointed out in Ruff et al. (2021), *anomaly* is usually used when the deviating sample is the point of interest in the study (e.g. suspicious bank withdrawal), *outlier*, when the sample is of no interest but deviates from the normality and as such should be removed (e.g. measurement error), while *novelty* commonly describes a sample that deviates from the observed normality until now, but should be considered as a new normal (e.g. dog of an unseen breed in a dog versus cat classification task).

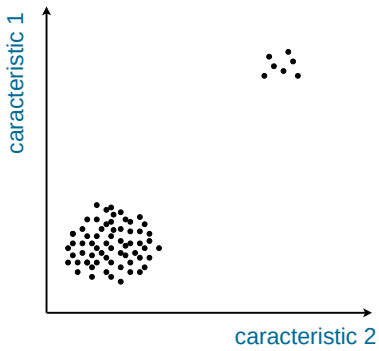

Figure I.1: Illustration of the importance of defining the concept of normality when interested in anomaly detection: one could consider the second cluster (top right) as outliers, or a second mode of the distribution, depending on the concept of normality.

Pedregosa et al. (2011b) make another semantic distinction in which anomaly detection includes both outlier detection and novelty detection. The term *outlier detection* in Pedregosa et al. (2011b) is used when the normal samples can be contaminated by some *outliers*, with no knowledge of which sample is an outlier and where the outliers are scarce (what Ruff et al. (2021) call *point anomalies*). The term *novelty* in Pedregosa et al. (2011b) is used when the normal samples are not contaminated and novelty can arise in clusters (what Ruff et al. (2021) call *group anomalies*).

The most common setup of anomaly detection is *unsupervised anomaly detection* (UAD), where no information is given on the data (that might be contaminated). Other setups include semi-supervised anomaly detection (SSAD) where a group of samples, usually small, is identified as anomalies; some authors call the unsupervised setup where the data is known to be unpolluted semi-supervised. Supervised anomaly detection, which is very uncommon, is the setup where the samples are all labeled as anomaly or normal, this setup is equivalent to binary classification and thus can be considered as an imbalanced classification problem.

As the unsupervised setup is the most commonly studied, in the remainder of the manuscript, the terms 'anomaly detection' and 'unsupervised anomaly detection' will be used interchangeably.

I.1.3 Formal definition of an anomaly

As the semantic definition of an anomaly is built on the opposition to normality, so is its formal definition. Given some data vector $\mathbf{x}$, belonging to some space $\mathcal{X}$ (usually $\subseteq \mathbb{R}^D$), we define the concept of *normality* by introducing a probability density $\mathbb{P}^{\odot}$ on $\mathcal{X}$ that is the ground truth law of normality. Samples that could be considered as abnormal lie in the low regions probability of $\mathbb{P}^{\odot}$: this is usually called the *concentration assumption* and is a core hypothesis of anomaly detection.

With the associated probability density function $p^{\odot}(\mathbf{x})$, we can consider the set of anomalies as the set of points that have probability below a threshold α, namely:

$$\mathcal{A} = \{\mathbf{x} \in \mathcal{X} | p^{\odot}(\mathbf{x}) \leq \alpha\}$$

The choice of α is crucial in anomaly detection and depends on the application: low values will give a larger normal set, that is more prone to miss anomalies, and have a higher false

negative[1] rate, whereas high values will give a tighter normal set thus miss-classifying normal samples and giving higher false positive rate.

The goal of anomaly detection methods is to estimate the probability density of the normal samples $p^{\odot}(\mathbf{x})$ (or normative distribution), or at least the support of the normal set $\mathcal{X} \setminus \mathcal{A}$, but not the hypothetical anomaly distribution $\mathbb{P}^{\oslash}$.

In the general unsupervised anomaly detection setup, we cannot sample points from $\mathbb{P}^{\odot}$, the training data will be composed of n samples $\mathbf{x}_i \in \mathcal{X}$ $(i = 1, ..., n)$ drawn from a distribution $\mathbb{P} = (1 - \eta)\mathbb{P}^{\odot} + \eta\mathbb{P}^{\oslash}$ with η the proportion of data contamination. For many methods, η is assumed to be very small or even 0, such that $\mathbb{P} \simeq \mathbb{P}^{\odot}$.

I.1.4 Rationale for anomaly detection

The task of anomaly is of great interest for multiple reasons. First, it allows the detection of unwanted occurrences without having to characterize them *a priori*: any sample deviating from the normality will be marked as an outlier. This is useful in a context where it would be difficult to obtain a clear listing of every possible anomalies because they are rare, difficult to define precisely, or even unknown at the time of the algorithm conception. In general, for statistical learning, it is easier to gather a large *normal* dataset than to gather anomalous samples. Anomaly detection is also a natural candidate to preprocess any kind of data used for statistical learning, to ensure the database is clean and will not perturbate any following algorithm.

These arguments are especially true for medical imaging where the number of rare pathology images are by definition rare, where the characterization of the considered pathology can be difficult due to a lack of medical consensus, and where we are certain that we have not yet discovered every disease.

I.1.5 Categorization of unsupervised anomaly detection method

As in Ruff et al. (2021), we partition the anomaly detection into three main families. Note that some methods do not fit exactly in those three categories (such as distance-based methods[2]), or are sometimes straddling between one family or the other, but the vast majority of methods developed in anomaly detection fit one of those three categories. Figure I.2 depicts these three families.

I.1.5.i Density estimation methods

The goal of density estimation methods is to directly model the probability density $p^{\odot}(\mathbf{x})$. The wide literature available on density estimation makes this family of methods a very good candidate for anomaly detection.

Density estimation techniques are usually separated into parametric and non-parametric methods. While the parametric methods have the advantage of being usually easier to estimate, they sometimes suffer from oversimplification and can miss some of the underlying complexity of the distributions.

One key weakness of the density estimation methods in anomaly detection is that they solve a more general problem (estimating the density distribution) than the one we are trying to solve

[1] We define as *positive* detections the finding of anomalies (not normal samples). Thus we use the symbols $\odot$ for normality and $\oslash$ for anomaly instead of $+$ and $-$ to avoid confusion.

[2] Though we could argue that the methods cited in Ruff et al. (2021) as distance-based (k-NN, LOF and iForest) are indirect density estimators and thus could fall into that category.

(detecting anomalies), which is most oftentimes harder, for instance, in low-density regions. Said differently, classification is an easier problem than regression, as for classification one only needs correct modeling near the classification frontier, whereas probability regression models need to fit the data everywhere. This is sometimes called Vapnik's principle[1].

These methods are sometimes called generative methods because one can then sample from the estimated distribution to produce a new sample, this feature, at first glance (see I.3.2.iv), seems of no use in the case of anomaly detection.

I.1.5.ii Support estimation methods

Another family of methods consists of estimating only the support of $p^{\odot}(\mathbf{x})$[2], i.e. $\mathcal{C} = \mathcal{X} \setminus \mathcal{A} = \{\mathbf{x}|p^{\odot}(\mathbf{x}) > 0\}$ this allows for solving a more simple problem, and the anomalies will be detected as the out-of-distribution samples. These methods are sometimes also called *one-class classification*, discriminative methods, or boundary-based methods.

I.1.5.iii Reconstruction methods

A third family of methods, which has no obvious link to the other two and is very popular among deep learning methods is the *reconstruction*-based methods. These methods are based on the estimation of a function Φ, which goal is to correctly reconstruct normal inputs $\mathbf{x} \sim p^{\odot}(\mathbf{x})$, i.e. $\Phi(\mathbf{x}) \simeq \mathbf{x}$. The assumption is then that anomalies, not seen during training, will be poorly reconstructed and thus detected, i.e. $\Phi(\mathbf{x}) \not\simeq \mathbf{x}$ when $\mathbf{x} \sim p^{\oslash}(\mathbf{x})$. These functions are often composed of an encoding function (reduction of dimension) and a decoding function (increase back to the original dimension), such that the information is compressed somehow, as otherwise, the function would trivially learn the identity.

We can see this family of methods as *projectors*, i.e. functions that will project any sample on the manifold of normal samples, thus, normal samples will not be moved too far away from their original position in the data space, whereas anomalies will be projected to become 'normal' and thus far away from their original representation. One can also argue that the function $d(\mathbf{x}) = ||\Phi(\mathbf{x}) - \mathbf{x}||$ can be seen as a 'distance from the normal data manifold' and thus as a distance-based method.

I.2 Metrics and datasets for Anomaly detection

I.2.1 Metrics for anomaly detection

As for any machine learning task, it is necessary to evaluate the performances of anomaly detection algorithms with relevant metrics. These metrics are very similar to the ones used in binary classification.

Evaluation will be carried out with a test dataset, composed of P positives and N negatives, to assess how many anomalies are correctly classified (True Positive, TP) or incorrectly classified (False Negative FN or miss) and how many normal samples are correctly classified (True Negative TN) or incorrectly classified (false positive FP or false alarm). Usually, any

[1] *When solving a problem of interest, do not solve a more general problem as an intermediate step* (Vapnik 2006).

[2] To be less restrictive, one can also estimate the minimum volume set (Scott and Nowak 2005, Schölkopf et al. 2001) of mass $1 - \alpha$: $\mathcal{C}_\alpha = \arg\min_{\mathcal{C}} \{\mu(\mathcal{C}) | \mathbb{P}(\mathcal{C}) \geq \alpha\} = \{\mathbf{x}|p(\mathbf{x}) > \tau_\alpha\}$ with τ_α the corresponding threshold and μ typically the Lebesgue measure. For values of α close to zero, this approximates to support estimation. Also, with the assumption that the data is corrupted with a fraction $\eta = \alpha$ of outliers that are located in the low probability region (concentration assumption), this is equivalent to support estimation.

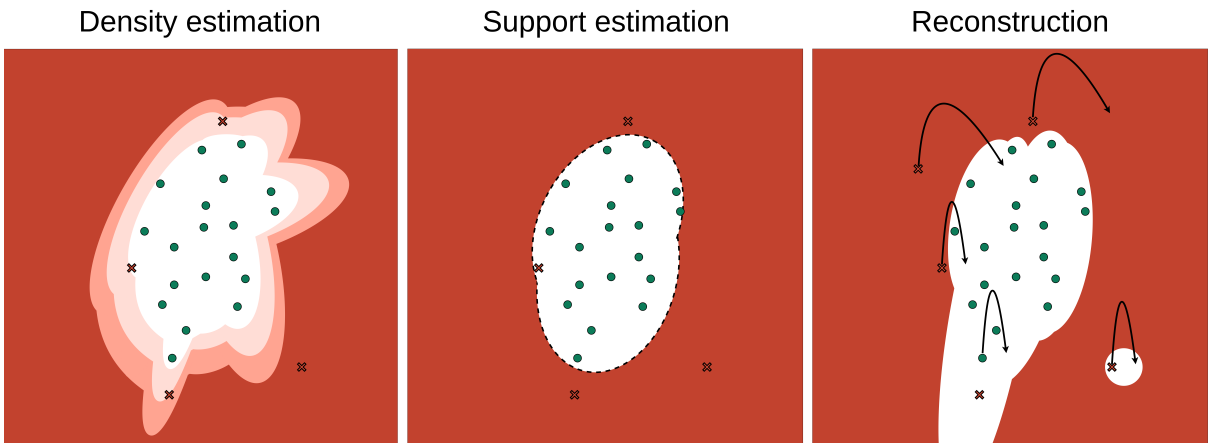

Figure I.2: Schematic comparison of the three families of anomaly detection methods, inspired by Ruff et al. (2021), red zones indicate zones detected as anomalous. Density estimation methods might overfit or underfit the tails of the distribution, or oversimplify the boundary. Support estimation methods might create boundaries that are either too loose or too constricted. Reconstruction methods might create artifact regions where anomalies are well reconstructed.

algorithm will attribute an anomaly score (e.g. in $\mathbb{R}$ or $[0, 1]$) to each test sample, one can then, with a specified threshold, classify each sample as *anomalous*/positive or *normal*/negative. An example is presented in figure I.3. It is important to note that the examples presented here use voxel-level evaluation metrics, where one voxel represents one sample. For images, it is also possible to evaluate at image-level, where one image is one sample. Image-level metrics answer the question 'Is there an anomaly in the image ?' whereas voxel-level metrics answer the question 'Where is the anomaly (if there is one) in the image ?'. The metrics presented hereunder can be used in both cases, but the presented examples will be voxel-level.

For evaluation of the performances, it is instructive to look at the confusion matrix (figure I.4), or to look at some ratios of the presented quantities, such as:

- *True positive rate* or *sensitivity* or *recall*: $TPR = \frac{TP}{TP+FN}$ the ratio between TP and positives ($P = TP + FN$), indicates the probability of detection.
- *False positive rate* or *false alarm ratio*: $FPR = \frac{FP}{FP+TN}$ the ratio between FP and negatives, indicates the probability of false alarm. Its complement is sometimes most known: the *specificity* or *true negative rate*.
- *Positive predicted value* or *precision*: $PPV = \frac{TP}{TP+FP}$ the ratio between TP and predicted positives ($PP = TP + FP$), indicates among the detections the ratio of correct ones.

A summary is given in figure I.5.

Many anomaly detection methods produce an anomaly score (i.e. a continuous score). The presented metrics involve selecting a threshold, above which a sample will be considered anomalous. This selection can be done by selecting a target false positive rate (e.g. in a credit card fraud setting where human operators could only review a fixed number of alarms per hour, one would set the false positive rate to said number) or a target true positive rate, or else.

This threshold selection will highly depend on the application, and can thus make comparison of the same metric across different applications difficult, as they will be biased.

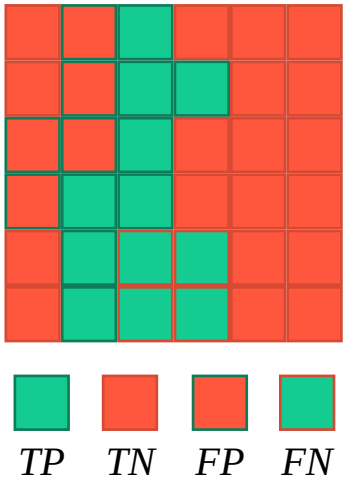

Figure I.3: Diagram of a binary segmented image's voxels, where its true value (positive or negative) is the interior color of each voxel, and its prediction is the contour of the voxel.

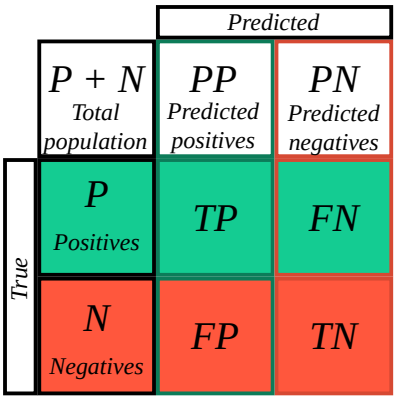

Figure I.4: Summary of the different quantities computed when looking at binary classification (and anomaly detection). The count of TP, FN, FP and FN (lower right quadrant) constitutes the *confusion matrix*.

For this reason, it is often of interest to compute these values for multiple thresholds, more precisely, for every threshold possible. This is the rationale behind a commonly used metric: $AU\ ROC$ (Area Under the ROC curve[1], also called AUC), where the ROC curve is the true positive rate plotted against the false positive rate. By measuring the area under this curve (the integral), one can have a global idea of the performance of the evaluated method, without the need of setting a threshold. More formally:

$$AU\ ROC = \int_{FPR=0}^{FPR=1} TPR(FPR)dFPR$$

While the $AU\ ROC$ has become a standard evaluation metric in classification, its usefulness is more arguable in the case of highly imbalanced datasets (Baur et al. 2021a, Davis and Goadrich 2006), which are very common in anomaly detection, where the positive class (anomalies) are very scarce. As an alternative, the $AU\ PR$ or $AU\ PRC$ (Area under the precision-recall curve), which evaluates the balance between precision (positive predicted value) and recall (true positive

[1]Note that ROC means *receiver operating characteristic* and so the last C is not for curve, as commonly thought.

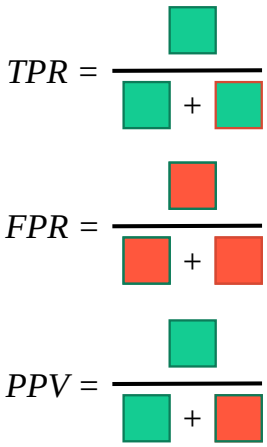

Figure I.5: Depiction of the TPR or *sensitivity* or *recall*, *false positive rate* or *false alarm ratio* and *positive predicted value* or *precision*.

rate), is less sensitive to this imbalance: in $AU\ ROC$ the false positive rate is evaluated against the true positive rate, thus a large change in the number of false positives can lead to a small change in the false positive rate (when there is a high number of true negatives) whereas in the $AU\ PR$ the false positives are compared to the true positives (in the precision), thus any small number of false positives will decrease the precision. More formally:

$$AU\ PR = \int_{TPR=0}^{TPR=1} PPV(TPR)dTPR$$

Put it simply: $AU\ PR$ (precision against true positive rate) is preferable to $AU\ ROC$ (true positive rate against false positive rate) when precision is more important than false positive rate, which would be the case in imbalanced datasets.

Said in a different fashion: we are interested in seeing if the first detections are the good ones; we can see that with precision but not with false positive rate.

A drawback of the $AU\ PR$ versus the $AU\ ROC$ is that for $AU\ ROC$, the random classifier will always have 0.5 $AU\ ROC$, while in $AU\ PR$ the random classifier will have $\frac{P}{N+P}$ as $AU\ PR$, making it a metric more difficult to compare on different datasets. An example of ROC and PR curves is given in figure I.6.

Another drawback of these metrics, when used for voxel-level evaluation, is that they do not account for the size of the detected area. As seen in figure I.7, an algorithm could increase its true positive rate by refining its detection of a large area and ignoring the detection of small ones, whereas one could argue that it could be more important to detect every positive (e.g. lesions), even if not perfectly segmented, than to have perfect segmentation of positives and ignore the smaller ones.

To account for this drawback, Bergmann et al. (2021) have popularized, in the anomaly detection community, the Per Region Overlap (PRO) curve, which draws the PRO against the false positive rate. The PRO is defined[1] as:

$$PRO = \frac{1}{N_c} \sum_{k=1}^{N_c} \frac{|P_k \cap PP|}{|P_k|}$$

[1] In Bergmann et al. (2021) it is actually defined as being a sum over the test images, the definition is equivalent when considering one image.

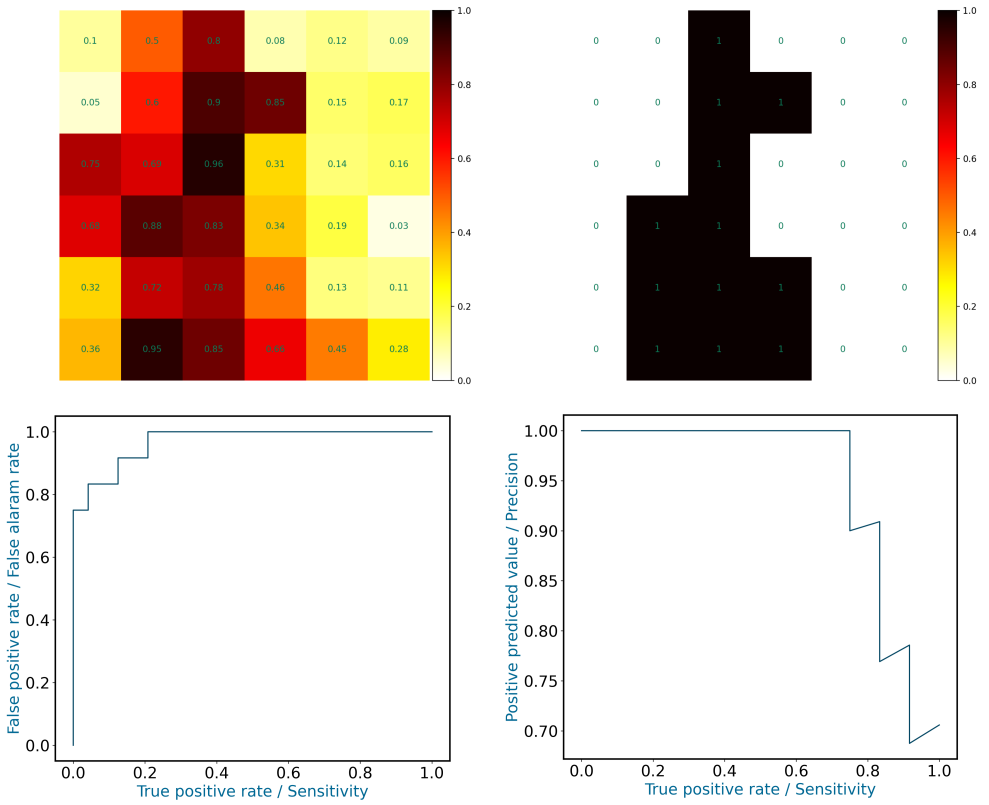

Figure I.6: Example (same as I.3) of an image anomaly score map (top left), its ground truth (top right), and the associated ROC curve (bottom left) and PR curve (bottom right)

Where N_c is the number of connected components of the positives in the image, P_k the k^{th} connected component of the positives (also called ground truth), and PP the predicted positives in the image.

As the sensitivity/true positive rate can also be defined as the ratio of the intersection of predicted positives and positives[1], the PRO effectively acts as a TPR normalized by the ground truth (positives) size, giving equal importance to the detection of small or large positives. We could also see the PRO as a 'clusterized' sensitivity, the sensitivity being normalized by the size of each cluster (ground truth). The PRO is then computed against the FPR, as for the $AU\ ROC$, to obtain the $AU\ PRO$:

$$AU\ PRO = \int_{FPR=0}^{FPR=1} PRO(FPR)dFPR$$

Bergmann et al. (2021) have also argued that in the case of imbalanced datasets with a large number of negatives, having FPR go to extreme values such as 90% gives degenerate binary maps that are not meaningful and not useful in practice. They propose to set the maximum FPR to 30% and as such, introduce $AU\ ROC$ 30 and $AU\ PRO$ 30 where the area under the curve is only computed up to 0.3 FPR, and then normalized by 0.3 to get results between 0 and 1 as the $AU\ ROC$ and $AU\ PRO$:

[1] $TPR = \frac{TP}{TP+FN} = \frac{P \cap PP}{P}$. See figure I.4.

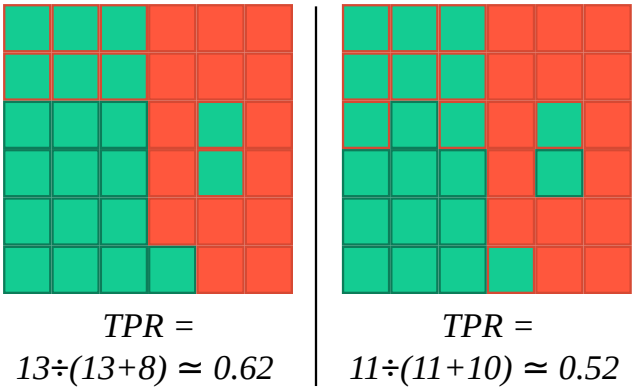

Figure I.7: Comparison of two classifications and their TPR: despite completely missing the small right area, the right classification achieves better TPR than the left classification. It can be argued that it is more important not to miss any positive area, even if its segmentation is not perfect.

$$AU\ PRO\ 30 = \frac{1}{0.3}\int_{FPR=0}^{FPR=0.3} PRO(FPR)dFPR \qquad AU\ ROC\ 30 = \frac{1}{0.3}\int_{FPR=0}^{FPR=0.3} TPR(FPR)dFPR$$

Note that the random classifier, achieves 0.15 $AU\ ROC\ 30$ and $AU\ PRO\ 30$.

A metric that is very commonly used in segmentation and thus in medical image segmentation is Dice's coefficient (Dice 1945), also called Sørensen index (Sorensen 1948):

$$Dice = \frac{2|P \cap PP|}{|P| + |PP|} = \frac{2TP}{2TP + FP + FN}$$

With P the positives (ground truth) and PP the predicted positives. It is essentially a measure of overlap between the positives and the predicted positives. There is no commonly admitted measure of the area under a curve of Dice against something else, but it is common in the anomaly detection community to measure the best achievable Dice, which is the maximum of the Dice over the possible thresholds.

We have described here metrics that can be used with any type of data, although the figures I.7 and I.3 show examples of these metrics being evaluated at the pixel level. It is common in anomaly detection to differentiate two tasks:

- Detecting if the image contains an anomaly or not: sometimes called *anomaly detection* or *anomaly screening* or *anomaly classification*.
- Detecting if each pixel is anomalous (is an anomaly or belongs to an anomaly) or normal, sometimes called *anomaly localization* or *anomaly detection* or *anomaly segmentation*.

In this present manuscript, we will make the difference between the two tasks by calling the first *image-level anomaly detection* and the second *pixel-level anomaly detection*

I.2.2 Datasets for anomaly detection

The rapid growth of statistical learning algorithm performances in recent years can be attributed to many factors, including but not limited to the large number of public databases released to train statistical algorithms and to evaluate them on common ground.

As a subdomain of statistical learning, anomaly detection also has reference databases on which the community evaluates the performances of its algorithms. Ruff et al. (2021) propose to classify the databases into three types:

- *K-classes out*: these datasets are based on multi-class classification datasets (including binary), where one class is artificially considered normal and the others are considered anomalous. As the classes are often very different, this can result in very obvious anomalies and thus not very realistic.

- *Synthetic*: a curated normal dataset is used, and synthetic, predefined anomalies are added to some normal samples. This can allow for generating a large number of anomalies of different types, but the real-world anomaly statistics, as they are hard to model, may differ greatly from the synthetic ones.

- *Real-world*: a dataset where normal samples have been gathered, as well as anomalous samples and fully annotated. This case is preferable although it might still not encompass every possible anomaly. Also, these datasets are sometimes so challenging that it might be hard to compare different methods' performances.

The constitution of real-world databases is a complex task because by definition the anomalies are rare events. It might be challenging to gather a variety of different anomalies, thus justifying the use of synthetic databases or K-classes out. Even real-world databases, because of the limited samples, will have a limited number of anomaly types, and will not be perfect estimates of a hypothetical 'real' database, i.e. some anomalies may have never been observed before they appear in practice. Note however that this will limit only the correctness of the evaluation and not the training (which does not use anomalies).

I.2.2.i Databases for general anomaly detection

We will briefly introduce datasets used in the image anomaly detection community, as it allows us to illustrate the three family types presented above, and also because they are sometimes used for the pre-processing of medical imaging algorithms. When interested in medical imaging anomaly detection, it is also of interest to broaden the applications to any images to benefit from the methodological developments done in computer vision.

- *K-classes out*:
  - MNIST (LeCun et al. 1998): a very popular dataset in machine learning and computer vision, containing 70,000 images of hand-written digits (from 0 to 9), with image size 28 by 28 pixels. Usually, in anomaly detection, it is interesting to see the difference in performances between test digits that look like the train digit and others that don't, e.g. when training on an 8, the gap of performances for detection of a 0 (looks like a 8) and a 2.
  - Forest Cover Type (Blackard 1998): this dataset provided by the US Forest Service is composed of 7 forest cover types, and presents 54 features (such as elevation, slope, hill shade, soil type, etc.). It has been used for instance in Liu et al. (2008).
  - Image-net (Deng et al. 2009): this popular dataset contains 1431167 natural RGB images spanning 1000 classes (including animals' race, fruits, objects, etc.) and with various image sizes (often resized to a fixed size). It is also widely used for pretraining purposes.

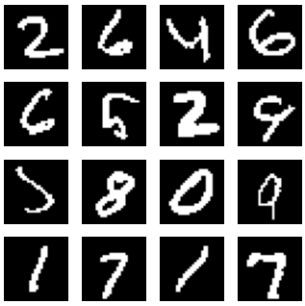

Figure I.8: Examples of the MNIST (LeCun et al. 1998) database of handwritten digits.

- *Synthetic*:
  - Mu and Gilmer (2019) introduced common perturbations to MNIST (shearing, rotation, motion blur), and Ruff et al. (2021) used the perturbed images as outliers.
  - Breunig et al. (2000) have used 2D Gaussian as the normal samples and for outliers, samples from a uniform distribution. Examples in Pedregosa et al. (2011a) also include normal samples as a mixture of Gaussian or simple geometric shapes (e.g. half-circles) and outliers as uniformly sampled in the domain.
- *Real-world*:
  - In Quinlan (1987), the ANNThyroid dataset is composed of 7200 samples composed of 21 features such as sex, pregnancy, previous surgery, etc., and contains pathological and healthy samples. It is used in Liu et al. (2008).
  - The MVTec AD dataset (Bergmann et al.; Bergmann et al. 2019; 2021) is a very popular dataset in the image anomaly detection community. It is composed of 5354 high-resolution (1024x1024) RGB images, of 15 classes: 10 objects (hazelnuts, pills, bottle, etc.) and 5 textures (wood, carpet, etc.) with 73 possible defaults over the 15 classes. The presence of multiple defect types for each class makes this dataset challenging and closer to real practice. It is presented in figure I.9.

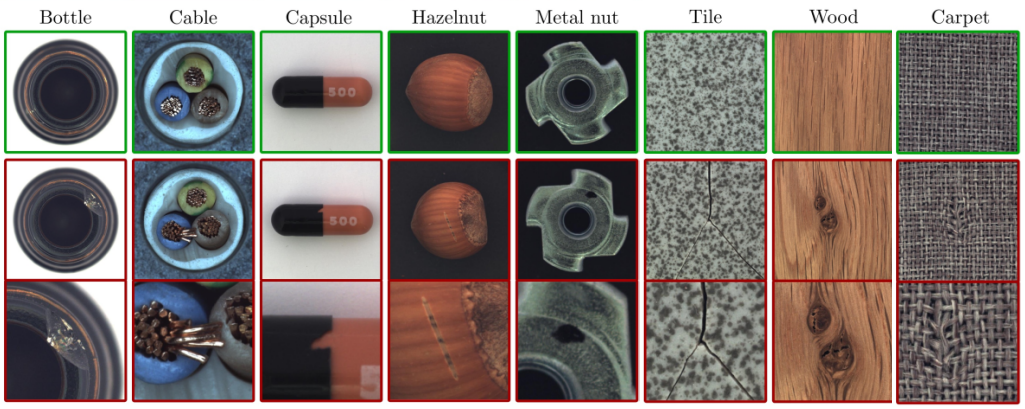

Figure I.9: Images of different MVTec AD objects and textures, normal samples (top line) are contoured in green and anomalous samples (middle line) and zooms (bottom) contoured in red. Image modified from Bergmann et al. (2019)

It is worth mentioning that synthetic datasets are often made on the spot for each task, making them maybe more fit to the task but also diminishing the possibility of comparison between methods.

I.2.2.ii Specific databases for medical imaging

We will introduce some of the commonly used datasets in medical image anomaly detection, as they will be referred to in the section I.3.2, but also to give a quick glance at the usual datasets and their weakness.

- *K-classes out*:
  - To the best of our knowledge no studies in anomaly detection have used k-classes out medical imaging datasets (except the case where the class is normal VS the rest, which amounts to a real-world setting)
- *Synthetic/real-world*:
  - The Medical Out-of-Distribution Analysis Challenge 2020 (MOOD Zimmerer et al. 2022) consists of two different tasks, anomaly detection on brain MRI and anomaly detection on chest CT-scan. The brain MRI dataset and the chest CT-scan dataset respectively contain 800 and 550 healthy scans. The validation set consists of 4 toy examples for each dataset. The test set is hidden from the participant of the challenge but contains among other things: corruptions, pathologies, and even non-medical images artificially added to the scans.
- *Real-world*:
  - The Parkinson progression marker initiative (PPMI Marek et al. 2018), consists of 3805 controls and patients[1]. The patients are at early pathological stages, for which brain structural abnormalities are subtle and hardly visible in standard structural MRI. Studies such as Muñoz-Ramírez et al. (2020) have tried identifying anomalies in the parkinsonian brains.
  - The Brain Tumor segmentation (BraTS Menze et al. 2015) is a yearly challenge for segmentation of brain tumors. The 2021 edition contains 8000 multi-parametric MRI brain scans of patients with gliomas (a type of brain tumor). It has been widely used for anomaly detection (Behrendt et al. 2023, Behrendt et al. 2022, Meissen et al. 2021a, Chen et al. 2021, Chen and Konukoglu 2018, Marimont and Tarroni 2021a, Kascenas et al. 2022b, Luo et al. 2023, Kascenas et al. 2022a, Chen and Konukoglu 2018, Zimmerer et al. 2019, Pinaya et al. 2022a, Zhao et al. 2022b, Pinaya et al. 2022b).
  - FastMRI+ (Zhao et al. 2022a), consists, first, of 1172 knee MR images with bounding boxes annotations for 22 different pathologies and secondly of 1001 brain MR images with bounding boxes for 21 different pathologies and image-level labels for 8 other pathologies. Bercea et al. (2023b) used this dataset's healthy brain MRI slices to train an unsupervised anomaly detection model based on auto-encoders, and used the pathological slices for testing. The wide range of possible anomalies makes this database interesting for anomaly detection.

[1] As of August 2023. Some of the patients have identified Parkinson's disease biomarkers or genetic variants but no symptoms yet.

It is important to note here the absence of real anomaly detection databases with the appropriate ground truth in the medical imaging literature. The real-world databases used in the literature are either private (in-house database) or suffer crucial flaws:

- The PPMI database does not contain voxel-level ground truth but only image-level ones, making it impossible to evaluate if the found anomalies are correct. Because the patients are still at an early stage of the pathology, it is not even sure that patients have detectable structural abnormalities.

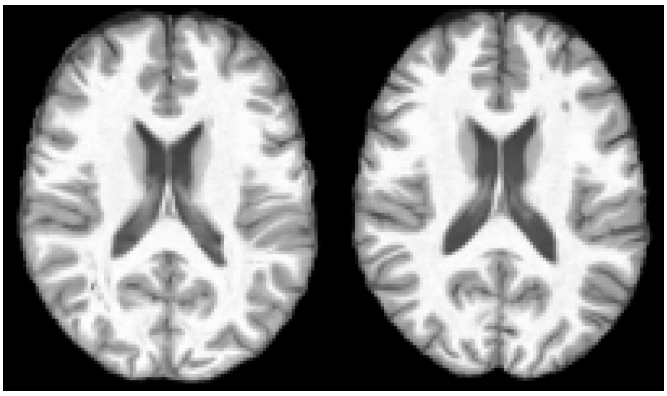

Figure I.10: Showcase of two slices of co-registered T1 MR images from a control (left) and a patient (right), both extracted from the PPMI database. There is no visible lesion, although the patient has been diagnosed with Parkinson's.

- FastMRI+ suffers from the same default as there are only image-level labels or bounding boxes. Although the bounding boxes are more precise than image-level labels, the pathologies to detect are often visible by the naked eye, i.e. not subtle (resection, edema, enlarged ventricles, etc.).

- The BraTS dataset, although possessing voxel-level ground truth, is arguably not a proper anomaly detection task, because the abnormalities are all very similar (gliomas[1]) and have very large volumes. It is not a requisite of anomaly detection to have small anomalies, but the task is generally considered interesting when the abnormalities make up for only a small fraction of the total data, are difficult to detect, and present variable patterns in the image.

Although only identified here on four databases, these challenges are present, to the best of our knowledge, in all the available medical image anomaly detection datasets[2] and make it challenging to properly evaluate and compare different anomaly detection methods in medical imaging. We will present in more detail these challenges in chapter II.

[1] Although differentiating between different parts of the tumor, the anomaly is still always a tumor.

[2] The MOOD challenge (Zimmerer et al. 2022) could be a very good candidate judging by the diversity of the anomalies in the test set, but as the test set is hidden to the participant and the validation set only contains 4 toy samples, the performances evaluation is severely limited.

I.3 State of the art methods in Unsupervised Anomaly Detection for images

There have been plenty of contributions to the field of anomaly detection in recent years. As stated before we will focus here on unsupervised anomaly detection (UAD, which we also simply call anomaly detection) and focus on methods that apply to images. Every method presented here is not bound to be used for images but the large literature on anomaly detection specific to computer vision, especially since the blossoming of deep learning, justifies the focus on UAD for images. We will thus not investigate methods that apply to, for instance, times series, graphs, etc.

We first present, in section I.3.1, what we think are 'fundamental methods' or elementary building blocks of the anomaly detection systems for images, with applications for medical or non-medical imaging. This selection, although arbitrary, will allow us to refer to this section to explain other methods in the following. We precise at the end of this section some methods that are state-of-the-art on the MVTecAD database, as we will be using these methods for comparison in section III.1.

Then, in section I.3.2, we present a more exhaustive state-of-the-art of the anomaly detection methods used for medical imaging. We conclude this chapter with table I.1 which summarizes this bibliography.

I.3.1 Fundamental methods used in UAD for computer vision

To present the fundamental methods used in UAD for computer vision, we re-use the classification introduced by Ruff et al. (2021), presented in section I.1.5.

I.3.1.i Reconstruction methods

Auto-encoders

In the ages of deep learning[1], a very common way of detecting anomalies and a very common baseline is auto-encoders.

Auto-encoders[2] (presented figure I.11) are a type of neural network with the primary use of compressing an input $\mathbf{x}$ into a *latent representation* $\mathbf{z}$ (also called latent code).

This is done with two separate networks, an *encoder* E and a *decoder* D (also called a *generator*). The encoder compresses the input into a latent representation and the decoder must produce an output $\hat{\mathbf{x}}$ (also called a reconstruction) which is the closest possible to the input from the latent representation.

The encoder and decoder are then trained by penalizing the difference between $\mathbf{x}$ and $\hat{\mathbf{x}}$, usually in the form of a pixel-to-pixel L^2 norm[3]. The two networks E and G will then be optimized by commonly used stochastic gradient descend (SGD) algorithm[4]. The end goal is thus to find the weights that minimize the difference between the input and the output (also called the reconstruction error):

[1] A large number of the presented methods make use of deep learning, which has seen its usage grow considerably, especially since ∼2014. We do not present the basis of deep learning and refer the novice reader to this popular book: Goodfellow et al. (2016).

[2] Which seem to date back to at least Ballard (1987).

[3] We obtain the mean squared error (MSE) by taking the average of the L^2 norm. Because the arg min is the same, they are used interchangeably.

[4] Recent popular optimization SGD algorithms include Kingma and Ba (2015) and Loshchilov and Hutter (2018).

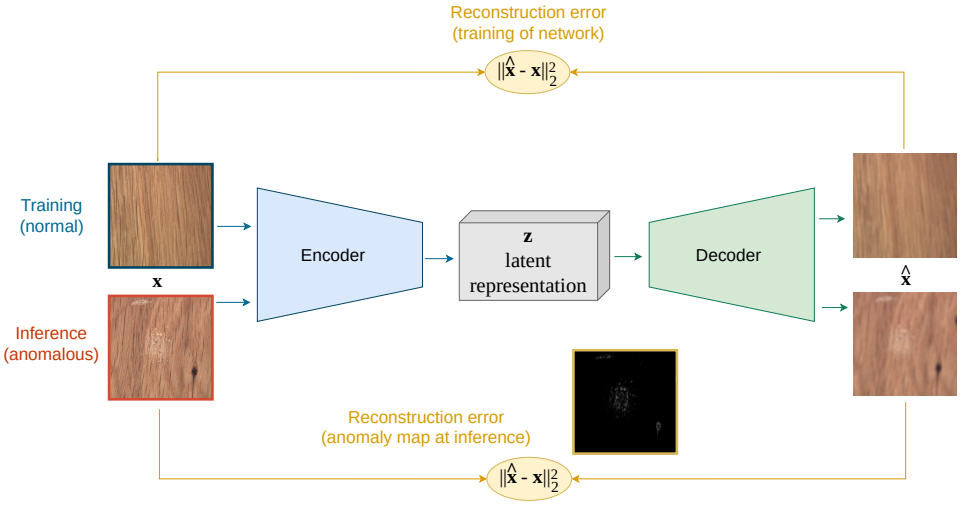

Figure I.11: Diagram of the training of an auto-encoder with L^2 reconstruction error, and inference with L^2 reconstruction error (it could be another criterion than the one optimized during training).

$$\min_{E,D} L_{AE}(\mathbf{x}) = ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 = ||\mathbf{x} - D(E(\mathbf{x}))||_2^2$$

Auto-encoders make use of the idea that the input data $\mathbf{x}$, although high dimensional, belongs to a lower dimension manifold embedded in the high dimensional space (Goodfellow et al. 2016). The goal is thus to recover this lower dimensional space with the encoder. This encoding/decoding process could be obtained by other means than deep learning, for instance with a principal component analysis transformation and by keeping only the principal axis that explains the most variance.

The main hypothesis is that when trained with normal samples only, at inference, the reconstruction error ($||\mathbf{x} - \hat{\mathbf{x}}||_2^2$) will be higher for abnormal samples (because not seen during training, i.e. projected into the 'normal manifold'). For images, this amounts to saying that pixels that contain anomalies will have higher reconstruction errors than pixels that don't.

It is worth noting that neural networks (including auto-encoders but not limited to) used to process images are often convolutional neural networks (CNN): neural networks in which most of the operations done are discrete convolutions, as they offer strong inductive bias on how the network should process the image (e.g. receptive field, filters, etc.).

U-net

A very popular neural network, widely used for segmentation and particularly in medical imaging[1] is the U-net (Ronneberger et al. 2015), presented figure I.12. It can be considered as a tweak of auto-encoder where skip connections are added between the encoding and decoding branches, such that the output can benefit from the high-abstraction-level encoded information and the precise and high-resolution features from the raw image. Also, the U-net generally outputs segmentation masks (and not the input image, which could be trivially recovered from the skip connections), and is trained in a supervised way. Thus, as is, it cannot be used for anomaly detection. Also, because it contains skip connections, the encoded latent vector does not represent the entirety of the input image, and can thus difficulty be used for feature extraction.

[1] The U-net was initially proposed for biomedical imaging.

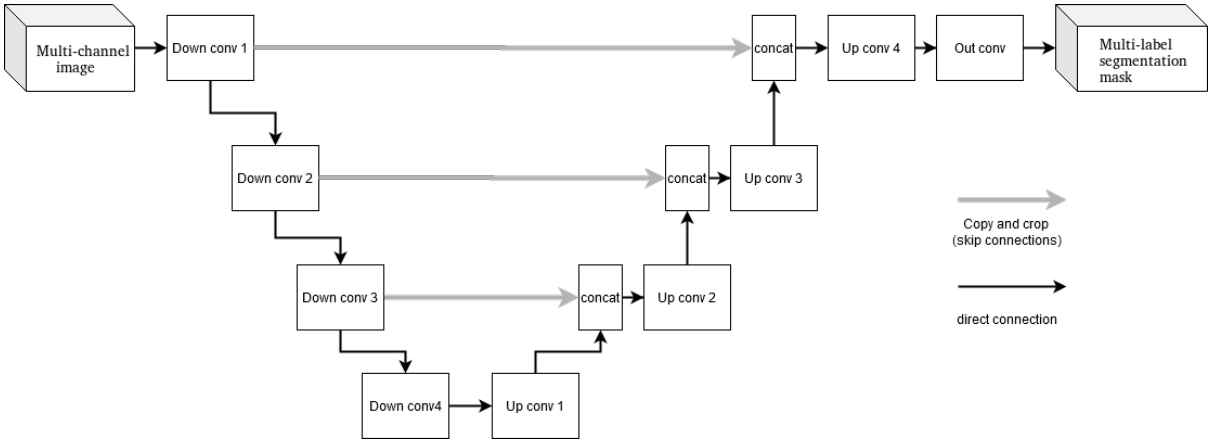

Figure I.12: Diagram of a standard U-net Ronneberger et al. (2015), composed of downsampling blocks, upsampling blocks, and skip connections. This architecture is supposed to have the semantic representation power of auto-encoders, with augmented accuracy of segmentations given by the skip connections.

We present later, in section I.3.2, UAD methods that use U-net as an indirect mean.

Variational auto-encoders

There is a popular variant of auto-encoders, called variational auto-encoders (VAE Kingma and Welling 2014, presented figure I.13), where instead of having a latent representation $\mathbf{z}$ after encoding, one has two outputs of the same dimension as $\mathbf{z}$: a mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ and the latent vector is obtained by a sampling operation $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Also, a KL divergence (defined below) term pushes the mean and standard deviation to be close to a standard normal distribution. One can also say that the posterior distribution $p(\mathbf{z}|\mathbf{x})$ obtained by the encoder is matched to a standard normal distribution (the prior distribution).

This is supposed to have a regularization effect on the latent space, because as there will be sampling around balls of center $\boldsymbol{\mu}$, and these balls are close due to the KL term, a more continuous space is obtained, and thus sampling in this latent space will give samples closer to the training distribution as one would obtain with a 'classical' auto-encoder (also called deterministic auto-encoders as there is no sampling, thus no stochastic process).

VAE can be used directly as reconstruction methods, when using the reconstruction error ($||\mathbf{x}-\hat{\mathbf{x}}||_2^2$), but can also be used for their structured latent space (more structured than classical AE), as a feature extraction method preliminary to other methods (such as density/support estimation methods)

The KL divergence (Kullback and Leibler 1951) D_{KL} between a discrete probability distribution $P(\mathbf{x})$ and $Q(\mathbf{x})$ is defined as:

$$D_{KL}(P||Q) = \sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{Q(\mathbf{x})}.$$

It is a statistical measure of how the probability distribution $P(\mathbf{x})$ is different from $Q(\mathbf{x})$. It will be 0 if the two distributions are equal and tend to $+\infty$ if they are completely different (disjointed supports). In the case of VAE, $P(\mathbf{x})$ is the output of the encoder, and $Q(\mathbf{x})$ is the standard normal distribution.

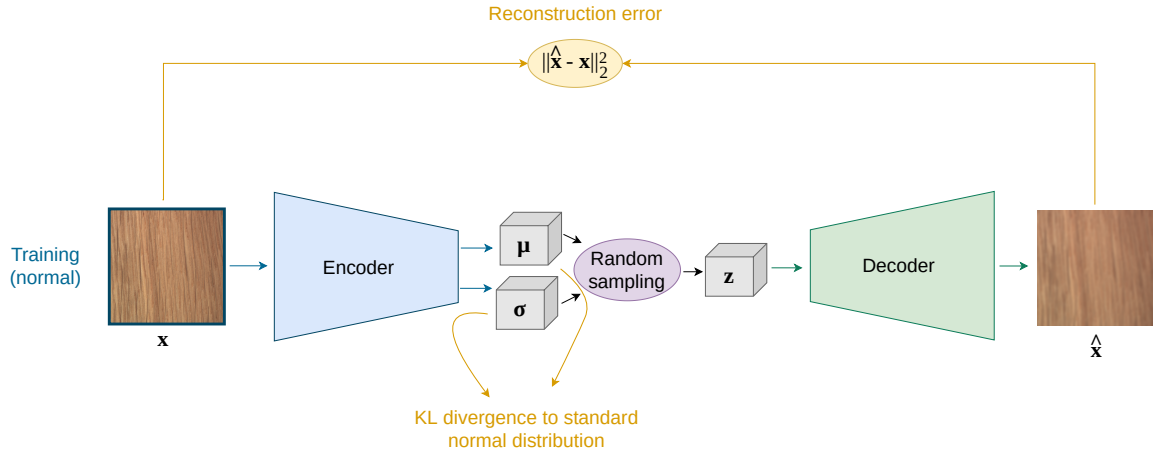

Figure I.13: Diagram of the training of a VAE (Kingma and Welling 2014). As all the generated $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are pushed toward $\mathcal{N}(0, 1)$, and thus brought closer, this ensures a form of continuity in the latent space.

VQ-VAE

Van Den Oord et al. (2017) have proposed another variant of the auto-encoder, called vector quantized-variational auto-encoder (VQ-VAE, presented figure I.14). VQ-VAEs are composed of an encoder and a decoder like the classical auto-encoder, but the latent representations are quantized through a discrete dictionary $\mathcal{E}_K = \{e_1, ..., e_K\}$ that is learned in the process. Every vector coordinates z^j output by the encoder is replaced by the closest dictionary element, i.e. $z^j \rightarrow e_k$ with $k = \arg\min_l ||z^j - e_l||_2^2$. The loss is the following :

$$L_{VQ-VAE}(\mathbf{x}) = ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 + ||sg[\mathbf{z}] - \mathbf{z_q}||_2^2 + \beta||\mathbf{z} - sg[\mathbf{z_q}]||_2^2 \tag{I.1}$$

where $\mathbf{z}$ is the vector output by the encoder, $\mathbf{z_q}$ its quantized version, and sg the stop gradient operation.

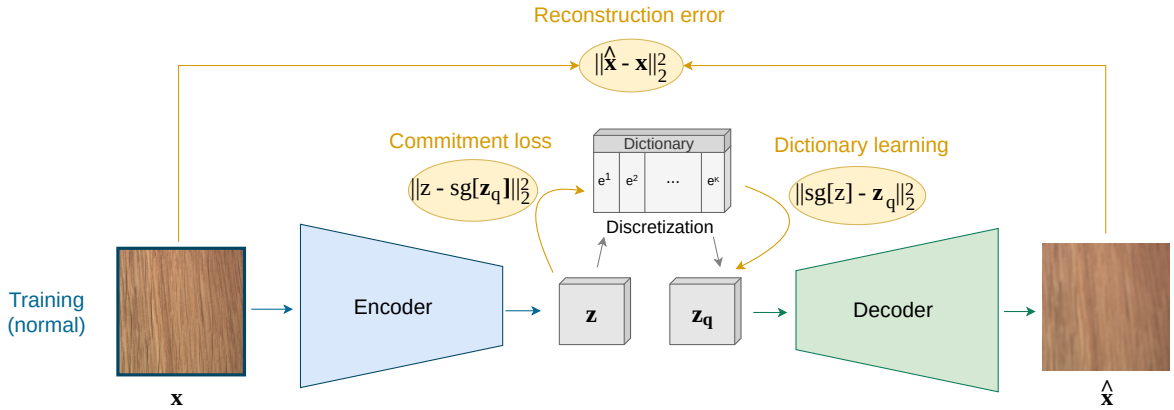

Figure I.14: Diagram if the training of a VQ-VAE (Van Den Oord et al. 2017), comprising the three loss terms presented in equation I.1.

We recognize the classical reconstruction error (in its L^2 norm form) $||\mathbf{x} - \hat{\mathbf{x}}||_2^2$, the second term $||sg[\mathbf{z}] - \mathbf{z_q}||_2^2$ allows learning the dictionary (no gradient flow on z), and the third term $||\mathbf{z} - sg[\mathbf{z_q}]||_2^2$ brings the encoder output closer to the dictionary. The second term only affects

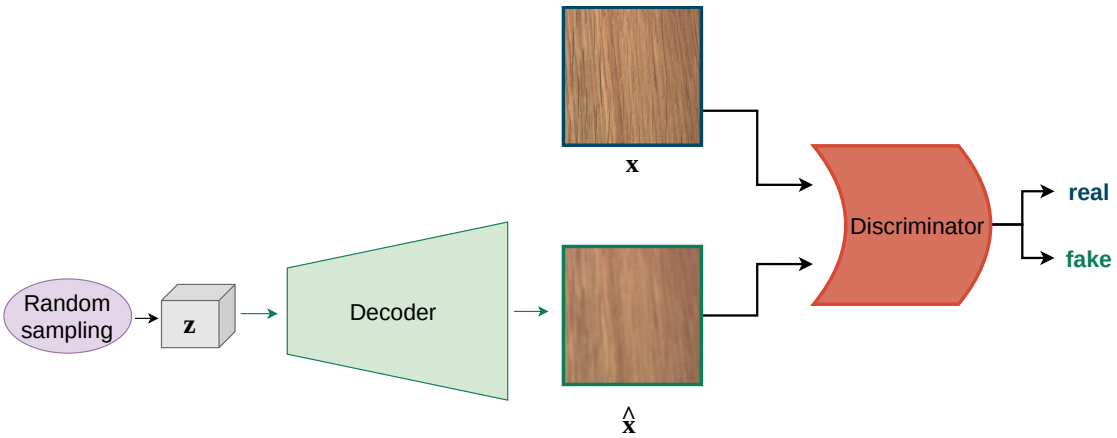

Figure I.15: Diagram of the training of a GAN (Goodfellow et al. 2014). The discriminator must classify between real images from the dataset and fake images generated by the decoder. The decoder must fool the discriminator.

the dictionary so no weight coefficient is needed. The first and third terms affect the encoder, the first term only affects the decoder.

The reconstruction error of VQ-VAE, as AE and VAE, can be readily used for anomaly detection. One can also make use of the quantized latent space they provide, as feature extractors, to use discrete density/support estimation methods.

Generative adversarial networks

Generative adversarial networks (GAN Goodfellow et al. 2014, presented figure I.15) are a type of network composed of a *generator* $\mathbf{D}$ (also called a decoder) that generates image $\hat{\mathbf{x}}$ from a latent prior distribution (usually gaussian) $p(\mathbf{z})$ and a *discriminator* C (or classifier[1]) that will try to differentiate between real samples $\mathbf{x}$ and generated samples $\hat{\mathbf{x}} = \mathbf{D}(\mathbf{z})$.

The two networks are trained in an adversarial way (min-max game), see the loss $L(C, \mathbf{D})$ in equation I.2. As the two network objectives are concurrent, each will be optimized in an alternating fashion, usually K learning steps for the discriminator alternating with one step for the generator. Note the similarity of the generator process, that decodes a latent code $\mathbf{z}$ into a generated image $\hat{\mathbf{x}}$ to the process of sampling in VAE, as such, these two architectures have long been competitors for image generation.

$$\min_{\mathbf{D}} \max_{C} \; L(C, \mathbf{D}) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log C(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - C(\mathbf{D}(\mathbf{z})))]. \tag{I.2}$$

The first term of equation I.2 pushes the discriminator to correctly classify the real samples, while the second term pushes it to classify generated samples as such (fake). The second term pushes the generator to 'fool' the discriminator into classifying a fake sample as real (as the second term is maximized by the discriminator but minimized by the generator). Note the total absence of any encoder in this process, contrary to VAE.

The discriminator of the GAN can be readily used for anomaly detection, but only at image-level: as is, it does not provide the part of the image responsible for classification as 'fake'.

[1] It is most common when referring to GANs to name the generator $\mathbf{G}$ and the discriminator D, but this conflicts with the decoder $\mathbf{D}$ presented in this manuscript.

Denoising models

Another family of reconstruction methods is *denoising models*: a network is trained to denoise the corrupted version $\tilde{\mathbf{x}}$ of the original image $\mathbf{x}$. At inference time, anomalies will be denoised and thus erased because they are less statistically probable than normal regions. By comparing denoised image $\hat{\mathbf{x}}$ and original image $\mathbf{x}$ pixel-wise, one can detect the anomalies. These methods have the following drawback: they need to ensure the noise statistics follow approximately the anomalies statistics, which amounts to a form of statistical modeling of the anomalies (see section I.1.4). Typical denoising models include denoising auto-encoders (Goodfellow et al. 2016), that are input with noisy images $\tilde{\mathbf{x}}$, but trained to reconstruct the noiseless image.

Diffusion models (Sohl-Dickstein et al. 2015, Ho et al. 2020) have become popular denoiser neural networks in recent years. Their principle is iterative and straightforward: at each timestep t, noise is added to the current noisy image $\mathbf{x}_t$. Iteratively noise is added to the base image $\mathbf{x} = \mathbf{x}_0$, until a time T where the image $\mathbf{x}_T$ is pure noise, this process is called the forward process. A network (often U-net-like) is trained to do the reverse process, i.e. from any noisy image $\mathbf{x}_t$, conditioned by the timestep t, recover the partially denoised image $\mathbf{x}_{t-1}$ (figure I.16 depicts the process). This network can be trained in a supervised manner because the added noise is known. At inference time, an image is partially noised until timestep L ($< T$), and the reversed process is applied to get a denoised image. Anomalies, not seen during training, are supposed to be erased by this process.

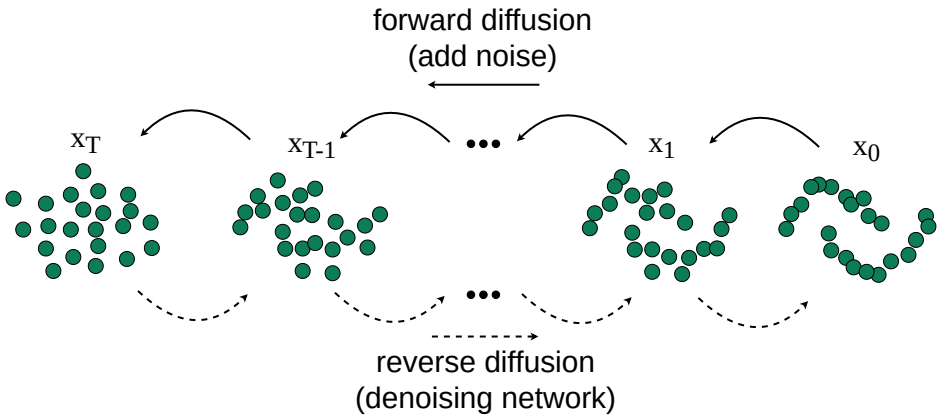

Figure I.16: Diffusion models learn to reverse the forward diffusion process.

I.3.1.ii Density estimation methods

Mixture models

The goal of mixture models, when used for anomaly detection, is to model the normal data distribution $p(\mathbf{x})$[1] as accurately as possible, with the goal that outliers would be in the low probability regions. These methods are not specific to images but have been used quite extensively (examples for UAD include Defard et al. 2021, Arnaud et al. 2018a, Prastawa 2004) to model voxels intensities or even multi-channel voxels.

The goal of a mixture model is to compute an estimate of the probability density $\hat{p}(\mathbf{x})$, as a mixture of carefully chosen distributions f:

[1]Here we denote the normal probability distribution $p(\mathbf{x})$, contrary to section I.1.3, for simplicity, and to encompass the case where the training data could be polluted by a small fraction η of outlier, i.e. $p(\mathbf{x}) = (1-\eta)p^{\odot}(\mathbf{x}) + \eta p^{\oslash}(\mathbf{x})$.

$$\hat{p}(\mathbf{x};\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x};\Phi_k) \qquad \text{with} \sum_{k=1}^{K} \pi_k = 1$$

where π_k is the mixture weight (or proportion) of the component k, $\mathbf{\Phi}$ the whole set of parameters of each probability distribution $f(\mathbf{x};\Phi_k)$. The parameters of the model can be estimated with a wide range of methods, for instance, likelihood maximization. Mixture models fall under the category of parametric models (one has to choose a specific distribution f and estimate its parameters). A popular choice of distribution is the Gaussian distribution, notably for its simplicity. An example of a Gaussian mixture model is presented in figure I.17.

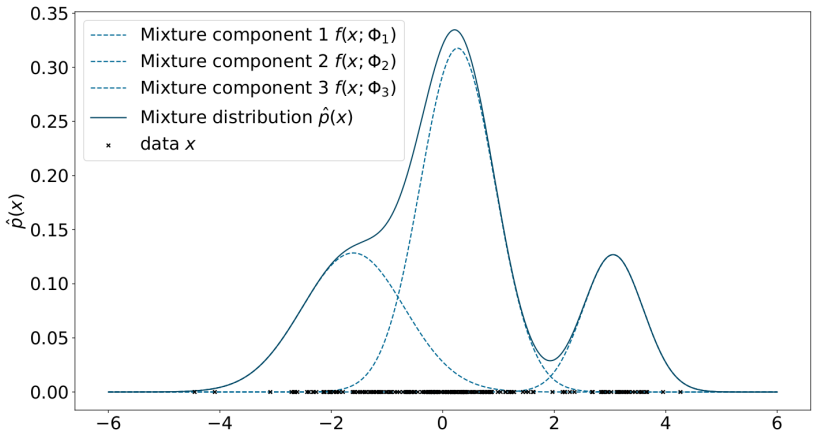

Figure I.17: Example of a 1-dimensional mixture model. The data points have been fitted with maximum likelihood estimation by the Gaussian mixture model $\hat{p}(\mathbf{x})$ with $K = 3$.

Kernel density estimation

The goal of kernel density estimation (also called the Parzen-Rosenblatt window estimator Parzen 1962, Rosenblatt 1956), is to approximate the distribution of the data $p(\mathbf{x})$, by the kernel density estimator $\hat{p}(\mathbf{x})$ defined as :

$$\hat{p}(\mathbf{x}) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

Where K is a non-negative function, called a kernel or a window, which will weight each data point $\mathbf{x}_i$ and h is a bandwidth parameter, which will determine the smoothing of the function or equivalently the range of interactions between evaluated points and existing data points.

These estimators fall under the category of non-parametric estimators, the main advantage being there is no assumption about the shape of the probability density function $p(\mathbf{x})$. This doesn't come with no hyperparameter choice though, since there is still the need to choose the kernel function and the bandwidth.

Normalizing Flows

Normalizing flows (Tabak and Turner 2013) are neural networks $\boldsymbol{f}$ that are built to transform a probability distribution $p_X(\mathbf{x})$ into a reference, prior probability distribution $p_Y(\mathbf{y})$. This mapping is built such that it needs to be bijective and invertible. $p_X(\mathbf{x})$ is the data distribution and $p_Y(\mathbf{y})$ is taken to be the standard normal distribution $\mathcal{N}(0,1)$. This process is represented in figure I.18 for a 1-dimensional example.

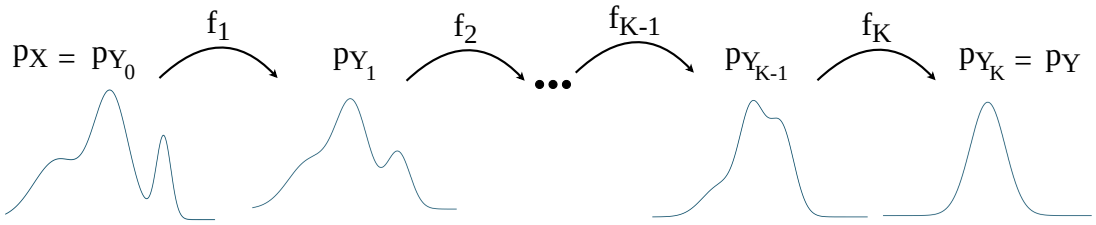

Figure I.18: Diagram of the normalizing flow process in dimension 1. The original data distribution $p_X(\mathbf{x})$ is transformed through bijective, invertible functions f_k to obtain $p_Y(\mathbf{y}) \sim \mathcal{N}(0, 1)$.

To ensure the mapping $\boldsymbol{f} : X \to Y$ is bijective and invertible, it is composed of multiple simple functions $\boldsymbol{f}_i$ that are bijective and invertible, thus $\boldsymbol{f} = \boldsymbol{f}_1 \circ \boldsymbol{f}_2 \circ ... \circ \boldsymbol{f}_K$.

The link between $p_X(\mathbf{x})$ and $p_Y(\mathbf{y})$ is obtained through a change of variables:

$$p_X(\mathbf{x}) = p_Y(\mathbf{y}) \left| \mathbf{det} \frac{\partial \boldsymbol{f}(\mathbf{x})}{\partial \mathbf{y}^T} \right|$$

with $\frac{\partial \boldsymbol{f}(\mathbf{x})}{\partial \mathbf{y}^T}$ the Jacobian matrix of the transformation $\boldsymbol{f}$, $\mathbf{y} = \boldsymbol{f}(\mathbf{x})$ and $\mathbf{x} = \boldsymbol{f}^{-1}(\mathbf{y})$.

With $\mathbf{y} = \mathbf{y}_K$ and $\mathbf{x} = \mathbf{y}_0$:

$$\mathbf{y} = \mathbf{y}_K = \boldsymbol{f}_K(\mathbf{y}_{K-1}) = (\boldsymbol{f}_{K-1} \circ \boldsymbol{f}_K)(\mathbf{y}_{K-2}) = ... = (\boldsymbol{f}_1 \circ ... \circ \boldsymbol{f}_K)(\mathbf{y}_0) = (\boldsymbol{f}_1 \circ ... \circ \boldsymbol{f}_K)(\mathbf{x})$$

the log-likelihood of samples $\mathbf{x}$ will be used as the loss function:

$$L(\mathbf{x}) = \log p_X(\mathbf{x}) = \log p_Y(\mathbf{y}) + \sum_{k=1}^{K} \log \left| \mathbf{det} \frac{\partial \boldsymbol{f}_i(\mathbf{y}_{k-1})}{\partial \mathbf{y}_{k-1}^T} \right|$$

The parameters of the functions $\boldsymbol{f}_i$ will be optimized by minimization of the loss function, through classical methods such as gradient descent.

The invertibility and bijectivity of the whole process, along with the usual simplicity of the functions $\boldsymbol{f}_i$, allows easy computation of the log-likelihood of samples $\mathbf{x}$, which allows the estimation of the probability density in the original space. It also allows easy sampling of $\mathbf{x} \sim p_X(\mathbf{x})$ by sampling from $\mathbf{y} \sim p_Y(\mathbf{y}) = \mathcal{N}(0, 1)$ and inverting the process, which can be used for generative models.

The main interest for density estimation is that it allows modeling complex non-parametric distributions, as the complexity emerges naturally from the composition of multiple simple functions.

I.3.1.iii Support estimation methods

One class support vector machines (OC-SVM)

One class SVMs have been introduced in Schölkopf et al. (2001). The end goal is to estimate the probability density support of the normal data $\mathbf{x}$.

To do so the normal data $\mathbf{x}$ is projected in a feature space[1] through a transformation $\boldsymbol{\Phi}(.)$, where it is separated from the origin with a maximum margin hyperplane of parameters $(\mathbf{w}, \rho)$.

[1]This step is not mandatory, but it is what makes this the method powerful as it allows generating highly non-linear support estimates.

In the original data space, this will amount to estimating the support of the $\mathbf{x}_i$, with a tolerance ν on the fraction of outlier present in the training data. The following problem is minimized to obtain the model's parameters :

$$\begin{aligned} \min_{\mathbf{w},\rho,\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\ \text{subject to} \quad & \langle \mathbf{w}, \boldsymbol{\Phi}(\mathbf{x}_i)\rangle \geq \rho - \xi_i \qquad i \in [1,n] \\ & \xi_i \geq 0 \qquad\qquad\qquad\quad i \in [1,n] \end{aligned}$$

After obtaining the optimal parameters, the decision function $f(\mathbf{x})$ can be retrieved and used to determine if sample $\mathbf{x}$ is normal ($f(\mathbf{x}) \geq 0$) or anomalous ($f(\mathbf{x}) < 0$). We will describe in more detail the functioning of OC-SVM later in section II.3.2.

Support vector data description (SVDD)

Support vector data description Tax and Duin (2004) bears the same principle as one class SVM, perhaps in a more intuitive way. Instead of finding a hyperplane that puts the data points on the right side, the goal is to find a ball of center $\mathbf{c}$ and radius R enclosing the normal data. This amounts to solving the following problem:

$$\begin{aligned} \min_{R,\mathbf{c},\xi} \quad & R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & ||\boldsymbol{\Phi}(\mathbf{x}_i) - \mathbf{c}||^2 \geq R^2 + \xi_i \quad i \in [1,n] \\ & \xi_i \geq 0 \qquad\qquad\qquad\quad i \in [1,n] \end{aligned}$$

We will describe in more detail later the functioning of SVDD in section IV.2.1.ii.

Deep SVDD

Ruff et al. (2018) proposed deep SVDD, where they replace the kernel mapping $\Phi(.)$ to an explicit, learned neural network mapping $\boldsymbol{\Phi}_W(.)$. They also simplify the problem of finding an enclosing ball, to only computing the distance to the center $\mathbf{c}$, which is set to be the barycenter of the representations of the training set, after mapping through the network W before it is trained (thus $\mathbf{c}$ is not learned). The problem then amounts to training the network to push points close to the center $\mathbf{c}$. An anomaly score is then naturally given by the distance of a point to the center.

We see that this method is a relaxation of a support estimation method in several ways. First, the concept of enclosing ball and its associated radius is removed, and only the distance to the center is estimated. Second, the center $\mathbf{c}$ is not learned during training, thus we could say that this modified method could be classified as a distance-based method (see section I.1.5). Third, they make no use of the kernel trick (see paragraph II.3.2:Dual problem), as the mapping to the latent space is learned explicitly.

I.3.1.iv Hybrid methods

There is a family of methods that combine density estimation and reconstruction, called restoration methods. In restoration, one usually tries to estimate the density of the normal distribution, uses this estimated density to resample the abnormal samples, and then compares this 'restored'/'healed' version $\hat{\mathbf{x}}$ to the original image $\mathbf{x}$. Because it compares a reconstructed/restored version of the image to the original one, we could argue that restoration is a reconstruction

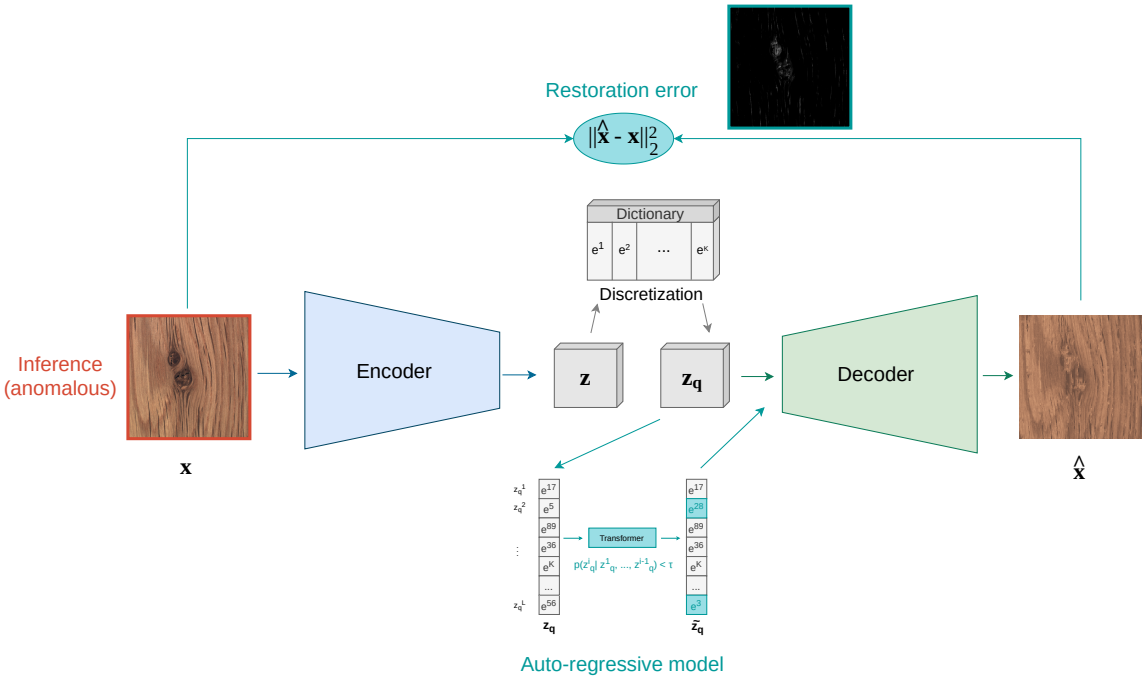

Figure I.19: Inference of a VQ-VAE + Transformer model. The auto-regressive model has been trained on latent representations of normal samples, prior to the inference.

method, but it uses density estimation for resampling. Usually, the anomaly is not estimated directly with the probability of the estimated density because the resampling operation usually allows fewer false positives, e.g. when erroneously resampling a healthy image patch, one could restore this patch into a patch that is similar to the original patch.

Some methods that we will see hereafter, combine density estimation/support estimation/reconstruction. In our opinion, these methods do not constitute hybrid methods, as they are either a concatenation of estimation/support estimation/reconstruction methods (one after the other) or a weighted sum of two or three of those methods, and thus not 'truly' hybrid.

VQ-VAE + autoregressive model

A possible way to perform restoration is to use a combination of VQ-VAE and auto-regressive models (such as PixelSnail Chen et al. (2018) or the very popular Transformer Vaswani et al. (2017a)). The VQ-VAE (presented I.3.1.i:VQ-VAE) allows to get discrete latent representation of the input image, while the auto-regressive model allows to restore latent vector coordinates when given other coordinates. A depiction of this process is presented in figure I.19.

More formally, once the latent vector is quantized (after encoding of the image), we model the probability of occurrence of a latent vector coordinate z^j conditioned by the previous coordinates, i.e. $p(z^j|z^1, ..., z^{j-1})$. This allows having a probability of occurrence of every z^j and allows resampling of the less probable samples during inference. After the resampling, the vector z is passed through the decoder to obtain a 'healed'/restored version of the anomalous image.

This auto-regressive modeling would not have been possible without the finite quantization provided by the VQ-VAE. Also note that predicting a coordinate in an image based on previous coordinates implies choosing a specific ordering (raster, zig-zag, etc.), this ordering can affect the performances but most importantly: different ordering can be combined and average to obtain a better restoration, at the cost of additional computational time.

Density estimation through auto-regressive models allows for detection directly in the latent space (density estimation) or in the image space through restoration (a hybrid of reconstruction and density estimation) or a combination of both (e.g. Wang et al. 2020).

I.3.1.v Partial state-of-the-art UAD for MVTecAD

As stated before, the MVTec AD database, presented in section I.2.2.i, by becoming essential to the domain of computer vision anomaly detection, has sped up the emergence of methodological developments in anomaly detection, by allowing the comparison of different methods on a common ground.

As a large number of recent developments have been evaluated on this database and as we will also use it for evaluation (section III.1), we wish to present some of the methods obtaining state-of-the-art performances on MVTecAD. Part of these methods have also been translated to the medical image analysis community.

Reconstruction methods

Bergmann et al. (2021) have shown the performances of auto-encoders to detect the anomalies on the MVTec AD datasets. The auto-encoder takes as input whole images from the training set (normal), is trained to minimize the reconstruction error, and at inference time when presented with images that contain anomalies, the pixels containing the anomalies exhibit larger reconstruction error.

Bergmann et al. (2021) have also applied the same strategy but replaced the L^2 norm in the training and inference with the structural similarity (*SSIM*, Wang et al. 2004). The *SSIM* is supposed to give a metric closer to the perceived quality than the pixel-wise L^2 norm. It has the drawback of working by patch (one must select window size and compute the *SSIM* over each window).

Density estimation

Auto-encoder can be used as an anomaly detection method (reconstruction method). Still, the encoder, by allowing the construction of a rich latent space (a large amount of information contained in a 'small' space) can also be used as a feature extractor (the latent space is sometimes called the feature space) as a first step for other methods. This is also true for neural networks (CNN for vision) trained for other tasks (such as classification for example)

Defard et al. (2021) used such a pre-trained CNN to extract features of patches of the whole image. For each patch, they used a concatenation of features of different layers of the CNN. Then for each position in the image, they estimate the density of the features by assuming it is generated by a multivariate Gaussian distribution and estimating its mean vector and covariance matrix, thus one estimation for each position in the image. Therefore, this procedure is a density estimation method for each patch position in the image, after extracting its feature through a CNN.

Yu et al. (2021) used the same principle: extracting features with a pre-trained CNN network or vision transformer Dosovitskiy et al. (2021) and then using Normalizing Flows (I.3.1.ii:Normalizing Flows) to estimate the probability density of the normal class. The outliers are then detected as having low probability.

Support estimation

Yi and Yoon (2021), first evaluated the deep SVDD (I.3.1.iii:Deep SVDD) method on MVTec and argued that the task was too difficult to learn a mapping of the whole image to a feature space able to detect anomalies. They thus proposed to work with patches, and instead of selecting one center (or multiple centers), they proposed to use siamese neural networks (Bromley et al. 1993) to enforce adjacent (in the image) patches to be close in the feature space. They then choose the anomaly score to be the distance to the closest training patch.

In the end, we believe that the proposed method is quite far from the SVDD technique, as the projection is not kernel-based, there is no center nor radius learned, and the anomaly score is derived only from the distance of the closest training sample. Arguably, this could be categorized as a distance-based method (thus a density estimation method).

Hybrid methods

Wang et al. (2020) used the approach presented in paragraph I.3.1.iv:VQ-VAE + autoregressive model, a VQ-VAE allowing discrete representation learning, combined with an autoregressive model (PixelSnail Chen et al. 2018) to restore the latent vectors, that are then decoded into 'healed' version of the image, compared with the input image to highlight anomalies.

I.3.2 State-of-the-art UAD in medical image analysis

Now that we have introduced what we believe are the fundamental methods and building blocks of unsupervised anomaly detection, we wish to perform a more exhaustive literature review, on UAD used for medical imaging. We believe the specific characteristics of medical imaging, such as limited databases, rare pathologies, difficult evaluation due to time-costly labels, etc. justify the need for a specific literature on the topic. Also, tackling real and challenging datasets such as medical images, can drive upstream fundamental research on UAD.

We once again rely on the classification developed by Ruff et al. (2021), which we will follow in this section. To perform this literature review, we exhaustively searched for every journal/-conference article containing the keywords 'anomaly detection', 'outlier detection', 'novelty detection', 'out of distribution detection', and 'medical'. We discarded articles where the anomaly detection was weakly/semi or completely supervised. We added some articles present in the bibliography of the articles we reviewed, which sometimes use different terminology. We also added every article matching the specification that we found in Ruff et al. (2021), but also in these two reviews: Fernando et al. (2021), Tschuchnig and Gadermayr (2022). We found an article dating back to 1995, but none prior. There may be articles predating this, but the term 'anomaly detection' was coined recently, making it difficult to find such articles.

This literature review is summarized at the end of the chapter, table I.1. For each method, we give a brief summary of the method, the features that are used for UAD, the databases used for training/testing, whether the method is machine learning/deep learning, density estimation/support estimation/reconstruction/others, the metrics reported, and if the evaluation is done at image-level or voxel level.

I.3.2.i Density estimation methods

Non-parametric models

Tarassenko (1995) were interested in the problem of detecting masses in mammography. After a pre-processing consisting of removing the curvilinear structures in the images and extracting

bounded contours with an in-house algorithm, they extracted classical image-processing features like standard deviation, edge gradient of the perimeter, or ratio of volume to area to obtain a feature vector $\mathbf{x}$ of dimension 5 for each bounded contour. They then used a Parzen-Rosenblatt window estimator (presented paragraph I.3.1.ii:Kernel density estimation) to estimate the probability density of the normal distribution. Once the distribution is estimated, they set a threshold of 5% false positive rate on a validation dataset to determine the threshold below which $\hat{p}(\mathbf{x})$ will be considered anomalous. They have evaluated 40 masses of the Mammographic Image Analysis Society digital mammogram database. Note here that they consider the whole distribution of data $p(\mathbf{x})$ to include both the normal class $p^{\odot}(\mathbf{x})$ and the outliers $p^{\oslash}(\mathbf{x})$, which would be highlighted by the KDE.

Prastawa (2004) also used a KDE model for brain tumor or edemadetection, as a part of a larger pipeline, where each voxel of a tissue type (Grey matter, cerebrospinal fluid, etc.) was used as a sample to estimate the probability density of the tissue type.

Parametric models

Van Leemput et al. (2001) and Arnaud et al. (2018a) each used a parametric estimation model, for multiple sclerosis detection in humans and brain tumor localization and characterization, respectively, both in MRI. They modeled voxels of a multi-channel MRI as a mixture model: each voxel probability density $p(\mathbf{x})$ (which has the dimension of the number of MRI channels) is modeled by the mixture distribution (presented paragraph I.3.1.ii:Mixture models).

In a medical imaging setting, the mixture weight typically represents affiliation to a tissue type, e.g. white matter, cerebrospinal fluid (CSF), etc. The distribution $f(\mathbf{x}; \Phi_k)$ is chosen to be Gaussian in Van Leemput et al. (2001) and a multiple scale t-distribution (Forbes and Wraith 2014) in Arnaud et al. (2018a), which generalizes the Gaussian multivariate distribution. An important difference between the two studies is that Van Leemput et al. (2001) fit the model on each MR image and determine the anomalies as voxels with low probabilities (with post-processing and robust-statistics methods specific to their applications), whereas Arnaud et al. (2018a) fit the model on a group of training healthy controls, and then apply it to test patients and controls.

Bowles et al. (2017) also used a Gaussian mixture model consisting of two components (one for white matter and one for grey matter) to assess the likelihood of a voxel belonging to the correct class and thus being abnormal or not. An important and non-trivial difference between these studies is the choice of the number of mixtures (K), which is sometimes fixed with *a priori* knowledge (e.g. number of tissue type) or sometimes with an information theory criteria such as the BIC (Schwarz 1978). This was used for unsupervised brain lesion segmentation in MRI.

Li et al. (2015) used a parametric Gaussian model (equivalently a mixture model with $K = 1$) to model hyperspectral image pixels for swine burn, and proved that a classifier's performances were improved when estimated anomalies were first removed from the training dataset. They applied this method to perform diagnosis on multispectral swine burn images.

Prastawa (2004), as a part of a larger pipeline mentioned above, used robust statistics to estimate the mean and covariance of healthy tissue classes, this amounts to assuming the data is composed of a mixture of Gaussian and thus is a parametric estimation method. In this work and others, a common step is to register a healthy atlas to the image being evaluated, thus obtaining priors on which tissue the voxel should belong.

Zimmerer et al. (2019) have used, among other criteria later described, a KL divergence term (presented section I.3.1.i:Variational auto-encoders), derived with respect to the input pixel, to obtain a 'pixel-wise KL divergence gradient map' to detect brain tumors.

Arguably, in our opinion, the KL divergence term constitutes a density estimation term, as the hypothesis for deriving the KL divergence term usually includes that the two distributions have simple parametric forms, under these hypotheses, one can use the KL divergence as a distance measure between the two distributions. Without these density modeling hypotheses, one could not derive the KL term.

Normalizing flows[1]

Zhao et al. (2022b) have used the same principle as Yu et al. (2021) (presented in paragraph I.3.1.ii:Normalizing Flows), i.e. extracting features with a convolutional neural network (encoder) and using normalizing flow for probability density estimation on these features for anomaly detection in brain MRI, chest x-ray and others. An interesting twist in Zhao et al. (2022b) is that the feature extractor is not fixed but trained alongside the normalizing flow, meaning the extracted features will be more relevant for the NF estimation. A decoder is also added, after normalization by the flow, to obtain reconstruction and pixel-wise error maps. Normalizing flows are arguably not parametric models, because they are supposed to model the input distribution $p(\mathbf{x})$, with no *a priori* on its shape.

Ciușdel et al. (2022) have also used a normalizing flow, without feature extraction, to detect synthetic abnormalities in the segmentation of the coronary artery lumen on coronary computed tomography angiography. They did not require feature extraction because they worked with small patch sizes (as normalizing flow networks cannot downsample images, due to the invertibility condition).

I.3.2.ii Support estimation methods

El Azami et al. (2016) used a one-class support vector machine (OC-SVM) to estimate the support of the distribution of individual voxels in maps obtained from white matter, gray matter, and cerebrospinal fluid (CSF) segmentation maps. This technique, along with a KDE of the anomaly score and other post-processing techniques, was used to detect patients with epileptogenic lesions in brain MRI.

Alaverdyan et al. (2020) have studied the same problem of anomaly detection for epileptogenic lesions, and used the same OC-SVM technique but the feature extracted were automatically learned by a patch-based auto-encoder, i.e. a patch that takes as input small patches. This method will be detailed in II.3.

As part of a larger pipeline, Bowles et al. (2017) used 2 OC-SVM models to highlight outliers, one for white matter and one for grey matter, used for unsupervised brain lesion segmentation in MRI. They take as input FLAIR voxels, synthesized FLAIR voxels, and partial volume probabilistic voxels as input, with the end goal of segmenting brain lesions.

To the best of our knowledge, El Azami et al. (2016), Alaverdyan et al. (2020), and Bowles et al. (2017) are the only methods that used support estimation for medical anomaly detection. They differ in the way they extract features, on the pre/post-processing but each used OC-SVM as the support estimator.

[1] We believe normalizing flows fall under the category of non-parametric models (they do not impose any shape on the estimated $p(\mathbf{x})$), but their importance in the literature justifies its own paragraph.

I.3.2.iii Reconstruction methods

Auto-encoders

Baur et al. (2021a) have done a comparative study of many auto-encoder-based methods, including VAE, all based on the residual error between the input image $\mathbf{x}$ and its reconstructed output $\hat{\mathbf{x}}$. They used a common architecture and evaluated the performances on in-house databases and on MSSEG (Commowick et al. 2018) and MSLUB (Lesjak et al. 2018) which are two publicly available brain MRI datasets containing patients with multiple sclerosis datasets. To begin with, they used the classical L^2 reconstruction error presented in paragraph :Auto-encoders.

Behrendt et al. (2022) and Luo et al. (2023) also used auto-encoders and L^2 or L^1 reconstruction error for an anomaly detection task (brain tumor segmentation on BraTS Menze et al. 2015).

Baur et al. (2021b) also used a tweaked version of this kind of auto-encoder, with some skip connections between the lower levels of encoder and decoder, making it a hybrid between the widely used U-net (presented paragraph I.3.1.i:U-net) and an auto-encoder. They used this network to detect white matter hyperintensities in brain MRI (WMH challenge Kuijf et al. 2019).

The same authors have used in Baur et al. (2020) an auto-encoder that reconstructs, instead of the whole image, successive high-frequency residuals obtained after smoothing, downsampling, and subtracting the original image. The main hypothesis of such a method is that the reconstruction of the high-frequency components of the anomaly will not be possible (e.g. it might be possible to reconstruct the hyperintense signal of a white matter hyperintensity but not its precise contours). This was applied to multiple sclerosis detection in MRI.

Baur et al. (2021a) have also studied Bayesian auto-encoders, where dropout (Hinton et al. 2012, Hertz (1991)), the process of shutting down randomly neurons of the neural network both at training and at inference, is applied. Applying different random dropout at inference allows to get multiple maps, and thus with the mean map get a more confident estimate. This, as hereabove, was applied to multiple sclerosis detection in MRI.

Seeböck et al. (2019) have used the standard deviation of sampled maps (from U-net-like architecture with dropout) to obtain regions where the prediction is uncertain, they used these regions, as a part of a more complex pipeline, to detect anomalies in retinal optical coherence tomography images.

Muñoz-Ramírez et al. (2020) have used VAE to detect anomalies in Parkinson patients' brain MRI, where they showed that more anomalies were detected in Parkinson patients than controls (PPMI database Marek et al. 2018).

Zimmerer et al. (2019) and Zhao et al. (2022b), among other anomy scoring methods, have used the reconstruction error from a VAE for brain tumor segmentation in MRI.

Adversarial models

Baur et al. (2021a) have also studied adversarial auto-encoder (AAE Makhzani et al. 2015) in their review. An AAE is very similar to a VAE, the key difference is that the term that matches the posterior distribution (the KL divergence term in VAE) is learned in an adversarial way: a discriminator network must learn to classify samples generated from the standard normal distribution from samples projected in the latent space by the auto-encoder. This adversarial training is very similar to the one of (GAN Goodfellow et al. 2014) that we will describe hereunder. This, as hereabove, was applied to multiple sclerosis detection in MRI.

Chen and Konukoglu (2018) have used VAE and AAE to study brain tumor detection in MRI with an added constraint on the latent representation, the reconstructed image $\hat{\mathbf{x}}$ must be projected close to the original latent representation $\mathbf{z}$, i.e. with $\mathbf{E}$ the encoder and $\mathbf{D}$ the decoder: $||\mathbf{E}(\mathbf{x}) - \mathbf{E}(\hat{\mathbf{x}})|| = ||\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{D}(\mathbf{E}(\mathbf{x})||$ must also be minimized. They argue that with AE, consistency in image space is obtained by minimizing $||\mathbf{x} - \hat{\mathbf{x}}||$ but so should be its latent representation.

Bercea et al. (2023b) have used soft-Intro VAE (Daniel and Tamar 2021), a VAE variation where the encoder and decoder are trained in an adversarial way, such that the encoder will learn to differentiate between generated and real images. They also add similarity metrics at the multiple layer levels of the encoder, to match the features of the latent representation of real image $\mathbf{x}$ and the feature of the latent representation of the reconstructed image $\hat{\mathbf{x}}$, like Chen and Konukoglu (2018). They evaluate their framework on multiple synthetic and real-world anomalies/pathologies of brain MRI.

When interested in anomaly detection, one desirable property is to be able to project an image into a latent space (feature space) where one can measure the distance to the normality[1]. This ability is granted by the encoder in a VAE framework but is missing in the GAN approach. This is resolved by the proposition of Adversarial Auto-encoders (Makhzani et al. 2015) described before.

Another proposed approach for solving this problem is Schlegl et al. (2019), which proposes f-AnoGAN: at first, a classical GAN is trained, yielding a generator/decoder $\mathbf{D}$ and a discriminator C. Secondly, the generator and discriminator weights are frozen, and the following loss is minimized to train the encoder $\mathbf{E}$:

$$\min_{\mathbf{E}} \; L(\mathbf{x}) = ||\mathbf{x} - \mathbf{D}(\mathbf{E}(\mathbf{x}))||_2^2 + \beta||\boldsymbol{f}(\mathbf{x}) - \boldsymbol{f}(\mathbf{D}(\mathbf{E}(\mathbf{x})))||_2^2$$

where in the first term we recognize the reconstruction error, β is a loss weight factor, and f is an intermediate feature map of the discriminator C. The authors found that using the second term improved the performance for their task, as it is important to match the reconstructed image $\hat{\mathbf{x}} = \mathbf{D}(\mathbf{E}(\mathbf{x}))$ to the original image $\mathbf{x}$ in image space but also in the feature space provided by the discriminator. This technique reminds the consistency loss term added by Chen and Konukoglu (2018), except it is in the discriminator feature space.

At inference, for voxel-level anomaly detection, the L^1 reconstruction error (sum of absolute difference) between $\hat{\mathbf{x}}$ and $\mathbf{x}$, $||\mathbf{x} - \mathbf{D}(\mathbf{E}(\mathbf{x}))||_1$ is used. The authors (Schlegl et al. 2019) used f-AnoGAN for anomaly detection in spectral-domain optical coherence tomography images.

Tian et al. (2021) also used a fine-tuned f-AnoGAN, among other methods, on top of a pre-trained encoder for anomaly detection in optical coherence tomography images, coloscopy images, fundus image and gastrointestinal images.

f-AnoGAN has also been successfully applied to the MVTec AD dataset by Bergmann et al. (2021). As U-net (Ronneberger et al. 2015), this transfer occurs from medical imaging to computer vision.

[1] One could argue that the discriminator of the GAN can be used for anomaly detection, which is true, but discrimination is done at the image level, with seemingly no way of extracting which image pixels belongs to the anomaly.

Denoising models

Kascenas et al. (2022a) have used a U-net model to denoise brain MR images, the U-net is trained to denoise multi-scale Gaussian noise during training and at inference when presented with brain tumor images (BraTS Menze et al. (2015)) is able to erase them.

Wyatt et al. (2022) have used such models with simplex noise, arguing that simplex noise is more structured than Gaussian noise and as such, should be able to 'repair' anomalous zones. They evaluated their methods on a public brain tumor MR image dataset (Pernet et al. 2016).

Behrendt et al. (2023) have used a similar method but proceeded to split the image into 4 quarters and denoise each of them separately, given the unperturbed other 3 patches, before stitching them to recover a full image. They evaluated performances on a brain tumor segmentation benchmark (Menze et al. (2015)).

Bercea et al. (2023a) have used a similar framework, where diffusion models were used in a more complex fashion with masking and stitching, and applied the method for ischemic stroke segmentation.

Reconstruction by regression

As part of a larger pipeline used for brain lesion segmentation, Bowles et al. (2017) have used a regression model to transform T1 MRI volume to FLAIR volumes, and then compared the synthesized FLAIR to the original FLAIR. Arguably, this also constitutes a reconstruction anomaly detection method. This is also notably the only reconstruction method we found not using deep learning models.

Bieder et al. (2022) have used a network that takes as input 3D patches and outputs the coordinates of the patch. This coordinate regression allows the detection of outliers (fractures and hemorrhages on brain CT scans) as the patches where the coordinate is poorly retrieved. This amounts to a different form of reconstruction from the ones presented before, not based on the reconstruction of the input volume voxel's intensity.

I.3.2.iv Hybrid methods

Restoration on quantized latent space

Pinaya et al. (2022b) have used a similar technique to Wang et al. (2020) (presented paragraph I.3.1.iv:VQ-VAE + autoregressive model), namely, a VQ-VAE is used for latent representations of brain MR images, then a Transformer-decoder Vaswani et al. (2017a) is used to predict the latent coordinates of the latent samples, which amounts to discrete density estimation, and allows resampling of the latent vectors, which then pass through the decoder to obtain a restored image. They also upsample the resampling mask (coordinates of the latent vector that have been resampled) to the image size to weigh the reconstruction error and give more importance to resampled zones. They evaluated their method on brain white matter hyperintensities (WMH Kuijf et al. 2019).

Marimont and Tarroni (2021b) used the same kind of method, with PixelSnail instead of the Transformer (like Wang et al. 2020) and evaluated their method on synthetic brain MRI anomalies.

Pinaya et al. (2022a), for hyperintense lesion detection, also used a VQ-VAE for latent representation learning, but then used a diffusion model (the same as in Wyatt et al. (2022)) to noise and denoise the latent vector obtained by the VQ-VAE. In this process, they estimate

at each denoising step of the noised latent vector, which parts have changed and use this mask to only restore, using the denoising process, the suspected parts (which we can call inpainting with diffusion models used to generate the mask and doing the inpainting).

Restoration on continuous latent space

Marimont and Tarroni (2021a) used a restoration technique for brain tumor detection in MRI, where the latent distribution of healthy images is learned implicitly during training with an auto-decoder (no encoder), by sampling a random latent vector, associating it with a voxel coordinate and optimizing its match to the true image after decoding. At inference time, an optimization is carried out to retrieve the latent vector that most matches the pathological image. Because the implicit distribution that is learned is the healthy control one, the optimization amounts to a restoration of the image. The difference between the original image and restoration is then evaluated. One key difference with the previous methods is that here the latent space is not quantized and thus an optimization procedure has to be carried out to retrieve the 'healed' version of a sample.

Chen et al. (2020) and Chen et al. (2021) have also performed restoration based on an optimization process to detect brain tumors or lesions after stroke. They used VAE or Gaussian Mixture VAE as the representation learning method, trained on healthy controls only. At inference time, they propose to use gradient descent to maximize the maximum *a posteriori* of a statistical model that treats the anomalous image as a normal image plus some noise, supposed to be the anomaly. Approximation of the gradient differs in Chen et al. (2020) or Chen et al. (2021).

I.3.2.v Other methods

In this section, we will present methods for unsupervised anomaly detection in medical imaging that do not fall under any of the three categories presented above.

Synthetic anomalies

In Tan et al. (2022) and Kascenas et al. (2022b) the authors both used supervised deep learning techniques, to train a network to discriminate between real image patches and synthetic anomalies.

Tan et al. (2022) have incorporated synthetic anomalies as foreign patches: a patch from another healthy sample is blended with the original patch with a factor α, and the network has to output the blending factor for each image pixel. At inference, high blending factors naturally constitute the anomaly map. This was used for synthetic anomalies detection on brain MRI and chest CT-scans.

Kascenas et al. (2022b) have used a similar technique, where a supervised network is trained to discriminate between the original patch and a negative pair, given the context of the surroundings of the patch. Negative pairs are created by using intensity and spatial transformations and by taking patches located elsewhere. This was used for brain tumor segmentation.

While in the end these constitute unsupervised anomaly detection methods, it can be argued that they break a core assumption of unsupervised anomaly detection that is: no *a priori* should be given about the anomalies that are to be seen, as it cannot be known what anomalies will look like at inference, one should not impose prior knowledge.

However, the task of anomaly detection is fundamentally challenging without incorporating *a priori*, and one could argue that any choice made in the engineering method of UAD (not

supervised) could be thought of as a form of *a priori* knowledge incorporation (also called inductive biases). In the end, these methods stand out from a methodological point of view but we think they are perfectly valid approaches to unsupervised anomaly detection.

Basic image processing

Meissen et al. (2021a) have used basic image-processing techniques, to highlight the fact that oftentimes in brain MR imaging, the lesions of interest are hyperintense, and thus can be highlighted by thresholding the pixel intensity. They only used histogram equalization to ensure that the same threshold could be shared among MRI scans and evaluated on brain tumors, brain white matter hyperintensities, and multiple sclerosis lesions in MRI. This was applied to hyperintense segmentations on a wide range of databases (WMH, BraTS, MSLUB, MSSEG).

I.3.2.vi Conclusion on UAD for medical image analysis

As the reader might have noticed, the density estimation methods presented in section I.3.2.i are the oldest (1995 to 2019 with the noticeable exception of 2 normalizing flow methods). They seem to have been the main approaches in medical anomaly detection for many years, whereas reconstruction-based methods presented in section I.3.2.iii have been more popular in recent years (2017 to 2023), and especially since the advent of deep learning. On the other hand, support estimation methods, presented in section I.3.2.ii seem to have been less explored than the other two families.

We will address in the next chapter why it would be of great interest to study support estimation methods in anomaly detection for medical imaging and why the task of anomaly detection is especially challenging in medical imaging. We present hereunder, a summary of the anomaly detection methods for medical imaging presented in this section. For each method, we give a brief summary of the method, the features that are used for UAD, the databases used for training/testing, whether the method is machine learning/deep learning, density estimation/-support estimation/reconstruction/others, the metrics reported, and if the evaluation is done at image-level or voxel level.

Authors	Method	Features	Database(s)	Density estimation / Support estimation / Reconstruction / Hybrid / Other	Deep learning / Machine learning / Image processing	Metrics reported	Image-level AD / Voxel-level AD
Pinaya et al. (2022b)	Restoration from VQ-VAE latent space with Transformer	Intensity of multi-channel voxels	Train on UKB 15k volumes with lowest lesional volume, test on WMH. Only 4 middle slices.	Hybrid (restoration)	Deep learning	Best Dice *AU PR*	Both
Baur et al. (2021b)	Reconstruction from AE with skip connections	Intensity of multi-channel voxels	Train on in-house healthy dataset, test on WMH (only 51/60)	Reconstruction	Deep learning	Best Dice *AU PR*	Voxel-level
Baur et al. (2021a)	Reconstruction from many varations of AE (VAE, Bayesian, AAE, fAnoGAN, etc.)	Intensity of multi-channel voxels	Train on in-house, test on in-house, another in-house and MSLUB and MSSEG	Reconstruction	Deep learning	*AU ROC* *AU PR* best Dice Others	Voxel-level
Wyatt et al. (2022)	Diffusion model with simplex noise	Intensity of multi-channel voxels	Train on NFBS dataset (healthy) and test on edimburg neuroimaging dataset (brain tumors, 22 patients) (4 slices containing tumor)	Reconstruction	Deep learning	Dice Precision Recall *AU ROC*	Both
Behrendt et al. (2023)	Diffusion model with quarter of image inpainting	Intensity of multi-channel voxels	Train on IXI and test on BraTS and MSLUB	Reconstruction	Deep learning	*AU PR* Best Dice others	Image-level
Behrendt et al. (2022)	Reconstruction from AE	Intensity of multi-channel voxels	Train on IXI and OASIS test on BraTS	Reconstruction	Deep learning	*AU ROC*	Image-level
Schlegl et al. (2019)	Reconstruction error from GAN and Encoder trained sequentially	Intensity of multi-channel voxels	Train/test on in-house 3D OCT volumes	Reconstruction	Deep learning	*AU ROC* *AU PR* Best Dice others	Both
Meissen et al. (2021a)	Binary thresholding after histogram equalization	Intensity of multi-channel voxels	No training because no NN, test on BraTS MSLUB, MSSEG, WMH. FLAIR images only	Other	Image processing	*AU ROC* *AU PR* Best Dice	Image-level
Chen et al. (2021)	Gradient descent to maximize statistical model of normal image + noise	Extracted from VAE or Gaussian mixture VAE	Train on CamCan and test on BraTS and ATLAS	Hybrid (restoration)	Deep learning	*AU ROC* *AU PR* Dice	Voxel-level

Authors	Method	Features	Database(s)	Density estimation / Support estimation / Reconstruction / Hybrid / Other	Deep learning / Machine learning / Image processing	Metrics reported	Image-level AD / Voxel-level AD
Arnaud et al. (2018b)	Mixture model of student distributions learned on quantitative MRI voxels	Intensity of multi-channel voxels	In house rat MRI dataset	Density estimation	Machine learning	Dice others	Voxel-level
Muñoz-Ramírez et al. (2020)	Reconstruction of VAE as input to binary classification	Intensity of multi-channel voxels	PPMI	Reconstruction	Deep learning	*AU ROC*	Image-level
Alaverdyan et al. (2020)	One-class SVM on features of siamese AE	Extracted from Siamese AE latent space	Train on in-house test on in-house	Support estimation	Deep learning + Machine learning	Sensitivity FPR	Voxel-level
El Azami et al. (2016)	One-class SVM on segmentation maps	Segmentation maps	Synthetic MRI anomalies + in-house epileptogenic patients	support estimation	Machine learning	*AU ROC* Sensitivity FPR	Both
Chen et al. (2020)	Gradient descent to maximize statistical model of normal image + noise	Extracted from VAE	Train on CamCan test on BraTS and ATLAS	Hybrid (restoration)	Deep learning	*AU ROC* Dice others	Voxel-level
Marimont and Tarroni (2021b)	Restoration from VQ-VAE latent space with PixelSnail	Intensity of multi-channel voxels	Train on mood train set test on mood valid set (4 images with big synthetic anomalies)	Hybrid (restoration)	Deep learning	*AU ROC* *AU PR* Dice	Both
Marimont and Tarroni (2021a)	Implicit learning of latent distribution with auto-decoder between coordinates and intensities	Intensity of mono-channel pixels + coordinates	Train on HCP test on BraTS	Hybrid (restoration)	Deep learning	*AU ROC* *AU PR* Dice others	Voxel-level
Tan et al. (2022)	Foreign patch blended with image, network regress the blending factor	Intensity of multi-channel voxels	Train on subset of MOOD Test on synthetic anomalies on subset of MOOD Train/test deeplesion (CT scans with lesions slices with no lesions used for train)	Other (supervised with fake anomalies)	Deep learning	*AUROC* best Dice *AUPR*	Both
Kascenas et al. (2022b)	Classification of true patch VS fake patch given the surrounding image	Features extracted from multi-scale CNN	Train on BraTS healthy slices test on BraTS	Other (supervised with fake anomalies)	Deep learning	*AU PR* Best Dice Sensitivity	Both

Authors	Method	Features	Database(s)	Density estimation / Support estimation / Reconstruction / Hybrid / Other	Deep learning / Machine learning / Image processing	Metrics reported	Image-level AD / Voxel-level AD
Kascenas et al. (2022a)	U-net trained to denoise noise added to controls during training. No noise added at inference	Intensity of multi-channel voxels	Train/test on BraTS, but for train only use slices that are healthy	Reconstruction	Deep learning	$AU\ PR$ Best Dice	Voxel-level
Luo et al. (2023)	Reconstruction error from AE	Intensity of multi-channel voxels	Train IXI test on BraTS and in-house	Reconstruction	Deep learning	AUROC AUPRC best Dice	Both
Bowles et al. (2017)	Combination of image synthesis Gaussian mixture models and one class SVM	Intensity of multi-channel voxels	In house dataset	Density estimaton + support estimation + Reconstruction	Machine learning	Dice ASSD precision recall	Voxel-level
Tarassenko (1995)	KDE estimation on classic image processing features	Classic image-processing features	24 in house mamograms	Density estimation	Machine learning	Sensitivity FPR	Image-level
Chen and Konukoglu (2018)	AAE and VAE with added latent space consistency	Intensity of single-channel voxels	Train on HCP test on BraTS	Reconstruction	Deep learning	$AU\ ROC$	Voxel-level
Seeböck et al. (2019)	Bayesian U-net to obtain uncertainty maps from standard deviation	Intensity of multi-channel voxels	in house OCT volumes	Reconstruction	Deep learning	Dice precision recall	Image-level lesion-level
Tian et al. (2021)	Pretrained image/patch-based encoder (contrastive learning) fine tuning of fAnoGAN or SSIM decoder on features	Intensity of multi-channel voxels	Train/test on 3 datasets : Hyper-Kvasir (gastrointestinal dataset), OCT (LAG), Liu et al colonoscopy datasets	Reconstruction	Deep learning	$AU\ ROC$ IoU	Both
Zimmerer et al. (2019)	VAE +derivative of KL div w.r.t. pixel to get "KL pixel wise map"	Intensity of multi-channel voxels	Fashionmnist, for medical Train on HCP and tested on BraTS	Reconstruction + Density estimation	Deep learning	$AU\ ROC$	Both
Baur et al. (2020)	Multiple AE operating on laplacian pyramid images (High frequency residuals)	Intensity of multi-channel voxels	Train on in-house, test on in-house, another in-house and MSLUB	Reconstruction	Deep learning	$AU\ PR$ Best dice	Both

Authors	Method	Features	Database(s)	Density estimation / Support estimation / Reconstruction / Hybrid / Other	Deep learning / Machine learning / Image processing	Metrics reported	Image-level AD / Voxel-level AD
Pinaya et al. (2022a)	Restoration with VQ-VAE + diffusion model	Intensity of single-channel voxels	Train/test medmnsit, train UKB FLAIR, test on UKB, WMH, MSLUB, BraTS	Hybrid (restoration)	Deep learning	Best dice $AU\ PR$	Both
Li et al. (2015)	Gaussian log-likelihood estimation	Intensity of multi-channel pixels (spectral data)	In-house classification task (supervised) tested with and without outlier removal	Density estimation	Machine learning	Accuracy of classification model	Image-level
Van Leemput et al. (2001)	GMM estimation (robust statistics) + post-processing with a priori infos on MS lesions	Intensity of multi-channel voxels	In house multiple-sclerosis patients (no controls, model supposed to be robust)	Density estimation	Machine learning	Varying threshold compared to expert lesion segmentation	Voxel-level
Prastawa (2004)	Robust covariance estimation	Intensity of multi-channel voxels	In-house dataset of 4 volumes with tumor or edema	Density estimation	Machine learning	IoU Haussdorf	Voxel-level
Bercea et al. (2023b)	Reconstruction error with soft introspective VAE (adversarial framework)	Intensity of single-channel voxels	Train on IXI + fastMRI+ test on fastMRI+	Reconstruction	Deep learning	$AU\ ROC$ $AU\ PR$	Both
Bercea et al. (2023a)	Diffusion models used for initial guess + resampling for restoration	Intensity of single-channel voxels	Train on IXI + fastMRI+ test on ATLASv2	Hybrid (restoration)	Deep learning	Best Dice $AU\ PR$	Both
Zhao et al. (2022b)	VAE+Normalizing flow in the bottleneck Anomaly score is reconstruction error + NF log-likelihood	Extracted from VAE	Multiple datasets including BraTS	Density estimation	Deep learning	$AU\ ROC$ Best Dice others	Image-level
Ciușdel et al. (2022)	Normalizing flow on small patches	Intensity of single-channel voxels	Two in-house datatsets one reserved for testing	Density estimation	Deep learning	$AU\ ROC$ Accuracy Precision others	Voxel-level

Table I.1: Summary of the UAD methods for medical image analysis presented in section I.3.2.

II | Problem Formulation

II.1 Introduction

In this chapter, we will formulate the problem we are trying to tackle. We will first attempt to outline the reasons why there is a need for literature on anomaly detection that is specific to medical imaging. To do so we will highlight the blind spots in the current literature and the scientific challenges it raises. Next, we will precise the problem we are trying to solve and describe a support-estimation-based pipeline that was previously studied and that will serve as one of the baselines for the following work. Finally, we will outline the contributions made in this thesis, how they relate to the previously established work, and how they can shed a little light on the literature's blind spots.

II.2 Challenges in UAD for medical imaging

Anomaly detection in medical imaging is still an emerging field of research: this is due to the fact that image processing is still a relatively young field of research, and anomaly detection, as a subpart of machine learning, has gained a lot of interest even more recently. Also, machine learning for computer vision, partly due to the emergence of deep learning, has gained a lot of attention since ∼2012, and by extension so has anomaly detection for images. As such, this emerging field of research is still being structured, and part of the research problem is to correctly cast the task.

Despite the vast amount of literature, there are still some blind spots we would like to highlight: some related to the databases and evaluation, and some directly related to the methods.

II.2.1 Challenges related to databases and evaluation

Difficulty of the task

As already seen in section I.2.2.ii, the BraTS dataset (Menze et al. 2015), is very commonly used for evaluation (13 out of 35 methods presented in section I.3.2, also summarized in table I.1). As we can see in figure II.1, this database comprises gliomas that are very large in volume, and easily visible, even to a non-specialist. Moreover, in the T2 MR image, they appear as hyperintense signals, and as such, a gross segmentation could be easily obtained by thresholding the voxels' intensity. Note that this challenge dataset initial task is to obtain the precise segmentation of the different tumoral tissues, which is a challenging task. This task is made easier when only the detection of whether a tumor is present or not is required, which is the common 'derived task' in anomaly detection for this database.

A problem of using such a database is that the derived task is relatively easy, as such comparing the performances of multiple algorithms on the task that they all succeeded quite well is not the most relevant, e.g. when comparing two algorithms, if they both found anomalies in the glioma, comparing performances amounts to evaluate what algorithm has a better-refined segmentation of such an anomaly. It is hard to see that this would translate into better clinical practice, as a clinician could be interested in detecting abnormalities that he or she would have missed but might not be interested in its precise segmentation (segmentation as a task is another research domain).

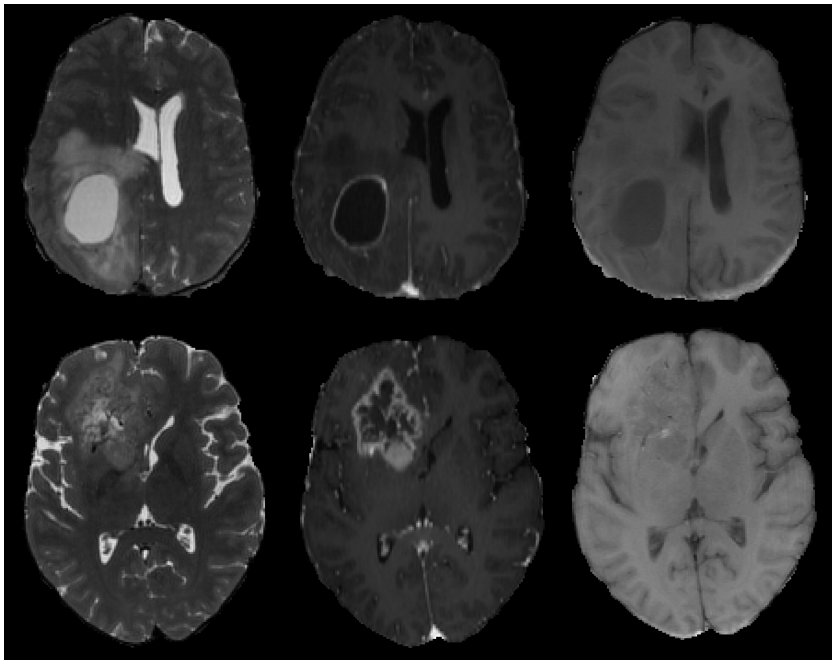

Figure II.1: Showcase slices of two patients (top and bottom) from the BraTS (Menze et al. 2015) database. T2 MR image (left), contrast-enhanced T1 image (center), and T1 image (right). The gliomas represent a large volume and are easily visible on each MR modality, especially in T2 and contrast-enhanced T1.

We therefore argue that datasets that contain anomalies that are too obvious (too large, too intense) are not good candidates for proper anomaly detection evaluation.

Hyperintense lesions

A related problem, that has been unveiled in Meissen et al. (2021a) and Meissen et al. (2021b), is present in the BraTS database, but not only, is that current literature often evaluates performances on anomalies that are much brighter than the rest of the image (hyperintense in MRI). Meissen et al. (2021a) achieved better performances than many state-of-the-art anomaly detection methods by only using the input MR (FLAIR) image as an anomaly score map, thus proving that by only thresholding the input image, one can highlight the anomalies. They further evaluate this behavior in Meissen et al. (2021b) where they show on synthetic experiments that there exists a 'blindspot' for anomaly detection, where if anomalies have the same intensity statistic as the rest of the dataset, they will be missed by classical reconstruction methods. They also show that texture is not an important feature when using these methods and that only the intensity matters.

Although the experimental findings made by Meissen et al. (2021b) are done with synthetic anomalies, we believe the conveyed message is accurate: 18 out of 35 of the methods presented (section I.2.2.ii or table I.1) evaluated their performances on a database containing only hyperintense anomalies. Thus, it remains to be proven that most of these methods work when presented with textured anomalies, or even hypointense anomalies.

Lack of proper ground truth

Another problem encountered in the literature is the databases that lack proper ground truth, for example, the PPMI (Marek et al. 2018) database used in previously presented works

only contains image databases with no proper ground truth. Another example is FastMRI+ (Zhao et al. 2022a) which contains image-level or bounding-boxes labels. This type of database makes the evaluation of pixel-level anomaly detection very difficult, only allowing for a proper evaluation of image-level anomaly detection. In the presented methods (section I.2.2.ii or table I.1) 8 out of 35 evaluate their performances only on image-level and 22 out of 35 consider image-level detection in their evaluation: it is therefore difficult to identify if the classification has been done due to correctly identifying the anomalies.

Despite the flaws of such databases, they still constitute interesting venues for the identification of potential new pathologies biomarkers[1], and often contained anomalies that are not hyperintensities as in PPMI (Marek et al. 2018) and FastMRI+ (Zhao et al. 2022a).

Single type of anomaly

A different common drawback present in some of the databases (Menze et al. 2015, Kuijf et al. 2019) used for evaluation is the presence of only one type of pathologies (respectively gliomas and white matter hyperintensities). The framework of anomaly detection allows for the detection of anything that deviates from the normality, and performance evaluation limited to a specific type of anomaly greatly increases the risk of overspecializing to this specific anomaly type, and thus poor generalization capability.

Domain-shift and confounding factors

Meissen et al. (2022) identified that domain shift (i.e. difference of data distribution between training set and testing set) might be a confounding factor of good performances in anomaly detection. This is especially true in their identified case where, on the Hyper-Kvasir dataset (Borgli et al. 2020), composed of gastro-intestinal videos where the lighting and camera angle are different when looking at an anomaly (polyp) because it is directed by the clinician.

On the other hand, confounding factors such as age can generate abnormalities such as brain shrinkage. For instance, if training on a control database of young healthy controls, and testing on patients that have Parkinson's disease, the brain shrinkage could be detected as anomalies but are not the anomalies of interest. Thus we argue that other anomalies, such as anatomical variabilities due to confounding factors, could negatively affect the performances.

Generalization capabilities

As pointed out in section I.2.2.ii, another obvious drawback of some databases used is the presence of synthetic anomalies only, these anomalies can be far from the true distribution and algorithms might not generalize well on real-world anomalies. Another generalization problem is the fairly common evaluation of only a subset of volume slices: this arguably overestimates the performance of the detection algorithms and hinders their clinical usability.

Conclusion

In the end, a lot of drawbacks of the databases used for the evaluation of anomaly detection tasks can be attributed to the fact that they are often conceived for other tasks than anomaly detection, and thus inherit the undesirable properties listed above. We believe that the blind spots described above are specific to the anomaly detection in medical imaging literature, and

[1] e.g. one could argue that if it is possible to discriminate Parkinson patient from healthy based on detected anomalies, these anomalies, if characterized, could constitute new biomarkers of the pathology, as in Muñoz-Ramírez et al. (2020).

therefore are important to address in this specific context, and could not be resolved or identified simply by looking at the anomaly detection for computer vision literature. A good example of this statement is that the MVTecAD (Bergmann et al. 2021) database used for computer-vision anomaly detection presents none of the weaknesses presented above.

II.2.2 Challenges related to the methods

Reconstruction methods

Meissen et al. (2021b) have pointed out that reconstruction-based methods were mostly hyperintensities detectors, as proven in another work by the same author (Meissen et al. 2021a) where they obtain better performances than state-of-the-art methods only by thresholding the input maps.

They also showed that when improving the ability of an auto-encoder to reconstruct the input images, anomalies were then also better reconstructed and that it decreased the anomaly detection performances. Baur et al. (2021a) had the same finding when studying auto-encoders. It seems that an auto-encoder either produces blurry reconstructions, and thus could miss the more subtle anomalies, or produce high-quality reconstructions, to the point that even anomalies never seen during training can be well reconstructed. As an example, this finding has been highlighted for computer vision in Tong et al. (2022), where they show the intuitive fact that for the MNIST database (LeCun et al. 1998), the 'all black' image, that is clearly an anomaly, can be very well reconstructed by an auto-encoder that has only seen classical digits during training.

Density estimation method

Density estimation methods (and restoration methods that use density estimation) have not shown such weaknesses, and are more interpretable as they output probabilities. But as stated before they solve a problem that is more general than the one we are trying to solve, i.e. the probabilistic model is regressed on all the data, but we only care about correct estimates of the classification frontier. Estimation of more complex models often leads to less robustness and more samples needed for correct estimation. Moreover, they often impose simplistic modeling assumptions, necessary for the tractability of the methods.

Support estimation methods

A third way would be the use of support estimation methods. We have seen in section I.3.2.ii that this third family has not yet been extensively studied for anomaly detection in medical imaging[1] and we propose to investigate this type of techniques as a way of solving the highlighted blind spots of the literature presented above.

II.3 One-class SVM on siamese auto-encoders latent space for anomaly detection

Now that we have presented the current challenges in UAD for medical imaging, we describe with more detail a support estimation method, proposed by Alaverdyan et al. (2020), as it will serve as a basis and a baseline to the following contributions chapters. This method is composed of multiple building blocks, two of which are of prime importance: first, the feature extraction step, and the outlier detection step. Figure II.2 summarizes the whole pipeline.

[1] In computer vision, although not demonstrated here, we found that density estimation and reconstruction methods have also been more studied than support estimation methods.

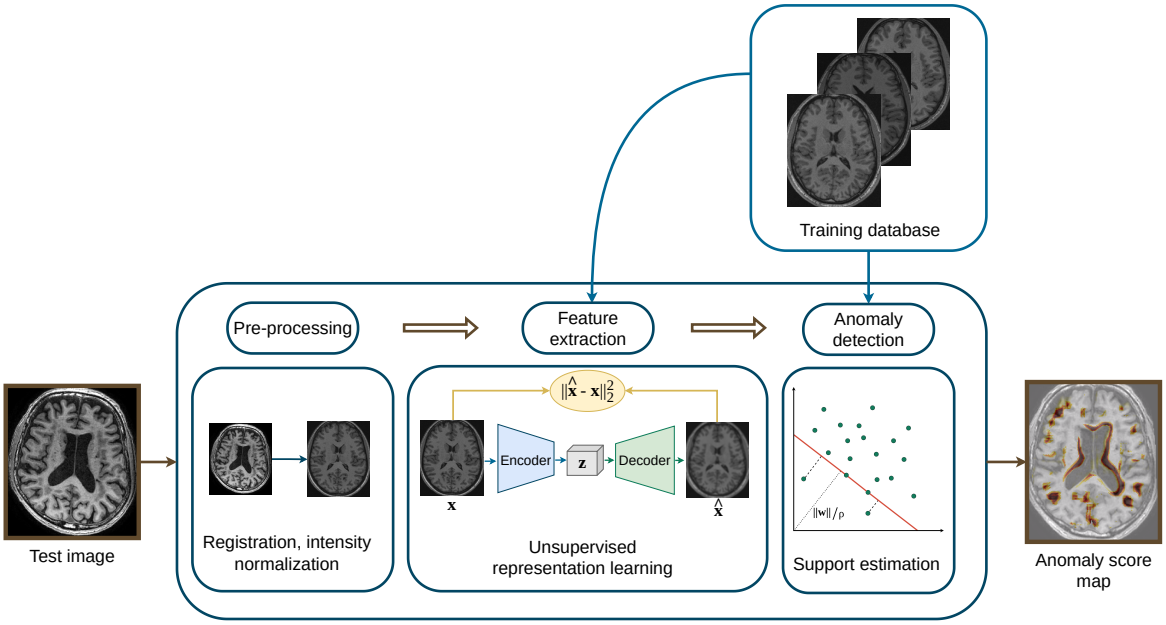

Figure II.2: Full pipeline of the method presented in Alaverdyan et al. (2020).

In the following sections (II.3.1 and II.3.2), we first detail the feature extraction step, which is done by training a siamese auto-encoder, and then detail the anomaly detection step, done by training a one-class SVM. The details of the integration of these two blocks into the whole pipeline and the pre-processing and post-processing steps are then described in section II.3.3. We close this section by showcasing the limits of this study (section II.3.5) which will introduce our contribution outline.

II.3.1 Siamese convolutional auto-encoder

We have seen in section I.3.1.i:Auto-encoders that auto-encoders can be used directly for anomaly detection, but also feature extraction: the auto-encoder is trained to reconstruct its input as a pretext task, and thus compressing the input information into a compact and rich latent code. This latent code can then be used for other tasks (e.g. anomaly detection).

Principle and uses

Siamese neural networks have been proposed by Bromley et al. (1993) for handwritten signature recognition. In their basic form, they only contain an encoder E and are called siamese because two representations $\mathbf{x}_1$ and $\mathbf{x}_2$, considered close in the input space, will be forced to have close latent representation $\mathbf{z_1}$ and $\mathbf{z_2}$ (with $\mathbf{z}_t = E(\mathbf{x_t})$). For example, two different handwritten signature images, but written signed by the same person, can be brought in the same neighborhood by minimization of the cosine similarity between $\mathbf{z_1}$ and $\mathbf{z_2}$ ($cos(\mathbf{z_1}, \mathbf{z_2}) = \frac{\mathbf{z_1} \cdot \mathbf{z_2}}{||\mathbf{z_1}||\_||\mathbf{z_2}||}$), this is what is done in Bromley et al. (1993).

To avoid the siamese network learning the trivial solution of mapping every input to the same latent code (known as collapsing), there is a need to introduce mechanisms to counter this behavior. One of which, called contrastive learning, is to map similar pairs to close locations in the latent space and to map dissimilar pairs to different locations (this is done for example in Chen and He (2021)).

Alaverdyan et al. (2020), which auto-encoder is presented in figure II.3, avoid collapsing by using a decoder and a reconstruction error term, as such, the encoder must learn to map similar

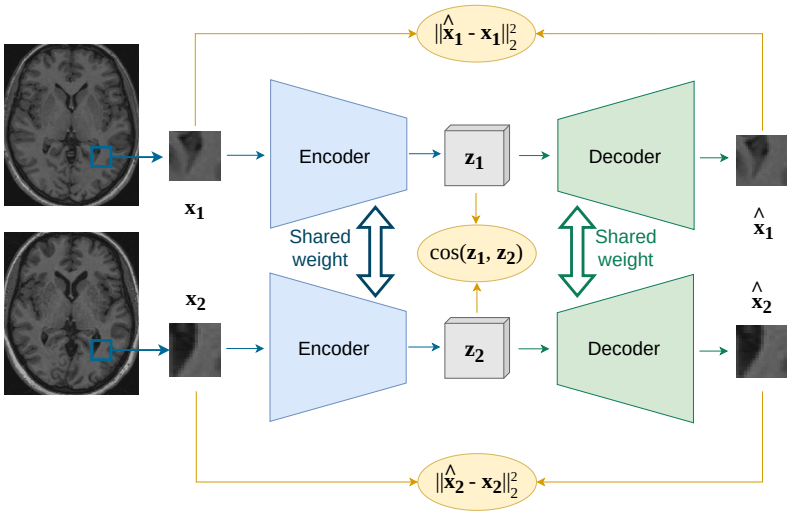

Figure II.3: Depiction of the siamese convolutional auto-encoder used in Alaverdyan et al. (2020). Patches of the same brain position but from different healthy controls are brought close in the latent space by the cosine similarity. Collapsing is avoided by the reconstruction error terms.

pairs to the same latent space neighborhood but still be able to use the latent code to provide a correct reconstruction. The loss function used is the following:

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

This term comprises the reconstruction loss for the pair and the cosine similarity between their latent representation[1]. We precise in section II.3.3 the notion of 'similar pairs', note that here no 'dissimilar pairs' are used.

This whole process can be viewed as two samples passing through a single encoder (and decoder if necessary), but can also be viewed, as two encoders sharing the same weights, each one processing one sample and bringing closer their latent codes, hence the term 'siamese'. This depiction is often more convenient to represent graphically (as in figure II.3) and is thus often used. Also note that the layers used siamese auto-encoders can vary, but when processing images are often composed of convolutions, as is commonly done for computer vision.

II.3.2 One-Class Support Vector Machine

The second building block we study is the support estimation step, which, strictly speaking, does the anomaly detection. In paragraph I.3.1.iii:One class support vector machines (OC-SVM), we have given a first insight into how one class SVM works. We will detail their functioning a little bit further.

General principle

Introduced in Schölkopf et al. (2001), and similar in principle to Tax and Duin (2004), one class SVM has the goal of estimating the support of the probability distribution that generated

[1]Note that any other similarity term could be used, such as the L^2 norm between z_1 and z_2. The cosine similarity is often preferred and has been used in the original article (Bromley et al. 1993), with no clear evidence of the benefits of such a choice.

samples $\mathbf{z}_i$ ($i \in [1, \ldots, n]$), or more generally, the minimum volume set of mass $1 - \alpha$ that will asymptotically contain a proportion $1 - \alpha$ of the samples.

To do so, the goal is to find a hyperplane that separates the data from the origin, in a feature space obtained by a transformation $\mathbf{\Phi}(.)$. A depiction of this problem is shown in figure II.4.

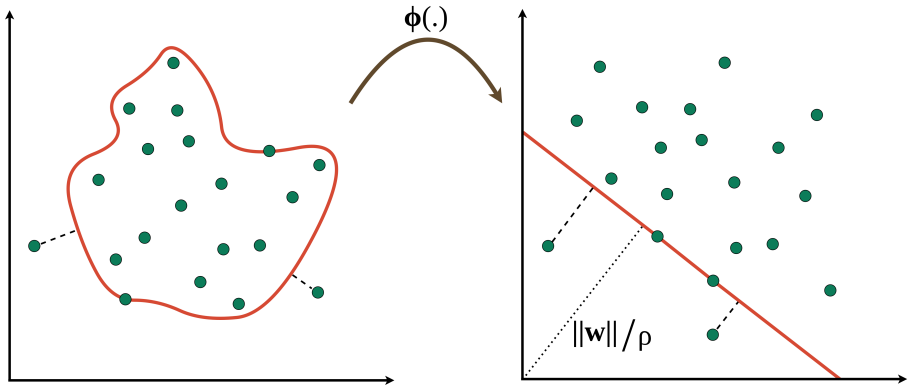

Figure II.4: Data points ($\mathbf{z}_i$, green), in the input space, being transformed by the implicit function $\mathbf{\Phi}(.)$ into a feature space where they can easily by separated from the origin by the maximum margin hyperplane ($(\mathbf{w}, \rho)$, in red). Some training data points are allowed to lie outside of the estimated support (corresponding to the wrong side of the hyperplane) thanks to the slack variables ξ_i.

The parameters to optimize are the vector normal to the hyperplane $\mathbf{w}$ and offset ρ, which completely defines the hyperplane. This convex optimization problem can be written as :

$$\begin{aligned} \min_{\mathbf{w},\rho} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} \quad & \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{z}_i) \rangle \geq \rho \quad i \in [1, n] \end{aligned} \tag{II.2}$$

We can see that in a classical support vector machine (Cortes and Vapnik 1995) fashion, the hyperplane to be found is the one that maximizes the margin to the origin ($||\mathbf{w}||$) under the constraint that the training points lie on the correct side of the hyperplane ($\langle \mathbf{w}, \mathbf{\Phi}(\mathbf{z}_i) \rangle \geq \rho$). At inference, for any new data point $\mathbf{z}$, one can characterize its signed distance to the hyperplane with the function:

$$f(z) = \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{z}) \rangle - \rho$$

This function will be negative if the point lies on the wrong side[1] of the hyperplane (if the point is an outlier and positive for the correct side (normal).

Relaxation

Because the training data can be polluted with outliers or simply noise, it is common to add the possibility for the training points to be located on the wrong side of the hyperplane, as long as they are not 'too distant'. This is materialized by the addition of 'slack variables' ξ_i, which allow miss-classification of training data points (i.e. allows violating the constraint, with the subtraction of ξ_i in the constraint), but penalizes this miss-classification (ξ_i is minimized in the objective):

[1] i.e. towards the origin.

$$
\begin{aligned}
\min_{\mathbf{w},\rho,\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\
\text{subject to} \quad & \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{z}_i)\rangle \geq \rho - \xi_i \qquad i \in [1,n] \\
& \xi_i \geq 0 \qquad\qquad\qquad\quad i \in [1,n]
\end{aligned}
\tag{II.3}
$$

Dual problem

Without entering into too many details[1], it is generally easier to solve a convex optimization problem by solving its dual. Using the Lagrange multipliers (the derivation can be found in Schölkopf et al. 2001, a similar derivation is shown in appendix A.1), we can show that the dual of the primal problem II.3 is the following problem:

$$
\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\mathbf{z}_i,\mathbf{z}_j) \\
\text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu n} \qquad i \in [1,n] \\
& \sum_{i=1}^{n}\alpha_i = 1
\end{aligned}
\tag{II.4}
$$

With the α_i being the dual variables to optimize (the Lagrange multipliers), and $k(.,.)$ the kernel function associated with the chosen reproducing kernel Hilbert space, meaning that when solving a kernel SVM problem, one must choose a kernel function $k(.,.)$ that will implicitly define the transformation $\mathbf{\Phi}(.)$. For rightfully chosen space[2] one can show that $\langle \mathbf{\Phi}(\mathbf{z}_i), \mathbf{\Phi}(\mathbf{z}_j)\rangle = k(\mathbf{z}_i,\mathbf{z}_j)$. By this trick, we can see that problem II.4 does not contain the transformation $\mathbf{\Phi}(.)$, and thus there is no need to know explicitly the transformation when solving through the dual. This is known as the kernel trick (Aizerman 1964). The reader will also notice that the problem II.4 is a quadratic program[3] and thus can be solved quite easily.

We can show[4] that the decision function can be rewritten as :

$$
f(\mathbf{z}) = \sum_{i=1}^{n}\alpha_i k(\mathbf{z}_i,\mathbf{z}) - \rho
$$

With $\rho = \sum_{i=1}^{n}\alpha_i k(\mathbf{z}_i,\mathbf{z}_l)$ for any l such that $0 < \alpha_l < \frac{1}{\nu n}$.

Note that in the end, this anomaly scoring function constitutes a distance to the hyperplane ($f : \mathcal{X} \to \mathbb{R}$) and not a proper probability estimate ($p : \mathcal{X} \to [0,1]$) of being an outlier.

Summary

We have seen that the estimation of the support of data points can be done with a convex optimization problem that maximizes the margin between data points and the origin. By looking at the dual problem and with the kernel trick, this can be done in a feature space, without knowing explicitly the transformation and solved efficiently as a quadratic program. Slack

[1] The interested reader will find great details about convex optimization in Boyd and Vandenberghe (2004)

[2] The interested reader can find more details about kernel functions in chapter 2 of Schölkopf and Smola (2002), and more mathematical details in general about one class SVM in chapter 8 of the same book.

[3] We refer again to Boyd and Vandenberghe (2004) for more details about convex optimization and quadratic programs.

[4] Again see Schölkopf and Smola (2002) for more details.

variables can also be added to relax the problem and allow data pollution while assuming a certain proportion of such pollution.

II.3.3 Implementation of Alaverdyan et al. (2020)

We now precise the whole pipeline[1] used by Alaverdyan et al. (2020) (presented figure II.2). First, all the subjects are registered to a common atlas (the registration procedure is presented in appendix B.1). The voxel intensities are normalized by removing the top 1% intensities and scaling the images to [0, 1].

II.3.3.i Feature extraction with siamese auto-encoder

For the siamese auto-encoder, the authors have used patches instead of images. First, this allows for recovering an anomaly score for each voxel of the image. Second, they argue that siamese neural networks have proven useful in the case where the number of classes (number of positions in the brain) is superior to the number of samples (number of healthy controls)[2].

The similarity optimized is the location in the brain, i.e. patches from different subjects, that are located at the same brain position will be brought close by the cosine similarity (this is depicted in figure II.3). Thus, the siamese auto-encoder will have to balance two goals: reconstruction of the input patches, and similarity in the latent space of same-localization patches.

A drawback of such a method is the need to pair brain positions, done by registration to a common atlas, which is in itself a very difficult task, making the whole process dependent on the quality of this registration step. Although working by patch, if the patch is large enough, allows to capture of some spatial context that might lighten this drawback.

Additionally, the auto-encoder is trained only with a subsample of all possible patches, approximately 4%, which in terms of brain coverage would be more because of the size of the patches (15×15).

II.3.3.ii Outlier detection with one class SVM on latent space

Once the auto-encoder is trained, the decoder can be discarded and the encoder is kept frozen (no weight changes) and used to extract latent representation $\mathbf{z}_i$ of brain patches $\mathbf{x}_i$. Here i represents a brain localization index, and must not be confused with the previously used $\mathbf{x_t}$ notation where $\mathbf{t} = \{\mathbf{1}, \mathbf{2}\}$ indexes the patch pair. The outlier detection process described hereafter for the $\mathbf{z}_i$ could be done directly on the $\mathbf{x}_i$, or with simpler feature extraction step as in El Azami et al. (2016) and Bowles et al. (2017).

Alaverdyan et al. (2020) have trained as many one class SVM as there is a voxel position in the brain, by taking as the training set the latent representation of all training controls at that position. More formally, for a given voxel position, the patches located at position i, $[\mathbf{x}_i^1, \ldots, \mathbf{x}_i^n]$, from the n healthy controls, are run through the encoder E to obtain latent representation $[\mathbf{z}_i^1, \ldots, \mathbf{z}_i^n]$ that are used to minimize the one class SVM problem II.3. They then obtain one decision function for each voxel position, and can then at inference use each decision function to attribute an anomaly score for each of the patient's brain voxels. In other words, the probability density support of the healthy latent representations is estimated for each voxel, with the one class SVM algorithm.

It is important to note that the feature extraction step is done by training on all the possible patch locations, contrary to the anomaly detection step, where one support estimation is done for each possible voxel position.

[1] We detail this pipeline even more formally way later in section IV.1.1.i.

[2] To illustrate this fact, their study contains 60 healthy controls and more than 1 million brain positions.

II.3.3.iii Post-processing

The authors used post-processing steps with the goal of obtaining cluster maps (binary maps with a fixed number of connected components, thus forming clusters). From the raw anomaly score maps, they binarize the maps by manually setting a threshold T which leaves only a small fraction of the voxels above the threshold. After binarization, they remove connected components that have a size smaller than 82. The threshold T is adapted such that at this step they approximately get 10 connected components (clusters). These clusters are then ranked according to a criterion mixing their size and mean anomaly score, thus clusters with large size and/or mean anomaly scores will be ranked higher. This post-processing step notably allows to production of a final result that is simpler to interpret for a clinician (a map with 10 cluster detections) than a raw anomaly score map.

II.3.4 Application to focal cortical dysplasia detection on MRI

Alaverdyan et al. (2020) used 75 healthy controls T1 and FLAIR MR images to train the siamese auto-encoders, and used the same controls to train the one class SVM. We detail this control database in section III.2.3.i. They then evaluated their performances on 21 intractable epilepsy patients (including 18 patients with no visible lesions) and achieved detection of 61% of the lesions when tolerating 8 false positives per patient (and 43% when tolerating 4 false positives per patient).

Comparison with the literature

The work from El Azami et al. (2016), anterior to Alaverdyan et al. (2020), uses the same building blocks, but for the feature extraction step, features are derived from the MRI with simple operations (addition, subtraction) from grey matter segmentation, white matter segmentation, and CSF segmentations obtained from T1 MRI. They use the same pre-processing, the same procedure for outlier detection, and the same post-processing step. The control database used for training was composed of either 37 or 40 healthy controls T1 MRI (two different experiments). Note that they do not use FLAIR maps.

They evaluated the method on 13 patients that had intractable epilepsy, where they showed detection of 10 out of 13 lesions (including 3/3 patients with visible lesions and 7/10 patients where the clinician did not see lesions), thus 77% sensitivity, allowing for an average of 3.2 false positives per patient. The authors also propose to evaluate their method on realistic synthetic lesions (two different types) added to 5 hold-out healthy controls.

In El Azami et al. (2016) and Bowles et al. (2017) (presented in section I.3.2.ii), which also use one class SVM), the feature extraction step is straightforward, and although based on clinical practice, might be specific to the pathology studied, and may miss some of the rich information contained in the raw MRIs. One of the goals of Alaverdyan et al. (2020) was to overcome this limitation with the use of auto-encoders to extract automatically the features contained in T1 and FLAIR MR images.

The detection performances reported by El Azami et al. (2016) and Alaverdyan et al. (2020) are lower than the ones usually reported for these tasks in computer vision or even medical imaging, especially when accounting for the high number of false positives per patient that they tolerate. However, they tackle detection tasks that are much harder than the ones presented in section I.3.2, i.e. where most of the MR images contain lesions that were not seen even by an expert radiologist. Notably, the performances are in the state-of-the-art range when the lesions are non-visible Sone (2021).

II.3.5 Limits of the study

The work presented above, despite being almost the only support estimation method applied to medical imaging for anomaly detection, has several limits that we highlight below and that will serve as the basis of the contributions we establish in the following chapters. We will divide the limits into limits linked to the method and limits linked to the evaluation.

II.3.5.i Limits linked to the method

The method presented uses a high number of independently trained one-class SVM (approximately 1.5 million), that are each trained on a single voxel coordinate, using every available control. This has several limitations, first the optimization is very long due to the large number of models, second, a core assumption of these models is the perfect alignment of the subjects, and thus perfect registration: non-linear registration is a very difficult task and is a research problem on its own, third the number of samples to train a SVM model is capped to the number of controls (other methods use many samples from different controls).

As the model is trained to detect deviation from normality at a very local scale, any deviation, whether pathological or benign (e.g. anatomical variability), would be detected as an anomaly, therefore intrinsically making the model poorly specific. Further experiments have also shown that the results have poor reproducibility, as the siamese network optimization is stochastic, a small variability of the network can greatly influence the rest of the pipeline, especially because of its complexity (many SVM fitting, post-processing).

The model training is dependent on the training database, i.e. there is no fine-tuning on the patient, there are no patient-specific characteristics integrated, and the performances might greatly suffer from domain shift.

Finally, the feature extraction and the outlier detection are decoupled, i.e. they are done in two different steps. There is no guarantee that the extracted features, but the pretext task of training an auto-encoder to reconstruct its input, will be relevant for the outlier detection task.

We summarize the method's weaknesses in the following list:

- Sensitivity to registration due to per-voxel model
- Sensitivity to training dataset size (number of healthy controls)
- No patient-specific characteristics
- Long optimization (1.5M SVMs)
- Intrinsic poor specificity
- Low reproducibility
- Feature extraction decoupled from outlier detection

II.3.5.ii Limits linked to the evaluation

The methods presented above use private databases for model tuning and evaluation, this makes it not possible to reproduce the results, and comparison to other methods is also impossible. The size of the database, although reasonable for medical imaging, is still small in regard to the usual number of samples used in machine learning, thus it can be hard to give strong statistical significance to the established results. For the evaluation database, it is noteworthy that the constitution of a large public epilepsy database is very challenging, mainly due to the

difficulty of obtaining accurate ground truth, as this disease and its causes are still not well understood. To the best of our knowledge, no such database exists. The low number of healthy controls, being from the same order of magnitude as the dimension of the space where the SVM is fitted, challenges the statistical robustness of the results. The size of the control database also questions the ability of the model to capture the anatomical variability of the healthy population. Also, the mean age of the control and test population were carefully matched, and as such, the generalizability of the performances on older/younger population might suffer.

The ground truth used was very approximate, thus counting a detection (true positive) or a false alarm (false positive) for a given cluster is arguable and has a vague definition. Despite the sensitivity being reported for a given false positive rate, more metrics could be computed to give a clearer understanding of the produced score maps, and the post-processing applied to obtain cluster loses the information that could be obtained by looking at the area under the curve of detection metrics and hinders the reproducibility of the analysis.

Further experiments have also shown that the epileptic patients studied, who had undergone surgery in the zone suspected to be responsible for epilepsy crises, had relapsed, and thus the ground truth (suspected zone) might not be valid.

We summarize the evaluation weaknesses in the following list:

- Private database
- Uncertain ground truth
- Uncertain patient outcome
- Small size of the patient database
- Few metrics evaluated
- Variability of the results
- Small size of the control database
- Control and patient databases age-matched

II.4 Contributions outline

We insist on the fact that the weakness of the presented study does not indicate in any manner that the studies are of poor quality. We have identified limits in every literature study, and hope that by shedding light on them we will propose contributions that are relevant and that can advance the scientific problem of anomaly detection in medical imaging. We detail hereafter the outline of the 3 contribution chapters.

Chapter III

To overcome the limitations due to the experimental evaluation (private database of limited size and uncertain ground truth), we propose in chapter III to evaluate the current model on three open databases: first, an industrial image database (MVTecAD Bergmann et al. 2021) with exact ground truth and multiple anomaly types (described section III.1.1). Secondly, an open medical imaging database that comprises hyperintense small brain lesions in MRI T1 and FLAIR images, referred to as white matter hyperintensities (WMH), with exact ground truth

(WMH Kuijf et al. 2019, described in section III.2.1). Thirdly, the PPMI database contains MRI acquisitions (T1, diffusion, etc.) of controls and de novo Parkinson patients, with ground truth at image level (described section III.3.1). By assessing the performances on open databases, we open the model to comparison with the literature and assess with more precision, thanks to the ground truth, the strengths and weaknesses of this approach.

Chapter IV

To address the problem of dependency on the training set (size and features extracted), partially the sensitivity to the registration, and the long optimization, we propose in chapter IV section IV.1 a new strategy for one class SVM training. To improve reproducibility, sensitivity, and specificity, we investigate in IV section IV.2 methods that allow conversion to probability, which allows for building ensemble models and performing score map calibration.

Chapter V

In chapter V, after moving to a more challenging detection task, we try to improve the latent space representation of input patches, with the goal of improving the sensitivity. In section V.2, we investigate classical methods to structure the latent space, such as the addition of variational regularization or positional encoding to the patches. With section V.3, we propose a new end-to-end model, that allows coupling of the feature extraction step and the outlier detection step. We then perform additional experiments in section V.4 to investigate performances of the models and latent space organization in the case of subtle lesion detection.

Summary

We summarize the contributions we made in this thesis in figure II.6. The contributions are highlighted (in green) in the way they differ from the original pipeline (figure II.2), although they not only relate to the original pipeline, as each contribution focuses on some of the weaknesses of said pipeline or blind spot of the presented literature review.

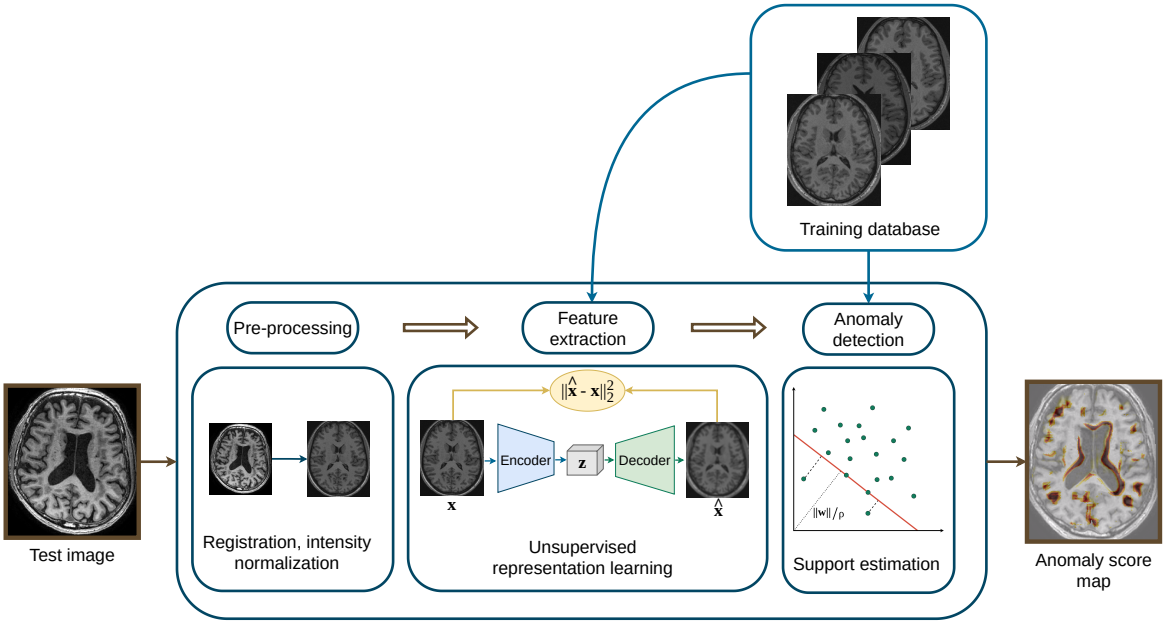

Figure II.5: Full pipeline of the method presented in Alaverdyan et al. (2020). Copy of figure II.2

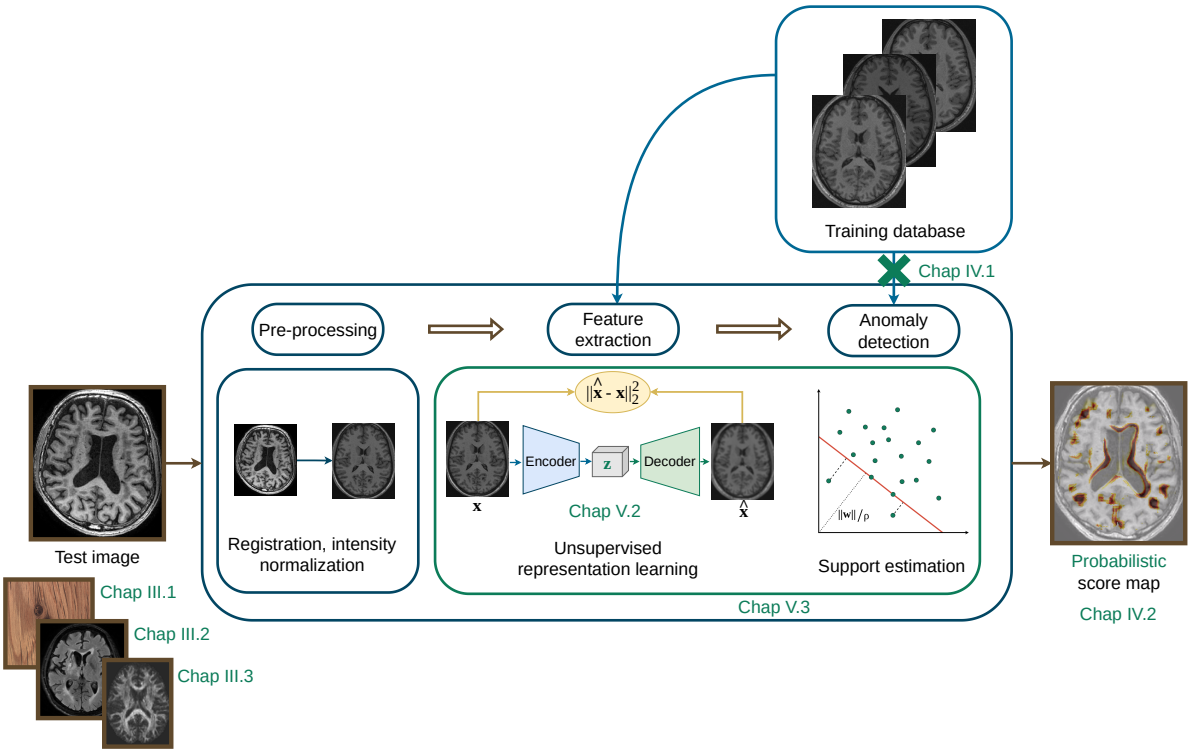

Figure II.6: Most of the contributions made in the different chapters and sections are summarized in this modification of figure II.2, copied hereabove, where the green color indicates said contributions.

III | Extension of the evaluation on public databases

We have seen in chapter II that it is essential, if possible, to carry out the evaluation of anomaly detection methods on databases that are public and have proper ground truth, and evaluate multiple metrics to highlight the strengths and weaknesses of each method. We do so with three public databases, first, an industrial image dataset with multiple anomalies and exact ground truth, secondly with an MRI brain neuroimaging dataset of patients with white matter lesions with exact ground truth, and finally with a brain MRI dataset of control subjects and *de novo* Parkinsonian patients with no semantic annotation but a patient-level annotation regarding the disease progression status.

We hope that this chapter will prove that support estimation methods, and especially methods that detect anomalies in the latent space of auto-encoders are viable alternatives to reconstruction methods, under the watch of diverse and heterogeneous datasets, which will prove the wide usability of these methods. We also hope that the clear evaluation of multiple metrics will help to identify the strengths and weaknesses of the proposed methods.

III.1 Anomaly detection for industrial images

A significant part of the state-of-the-art we covered in chapter I was evaluated on the MVTecAD public dataset that we present in section III.1.1. We wish to investigate the performance of the support estimation method presented in section II.3 on this computer vision dataset. This first allows optimizing the auto-encoder architecture in section III.1.2. This is made possible by the exact ground truth provided with this dataset, which allows measuring the performances with multiple metrics. Finally, in section III.1.3, we benchmark multiple methods using this auto-encoder: reconstruction, support estimation, and restoration, and extend the comparison with two state-of-the-art methods.

III.1.1 MVTec anomaly detection dataset

As we have seen in section I.2.2.i, MVTecAD (Bergmann et al. 2021) has been a very popular dataset for anomaly detection in computer vision. It contains 5354 high resolution[1] RGB images that are divided into 15 categories: 5 textures (*carpet*, *grid*, *leather*, *tile* and *wood*) and 10 objects (*bottle*, *cable*, *capsule*, *hazelnut*, *metal nut*, *pill*, *screw*, *toothbrush*, *transistor* and *zipper*). The objects have more precise edges and positions than the texture, which has regular patterns all over the image.

Each category is divided into normal images, used for training, and test images that contain normal images and anomalous images, with on average 5 types of anomalies.

We focus our evaluation on 2 arbitrary texture categories, *wood* and *carpet*, as we think they already give a correct view of the performances. The choice of textures rather than objects is motivated later.

The *wood* category comprises 247 normal train images, 19 normal test images, and 60 defective images with 5 defect types: holes, liquid droplets, scratches, color stains, and a fifth category with a combination of the previous defects and wood knots. Figure III.1 showcases examples of normal and the five defect categories.

[1] For most objects and texture 1024×1024 pixels, and for other individual categories 700×700 (*metal nut*), 800×800 (*pill*), 840×840 (*tile*), 900×900 (*bottle*) and 1000×1000 (*capsule*).

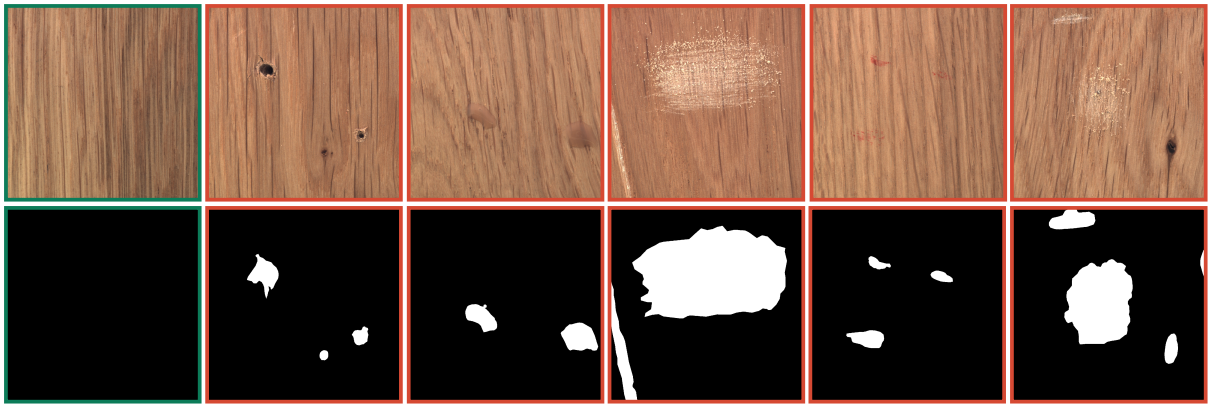

Figure III.1: Examples drawn from the category *wood* of the MVTecAD dataset. Top, from left to right, labels: normal, holes, liquid droplets, scratches, color stains, and combined. Bottom: exact pixel ground truth of anomalies.

The *carpet* category comprises 280 normal train images, 28 normal test images, and 89 defective images with 5 defect types: holes, color stains, cuts, metal contamination, and threads. Figure III.2 showcases examples of normal and the five defect categories.

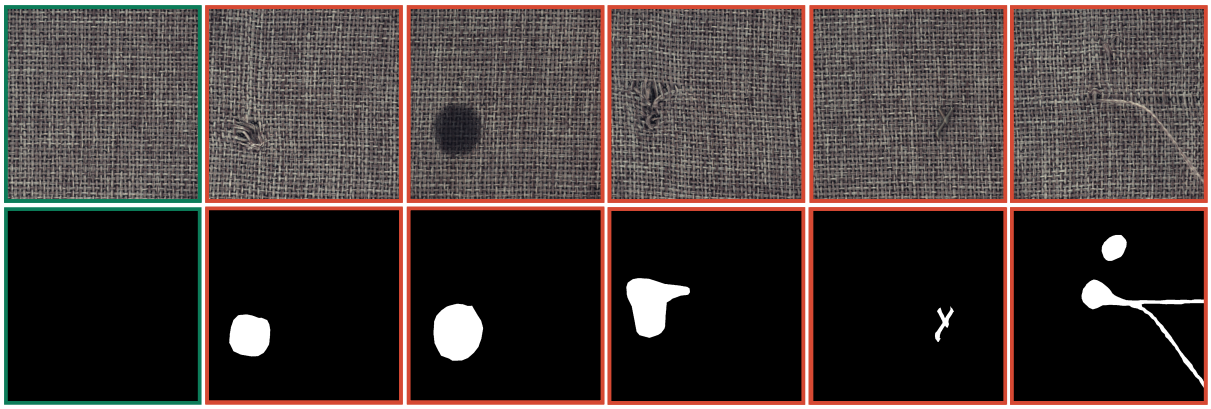

Figure III.2: Examples drawn from the category *carpet* of the MVTecAD dataset. Top, from left to right, labels: normal, holes, color stains, cuts, metal contamination, and threads. Bottom: exact pixel ground truth of anomalies.

As pointed out in section I.2.2.i, the variability of the defect, whether in texture, intensity, or position, makes this dataset challenging, and its exact ground truth makes the evaluation very reliable.

III.1.2 Evolution of the auto-encoder architecture

In the following section, we use a similar pipeline as the one described in section II.3.3. It is composed of a representation learning step done with a siamese auto-encoder (SAE), which we described in section II.3.1 and an anomaly detection step itself, done with a one-class SVM (OC-SVM). We detailed in section II.3.3 the implementation that was done by the authors, in the form of patch-based SAE, with the similarity term based on the position of the patches, and the training of the OC-SVM, which was done by training one OC-SVM per brain localization, which we call ***loc*OC-SVM**. Hereafter, we describe how we adapt the method to perform anomaly detection on MVTecAD, and propose several alternative auto-encoder architectures to perform

the representation learning step. We thus focus only on the representation learning step, and leave the outlier detection step unchanged.

III.1.2.i Methods

The original auto-encoder proposed by Alaverdyan et al. (2020) has a subtle flaw: the use of an even max-pooling kernel on an odd feature map[1] was causing the discard of the last column and last row of the feature map obtained after the first convolution. We wanted to remedy this issue and thus propose 5 alternative auto-encoder architectures, some close to the original one and some more distinct. Architectures of the encoders are presented in figure III.3, the decoders constructed to be symmetric of the encoders for each architecture. The original architecture, which we name 'ConvSiamAlaverdyan' is also briefly presented (which has a decoder slightly different than the symmetric of the encoder).

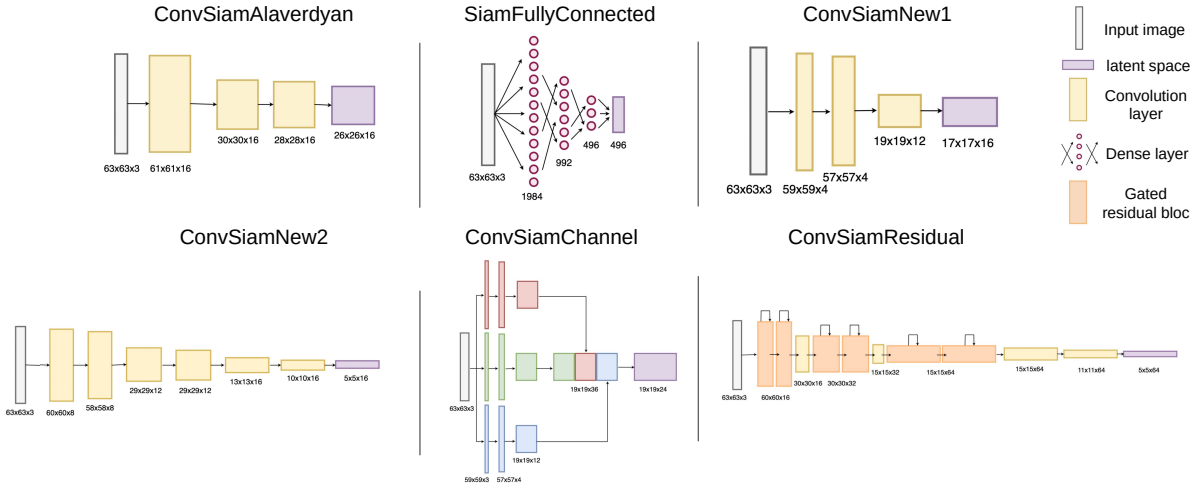

Figure III.3: Original encoder from Alaverdyan et al. (2020), named 'ConvSiamAlaverdyan', and 5 proposed alternative architectures. Decoders are constructed to be the symmetric of each encoder. Every auto-encoder is trained with patches and with the siamese loss term.

- ConvSiamAlaverdyan: the encoder is composed of 3 convolutional blocks, a maxpooling block, ReLU activation function. The decoder is not the exact symmetric of the encoder, no batch normalization (Ioffe and Szegedy 2015) is applied.
- ConvSiamNew1: the encoder is composed of 4 convolutional blocks, GeLU activation functions, BatchNormalization and the decoder is the exact symmetric of the encoder.
- ConvSiamNew2: the encoder is composed of 3 convolutional blocks, ReLU activation functions, BatchNormalization and the decoder is the exact symmetric of the encoder.
- ConvSiamResidual: the encoder is composed of 7 convolutional blocks and 3 residual blocks (as in He et al. 2016), ELU activation functions, batch normalization and the decoder is symmetric of the encoder.
- ConvSiamChannel: each channel of the image goes through an independent path of 3 convolutional blocks, and is then concatenated and run through a dense layer. GeLU

[1] Precisely, the feature map of size 13×13 obtained after the first convolution was taken as the input of a 2×2 max-pooling operation, therefore discarding the 13^{th} row and 13^{th} column of this feature map.

activation functions and batch normalization are used, the decoder is the symmetric of the encoder.

- SiamFullyConnected: the encoder is simply composed of 4 fully connected layers, the decoder is symmetric of the encoder, and GeLU and batch normalization are used.

The idea behind ConvSiamNew1&2 was to propose a simple architecture that did not have the same flaw as ConvSiamAlaverdyan, adding batch normalization as it has proved its worth, add a more effective activation function, and also remove the maxpooling operation, as it is believed to have the capacity to generate invariance (chapter 9 Goodfellow et al. 2016) which might not be useful for very small patches. We tested ConvSiamResidual to use the residual blocks that were proven to be very effective. The idea behind ConvSiamChannel was to put more weight into each color channel that could be processed individually and then aggregated. We also tested a fully connected network as a very simple baseline for this task.

III.1.2.ii Experiments

For each proposed auto-encoder, we train the network with the following siamese loss, as presented in section II.3.1:

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

Each SAE is trained with pairs of patches, with this time size of 63×63, as the images have higher resolution than brain MRI. A pair of patches is constituted by any two patches of the same image or different images, with the idea that texture is invariant to any translations, and so any patch can be paired with any other patch[1]. Note that the choice to focus the evaluation on texture categories was caused by the design of the SAE, which implies pairing patches together, choosing texture categories neutralizes this pairing problem: we can assume that every patch can be paired with any other, whereas if we were to evaluate on objects, we would need to define position in the object, to do this appropriate pairing.

We train each SAE with 1000 patches sampled from each training image, of size 63×63[2], with training batch sizes of 100 and 200 for testing, for 10 epochs with Adam optimizer (learning rate of 0.001). The similarity coefficient for the siamese term is set to 0.005. Best validation loss is used to select the optimal model parameters.

After training the SAE, one OC-SVM is trained to estimate the support of the normal patches latent distribution. It is trained on a subsample of 200 latent representations, with the gamma RBF coefficient set to the inverse of the product between the variance and the dimension of the latent space. The hyperparameter of the OC-SVM ν is set to 0.03.

Note that the training of the OC-SVM differs from the SAE+*loc*OC-SVM method presented in section II.3.3, here only one OC-SVM is trained as if there was only one localization possible (which is a reasonable assumption considering we study textures). We name this simplified setup SAE+OC-SVM.

We evaluate the area under (AU) the ROC curve, precision-recall (PR) curve, per region overlap (PRO) curve, and for ROC and PRO, their area limited to 30% false positive rates, as above this threshold the map can be considered degenerate when the anomalies are very scarce in the dataset (see chapter I section I.2.1).

[1] This pairing will have more sense when looking at co-registered patients in the following sections.

[2] With the medical image experiments done in this thesis, we use size 15×15. Here we use size 63×63, as the resolution of the images on MVTecAD is much higher than on standard MR images. We believe this roughly corresponds to the same field of view.

III.1.2.iii Results and discussion

In table III.1, we evaluate the detection performances of the different auto-encoders (when using their encoder in conjunction with one class SVM) on *wood*, table III.2 presents the same results for *carpet*.

MVTecAD *wood*	*AU ROC*	*AU ROC 30*	*AU PR*	*AU PRO*	*AU PRO 30*
ConvSiamAlaverdyan	0.66	0.45	0.20	0.70	0.50
ConvSiamNew1	0.84	0.63	0.48	0.87	0.67
ConvSiamNew2	0.72	0.27	0.32	0.75	0.27
ConvSiamResidual	0.69	0.50	0.31	0.77	0.59
ConvSiamChannel	0.82	0.77	0.33	0.82	0.77
SiamFullyConnected	0.53	0.30	0.14	0.45	0.15

Table III.1: Pixel-level anomaly detection performances on the subset *wood* of MVTecAD, for SAE+OC-SVM method, used with different auto-encoder architectures.

MVTecAD *carpet*	*AU ROC*	*AU ROC 30*	*AU PR*	*AU PRO*	*AU PRO 30*
ConvSiamAlaverdyan	0.51	0.16	0.05	0.49	0.16
ConvSiamNew1	0.37	0.08	0.01	0.35	0.09
ConvSiamNew2	0.39	0.08	0.01	0.38	0.08
ConvSiamResidual	0.65	0.30	0.05	0.67	0.32
ConvSiamChannel	0.5	0.14	0.02	0.48	0.12
SiamFullyConnected	0.38	0.08	0.02	0.35	0.09

Table III.2: Pixel-level anomaly detection performances on the subset *carpet* of MVTecAD, for SAE+OC-SVM method, used with different auto-encoder architectures.

We recall that the *AU ROC* and *AU PRO* of a random classifier would be 0.5, while its *AU ROC 30* and *AU PRO 30* would be 0.15. The random classifier *AU PR* would be 0.04 for *wood* and 0.02 for *carpet*.

We can see that for *wood*, ConvSiamNew1 outperforms every model for *AU ROC*, *AU PR*, and *AU PRO*. For *AU ROC 30* and *AU PRO 30*, ConvSiamChannel has the best performances. On *carpet*, we see that no model reaches above chance performances, except ConvSiamAlaverdyan on *AU PR* and ConvSiamResidual on all metrics. Figure III.4 presents a visual comparison of the anomaly maps obtained for every model on samples from *wood*.

The best-performing model on *wood* is arguably ConvSiamNew1, as it outperforms all the other models on 3 out of 5 metrics, and as the best model on the 2 other metrics (ConvSiamChannel) drops significantly for these 3 metrics. On *carpet*, the results are arguably non-significant, as almost all models struggle to get at chance level, and the model that stands out (ConvSiamResidual) doesn't have outstanding performances. As such, in the remainder of the manuscript, we use ConvSiamNew1 as the default auto-encoder for every experiment. In figure III.4 we see that ConvSiamNew1 detects defaults that were missed by the model from Alaverdyan et al. (2020), such as the top scratch on the rightmost image, and that it seems to have less false positives in these images, which is confirmed quantitatively by the metrics. We see with this image that there seems to be great variability between defect types, and a more in-depth analysis could be carried out to select the best model.

An interesting difference between ConvSiamAlaverdyan and ConvSiamNew1 is that when using 15×15 patches, ConvSiamAlaverdyan has a latent space of size 64, and ConvSiamNew1

has a latent space of size 16 (272 versus 416 when using 63×63 patches). This downsize in dimensionality doesn't seem to have an impact on the presented performances and could be useful as it is often more difficult to learn support or density in higher dimensional spaces.

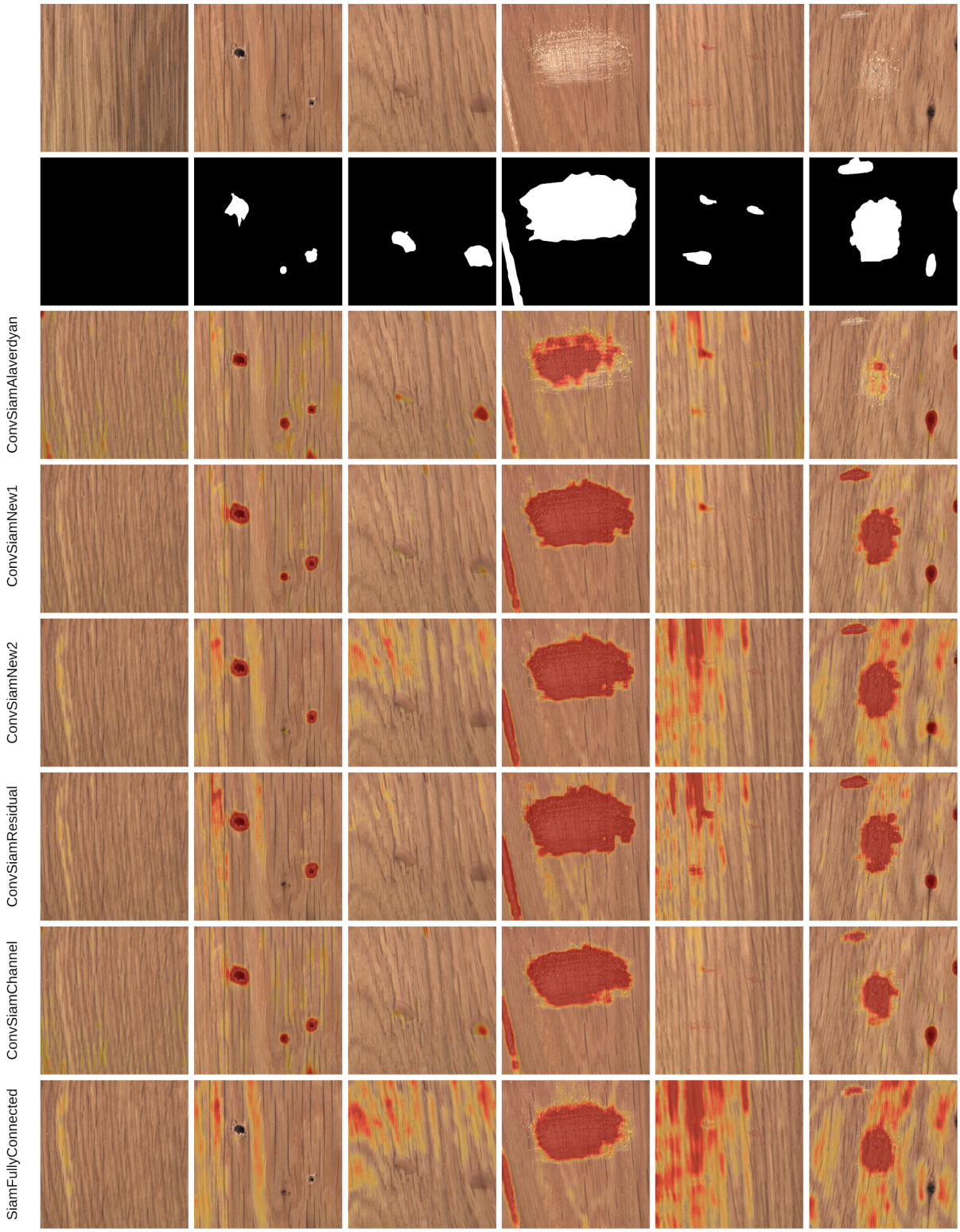

Figure III.4: Visual comparison of the studied auto-encoders when used with one class SVM (SAE+OC-SVM) to produce anomaly maps. The top two rows are the input image and ground truth, following rows are a superposition of the anomaly score map (more red means more anomalous) and input image for every model. Each column showcases a defect type (or normal image for the 1st column)

III.1.3 Comparison with state-of-the-art methods

Now that we have chosen a more optimal auto-encoder architecture, we wish to compare the anomaly detection performances on *wood* and *carpet* of this auto-encoder, for different methods (reconstruction, restoration, support estimation) and with other state-of-the-art methods.

This work was presented at GRETSI 2023 (Pinon et al. 2023b).

III.1.3.i Methods

SAE-based methods

As in the previous section, we study the SAE+OC-SVM method. We also study the use of the reconstruction error of the SAE, named SAE recons., the anomaly score per voxel is the mean squared error between the central pixel of the reconstructed patch and the original pixel.

Auto-encoder restoration with Transformer

To put into practice a restoration method, we chose a discrete restoration method, that imposes two things: the quantization of the auto-encoder model, and the training of an auto-regressive model (see chapter I section I.3.2.iv, where we present VQ-VAE and auto-regressive models).

As in Van Den Oord et al. (2017), we use a VQ-VAE for discrete representation learning. We dropped the siamese framework for these experiments: for simplicity and because preliminary experiments combining siamese architecture with quantization were not conclusive in terms of performance. For the auto-regressive model, we use a Transformer-decoder architecture Vaswani et al. (2017b), that allows predicting the probability of each latent coordinate vector, thus allowing resampling to produce a 'healed' version of the latent vector. We call this method VQ-VAE restoration.

VQ-VAE-based methods

As we implemented a quantized auto-encoder, we take this opportunity to study its reconstruction error (VQ-VAE recons.) and its combination with a one-class SVM (VQ-VAE+OC-SVM).

One class SVM with pre-trained feature extractor

The performances of convolutional neural networks pre-trained with large image datasets are well proven in the literature and especially in anomaly detection for MVTecAD, where the top performing models (e.g. FastFlow Yu et al. 2021 and PaDiM Defard et al. 2021) are pretrained on Image-net Deng et al. (2009). To evaluate the strength of this pretraining, we propose to evaluate the performances of a one-class SVM trained on feature extracted with a Resnet-50 (He et al. 2016). This Resnet, trained on Image-net (not fine-tuned on MVTecAD), should extract high-quality features and the one-class SVM is tuned specifically on MVTecAD. We call this method Resnet (pre-trained)+ OC-SVM.

PaDiM

As presented in paragraph I.3.1.v:Density estimation, the model proposed by Defard et al. (2021) uses a Resnet-50 that processes patches of the input image and uses multiple layers of the Resnet outputs as the features to estimate a multivariate gaussian distribution for each patch position in the input image. This method can take advantage of the pre-training of the Resnet on Image-net.

FastFlow

As presented also in paragraph I.3.1.v:Density estimation, Yu et al. (2021) also uses a Resnet-50 for feature extraction, and uses extracted feature as an input to a normalizing flow (Tabak and Turner 2013), that estimate the probability density by learning a mapping of the feature to a space where the training data would be gaussian.

III.1.3.ii Experiments

For the SAE-based methods, the parameters are the same as presented in section III.1.2.ii. Methods that used SVM also have the same parameters as in this section. For the VQ-VAE, the codebook size is chosen to be 512. The transformer, for restoration, is trained during 50 epochs, with sequences of 100 patches per image. Adam optimizer is used with learning rate $1e^{-4}$. The threshold of the probability for resampling (restoration) is set to 0.02. FastFlow and PaDiM are implemented with the help of the Anomalib toolbox Akcay et al. (2022), with the same hyperparameters as in their original studies.

III.1.3.iii Results and discussion

Figure III.5 presents reconstruction obtained with normal auto-encoder, quantized auto-encoder, and a restoration obtained by resampling the latent quantized vector for *wood* images. We see that the wood knot, not present in the training set, is 'healed'. There is also a difference in contrast when reconstructing through a quantized auto-encoder.

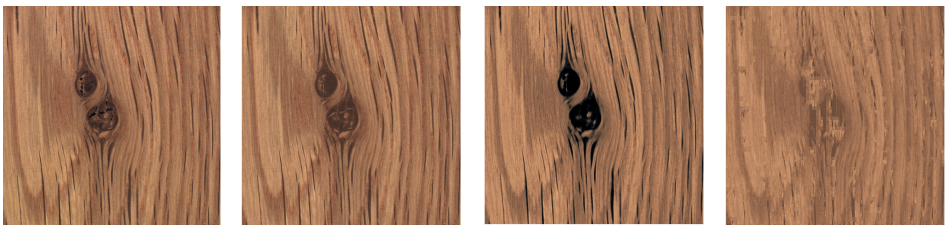

Figure III.5: From left to right: input image, reconstruction by SAE, reconstruction by VQ-VAE, and restoration by VQ-VAE + Transformer.

For both PaDiM and FastFlow, we evaluate the performances without pre-training of the feature extractor, using the Anomalib library (Akcay et al. 2022), and report the performances found in the original papers for the pre-trained version. The results are presented in table III.3 for *wood* and III.4 for *carpet*. Figure III.6 shows the different anomaly score maps of the presented models on a subset of *wood*. As the validation set is fixed, all the experiments' performances can be compared.

On *wood*, we see that PaDiM and FastFlow are the best-performing models on all the metrics, with an advantage for FastFlow on the PRO metrics. This shows the strength of the models of the literature, even with no pre-training. When the models are pre-trained, on the available metrics they seem to outperform by a large margin the non-pre-trained counterpart. However, we see that the pre-trained ResNet + OC-SVM doesn't outperform our proposed models, meaning that the strength of the literature model does not only reside in the pre-training on a large computer vision database. Among the SAE-based models, the SAE+OC-SVM (i.e. our standard pipeline) seems to be the best candidate on all of the metrics and has performances that are close to the literature models.

By visual inspection of figure III.6 the reconstruction and restoration methods seem to produce very sharp anomaly maps, where some parts of the anomalies are sometimes missed or not

MVTecAD *wood*	*AU ROC*	*AU ROC 30*	*AU PR*	*AU PRO*	*AU PRO 30*
SAE recons.	0.70	0.45	0.29	0.75	0.54
VQ-VAE recons.	0.71	0.47	0.26	0.55	0.26
SAE + OC-SVM	0.80	0.57	0.35	0.82	0.62
VQ-VAE + OC-SVM	0.75	0.50	0.29	0.78	0.54
VQ-VAE restoration	0.76	0.56	0.27	0.83	0.63
Resnet (pre-trained) + OC-SVM	0.69	0.37	0.25	0.74	0.39
FastFlow (not pre-trained)	0.85	0.65	0.41	0.90	0.72
PaDiM (not pre-trained)	0.85	0.61	0.32	0.83	0.59
FastFlow (pre-trained) (from Yu et al. (2021))	0.97	/	/	/	/
PaDiM (pre-trained) (from Defard et al. (2021))	0.95	/	/	/	0.91

Table III.3: Anomaly detection performances on the subset *wood* of the MVTecAD dataset (Bergmann et al. 2021) for the presented models and metrics. The two last rows are taken directly from the literature, / indicates that the metric was not reported.

MVTecAD *carpet*	*AU ROC*	*AU ROC 30*	*AU PR*	*AU PRO*	*AU PRO 30*
SAE recons.	0.51	0.17	0.02	0.57	0.25
VQ-VAE recons.	0.51	0.16	0.02	0.51	0.16
SAE + OC-SVM	0.47	0.12	0.02	0.47	0.14
VQ-VAE + OC-SVM	0.54	0.15	0.02	0.57	0.21
VQ-VAE restoration	0.79	0.45	0.05	0.76	0.42
Resnet (pre-trained) + OC-SVM	0.35	0.06	0.01	0.37	0.07
FastFlow (not pre-trained)	0.83	0.59	0.24	0.85	0.60
PaDiM (not pre-trained)	0.75	0.41	0.14	0.69	0.34
FastFlow (pre-trained) (from Yu et al. (2021))	0.99	/	/	/	/
PaDiM (pre-trained) (from Defard et al. (2021))	0.99	/	/	/	0.96

Table III.4: Anomaly detection performances on the subset *carpet* of the MVTecAD dataset (Bergmann et al. 2021) for the presented models and metrics. The two last rows are taken directly from the literature, / indicates that the metric was not reported.

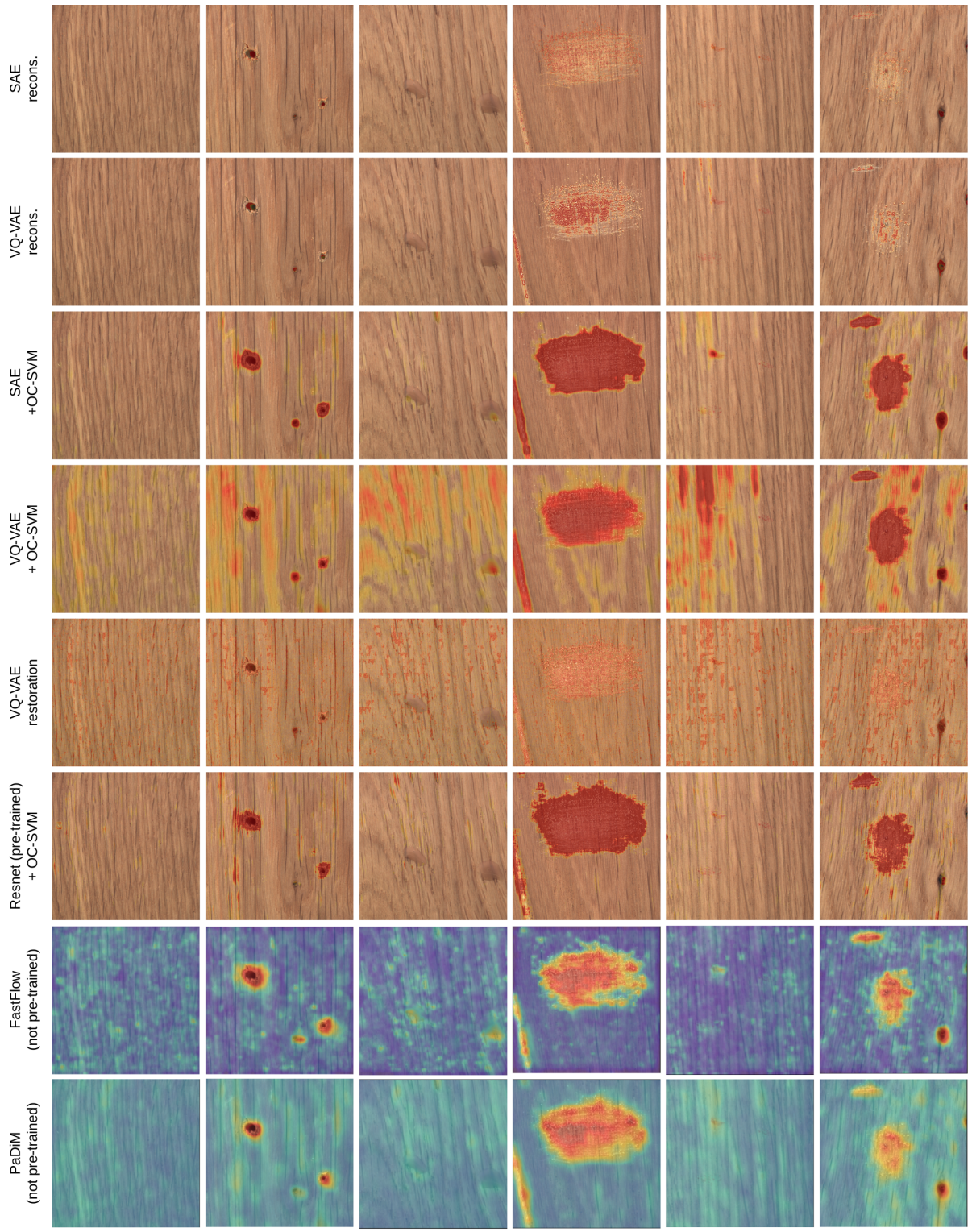

Figure III.6: Visual comparison of the different studied methods used to produce anomaly maps. Images are superposition of the anomaly score map (from transparent to red, red means more anomalous) and input image for every model. Last two columns have different colormap (from blue to red, red means more anomalous). Each column showcase a defect type (or normal), ground truth can be visualized figure III.4.

fully covered. On the other end, one class SVM methods tend to produce coarser maps, where anomalies are not missed but the false positives seem more numerous. The quantization doesn't seem to affect the reconstruction methods, and the quantitative performances corroborate this result. For the one class SVM, the quantization seems to increase the number of false positives, in fact, one would not see any advantages for support estimation to quantize the latent space, as it leads to some loss of information. FastFlow and PaDiM produce maps that are sharper than one class SVM methods and coarser than reconstruction methods, leading to better quantitative performances in the end. We see that overall, all the methods struggle to detect liquid droplets, and never miss holes.

On *carpet*, the findings are different, as for SAE-based methods almost all results are at chance level, except for the restoration method, which seems a good candidate for this subset, and that tackles PaDiM and FastFlow in terms of performance. We see again that PaDiM and FastFlow, when pre-trained outperform the rest of the methods by a large margin.

The findings on the performances of PaDiM and FastFlow compared to our methods have to be mitigated for several reasons. First, PaDiM and FastFlow are trained and evaluated (both in the original paper and in Anomalib) on downsampled images (from 1024×1024 to 256×256), and secondly, because they apply post-processing on the anomaly maps (for instance gaussian smoothing and averaging multiple maps for PaDiM). This post-processing work was not done for the proposed SAE-based methods and as such we believe the anomaly maps could be refined to push the performances.

III.1.4 Conclusion and perspectives

We have seen the performances of the proposed and optimized auto-encoder architecture when used for reconstruction/restoration/support estimation, on two subsets of the MVTecAD database. We found that the support estimation technique was a good candidate when combined with our patch-based auto-encoder. For a more complete evaluation, it would be interesting to look at a density estimation technique combined with the proposed auto-encoder architecture. For a more in-depth comparison, it would be beneficial to work on the post-processing of the anomaly score maps, to see if we can close the gap with the state-of-the-art performances. Even though we do not reach such performances, we believe that with more work on the post-processing of the anomaly score maps we could reach this goal and thus maybe give more strength to our analysis.

We believe the SAE+OC-SVM is a good candidate for anomaly detection, as the principle is relatively simple and straightforward, the model has a low number of parameters (because it is patch-based), meaning it is lighter, easier, and faster to train, unlike FastFlow and PaDiM.

Evaluating the performances by looking at more than one metric still allowed us to give more strength to the analysis and highlight differences between the studied methods. One of the weaknesses of the proposed analysis is that we didn't look at the metrics on each defect type (as often done in the literature). We saw with the qualitative analysis that some defects seemed to be often missed (e.g. liquid droplets) while others seemed always correctly segmented (e.g. holes)[1].

[1] As an example the SAE+OC-SVM, on *wood* achieves an $AU\ PRO$ of 0.94 on the holes and only 0.49 on the liquid droplets.

III.2 Detection of white matter hyperintensities (WMH) in brain MRI

Now that we have optimized the patch-based auto-encoder architecture and tested it with different methods (restoration/support estimation/reconstruction) on a public database, we wish to continue the public database evaluation, but this time on a neuroimaging task, as it is the purpose of this thesis. We do so on a public leukoaraiosis MRI database that we present in the first section. We then limit the study to the SAE+*loc*OC-SVM, because first, the method seemed promising from the results on MVTecAD, second, support estimation methods are not thoroughly studied in the literature as we demonstrated in chapter I, and third, we compare to two state of the art methods that are reconstruction based and restoration based.

This work was partially presented at MIDL 2023 Pinon et al. (2023c).

III.2.1 WMH segmentation challenge

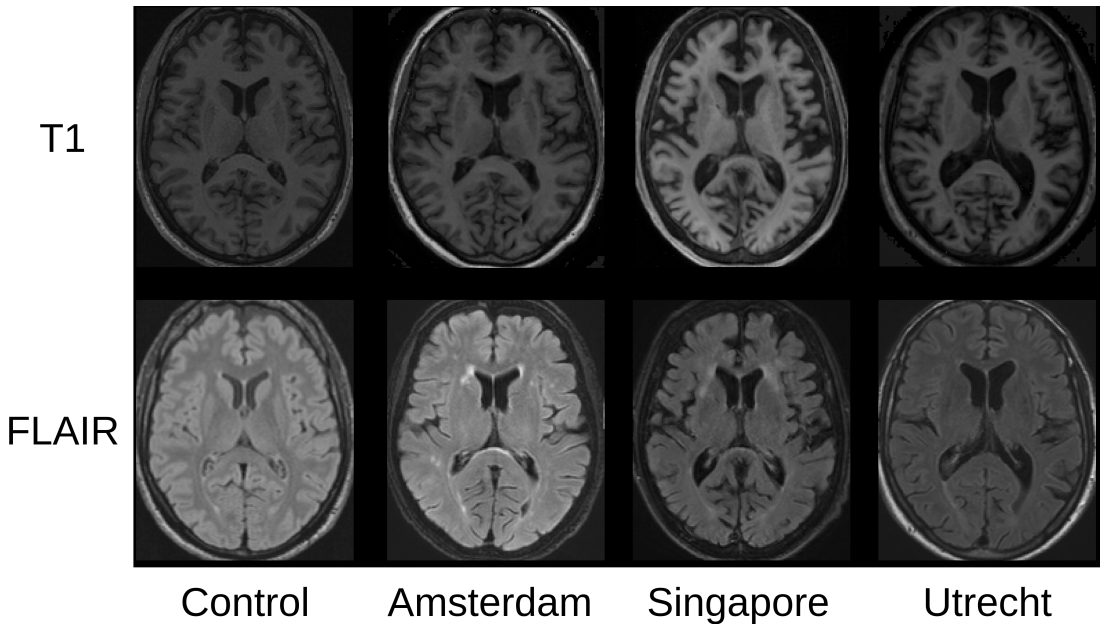

Figure III.7: Showcase of T1 and FLAIR images of the CERMEP control database and of the 3 hospitals of the WMH database

The white matter hyperintensities segmentation challenge (Kuijf et al. 2019), is a MICCAI conference challenge that was active from 2017 to 2022. The goal of the challenge was to segment white matter hyperintensities in brain MRI, with train data consisting of MRI volumes and ground truth segmentation, and test data hidden from the participants. The challenge was thus a supervised segmentation task.

It has been used as an unsupervised anomaly detection task (pixel-wise anomaly detection) in Baur et al. (2021b), Pinaya et al. (2022b), Pinaya et al. (2022a) by training on another healthy dataset, and using the WMH dataset only for evaluation (and Meissen et al. 2021a with no training dataset).

The dataset[1] contains MRI scans from 60 patients: 20 patients from 3 different hospitals, each with different MRI vendors, different voxel sizes, relaxation, and echo time. Each patient

[1]We mean by 'the dataset', the training part of the dataset, that is available to the participants.

scan consists of one T1 volume, one FLAIR volume, and the associated ground truth mask that contains the precise segmentation of the white matter hyperintensities, and also the segmentation of any other pathologies. Participants in the challenge were asked to segment the WMH, and the other pathologies were ignored during evaluation[1]. The segmentation of WMH and other pathologies was done by an expert observer and corrected by consensus by a second expert observer. Images were pre-processed with bias field correction, resampling, registration of T1 to FLAIR, and defacing.

It is worth noting that the patients' ages included in this study are 70.1±9.3 (mean±std) years old, which is considerably older than the mean age in Europe for instance[2]. This seems concordant with the studied pathology but plays a role in the domain shift between the potential training set and this testing set. Also, most of the WMH are, obviously, in the white matter, giving an an priori strength to models that attribute higher scores in white matter than elsewhere. 50% of the subjects were male.

This dataset has the advantage of containing a wide range of lesion volume (from 0.78 cm^3 to 195.15 cm^3), and other pathologies, that are precisely segmented. It also covers three different hospitals and scanners, old population, thus making the task difficult because of the domain shift, but close to the real clinical practice. Its main drawback for anomaly detection is the over-representation of one type of anomaly, which is also easy to find because of its hyperintensity.

III.2.2 Methods

In this section, we study three methods, the SAE+*loc*OC-SVM, presented in section II.3.3, which is a support estimation method, and two methods presented in paragraph I.3.2.iii:Auto-encoders and I.3.2.iv:Restoration on quantized latent space a reconstruction method and a restoration method.

III.2.2.i SAE+*loc*OC-SVM

The auto-encoder is the optimized architecture from section III.1.3. To set up the siamese auto-encoder, we need a one-to-one correspondence of voxel positions from one subject to another. To do this we register each subject (train or test) to a common atlas. The registration pipeline is detailed in appendix B.

A pair of patches is thus constituted of two patches that are located in the same brain position, from two different healthy subjects. We hope that the latent space will gain some structure by putting patches that should be similar (because in the same position) in the same neighborhood in the latent space. Then the support estimation is done with one one-class SVM per brain location, allowing learning a normal class boundary that is specific to the position, despite the fact the training set is reduced to the number of training patients.

III.2.2.ii AE reconstruction error Baur et al. (2021b)

Baur et al. (2021b) have proposed to use a reconstruction method, using an auto-encoder, to tackle this problem, as presented in section I.3.2.iii:Auto-encoders.

The auto-encoder is trained on a healthy control database, and presented with the patients at inference, with the hope that the pathological areas, not seen during training, will be less well

[1]We believe it is also what is done in Baur et al. (2021b) and Pinaya et al. (2022b), as it was the goal task in the original challenge, but no information is given about this matter in their work.

[2]The EU reports (European Union 2022) a median age of 44.4 years in 2022, even if the median is not the same as the mean we can guess that with the very high number of EU citizen and the central limit theorem that the mean is not too far from the median.

reconstructed. Their auto-encoder includes 2 skip connections, making it a hybrid between a U-net and an auto-encoder. They apply data augmentation during training and median filtering on the output anomaly maps.

In their original study, they trained on an in-house healthy dataset of 109 controls, using FLAIR sequences only. They evaluated on WMH, on 51 out of 60 patients[1] and report pixel-level *AU PR* and *Dice*, with the threshold determined with a held-out validation dataset composed of patients (this metric is thus very close to best achievable *Dice*).

We re-implement this method, using the same auto-encoder, data augmentation, and post-processing. We use T1 and FLAIR as input channels, to achieve fair comparison with our method.

III.2.2.iii VQ-VAE + Transformer restoration Pinaya et al. (2022b)

Pinaya et al. (2022b) have proposed to use a quantized auto-encoder for representation learning, and then a transformer model to perform restoration on the quantized latent space of the auto-encoder, as presented in section I.3.2.iv:Restoration on quantized latent space.

As all the latent vectors are not resampled (restored), this produces a resampling mask in the latent space(indicating which latent vector was resampled). The final restoration error is weighted by the upsampling of the resampling mask.

They limit their study to training and testing on the 4 central slices of FLAIR MRI volumes. They use 15,000 pseudo-healthy FLAIR MRI volumes from the UK biobank database (Sudlow et al. 2015) for training, data augmentation, and test on WMH patients. They report the pixel-level *AU PR* and best achievable *Dice*. We re-implement this method, using the same quantized auto-encoder architecture and data augmentation strategy. We use a different transformer architecture Choromanski et al. (2021) than the one used by the authors Dosovitskiy et al. (2021) as we found that the original one led to worse performances and higher computation time.

Once encoded in the latent space, the latent representation still retains its 2-dimensional structure. The model used for resampling (Transformer), as with all the auto-regressive models, needs to organize these latent representations as a list of tokens, thus a 1-dimensional ordering. To transition from 2D to 1D, multiple orderings are possible. The authors used multiple orderings (32) and aggregated the results into a single anomaly score map. While this improved the performances, it amplifies the computational time significantly; thus the combination of multiple orderings was not evaluated for computational reasons. We use T1 and FLAIR as input channels, to achieve fair comparison with our method.

III.2.3 Experiments

We train every method presented above on the CERMEP control database, presented hereunder. For the testing, every patient of the WMH dataset is used. We evaluate the detection metrics voxel-wise, on the whole volume of every patient. 25 000 patches, of size 15×15, are sampled from each control subject. The SAE (with architecture ConSiamNew1) is trained for 10 epochs, with best model selection on validation loss. The validation set is composed of 15 out of 75 control subjects. The similarity coefficient for the siamese term is set to 0.005. The hyperparameter of the OC-SVM ν is set to 0.03.

We study the same metrics as in the previous experiments on MVTecAD and also look at the best achievable *Dice* ($\lceil Dice \rceil$) as it is a common metric in the anomaly detection in neuroimaging community (and reported in Baur et al. 2021b and Pinaya et al. 2022b). We evaluate the detection of WMH and other pathologies indistinctly[2].

[1] We didn't find any information on which patients were used.

[2] A separated analysis will be carried out in chapter V.

III.2.3.i CERMEP Control database

What we call the CERMEP[1] database (Mérida et al. 2021) is a semi-public (available on demand) dataset composed of 75 healthy controls T1 and FLAIR image acquired on the same scanner. The same registration process was applied to these training controls as the WMH patients and is detailed in appendix B.1. These healthy controls are relatively younger (38±11.5 years) than the WMH patients.

III.2.4 Results and discussion

The results are presented in table III.5, where we present the results averaged on every patient[2] from every hospital. Figure III.8 showcases the anomaly maps of the three methods, along with the input FLAIR and the ground truth of the WMH. The dynamic range of these anomaly score maps is adapted, for visualization purposes, to showcase with more detail the most anomalous scores[3].

WMH (T1+FLAIR) 3 hospitals	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM
AU ROC	0.69 ± 0.13	0.53 ± 0.09	0.52 ± 0.19
AU ROC 30	0.40 ± 0.20	0.20 ± 0.12	0.19 ± 0.16
AU PR	0.065 ± 0.079	0.028 ± 0.030	0.023 ± 0.031
AU PRO	0.55 ± 0.10	0.50 ± 0.08	0.43 ± 0.17
AU PRO 30	0.19 ± 0.13	0.15 ± 0.07	0.09 ± 0.13
$\lceil Dice \rceil$	0.11 ± 0.10	0.06 ± 0.05	0.05 ± 0.05

Table III.5: Mean (± std) metric on every patient from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.007±0.006.

We see on the quantitative side that the re-implementation of the VQ-VAE + transformer method is superior for every metric. The AE method and our proposal are around chance levels for nearly every metric except *AU PR*, indicating a good detection of the most anomalous scores. We show in table III.6 the detailed performances for each hospital. We see that the performance of the VQ-VAE + transformer drops significantly for the Utrecht hospital. On the other hand, for Amsterdam, AE and the SAE+*loc*OC-SVM reach above-chance performances.

We see in figure III.8 that the VQ-VAE + transformer method seems to produce sharp anomaly maps, that correctly segment the anomalies but have a large number of false positives. The AE method seems to miss most of the anomalies and produce homogeneous maps. The SAE+*loc*OC-SVM seems to detect correctly the WMH, but with imprecise segmentation, and localized false positives on the ventricles or the cortex border (where there might be brain shrinkage because of the old age of the WMH population compared to the control dataset).

[1] From the name of the institute where the data was acquired: Centre d'Etude et de Recherche Multimodal Et Pluridisciplinaire en imagerie du vivant. Approval number 2012-A00516-37 and 2014-00610-56.

[2] Note that as a consequence, the standard deviation presented corresponds to the variation of the performances among the patients (which we could call aleatoric uncertainty) and not the variation of model performances among some cross-validation (which we could call epistemic uncertainty). As the patients have large differences in terms of lesion size, MRI parameters, pathologies, etc., it is not surprising that the variance of the performance can be quite large.

[3] For the SVM methods, a natural threshold is used: scores above 0 (on the correct side of the hyperplane, i.e. inside the support) are displayed as all white. No other processing is used for SVM. For the other methods, a more 'manual' approach is used to fix the dynamic range.

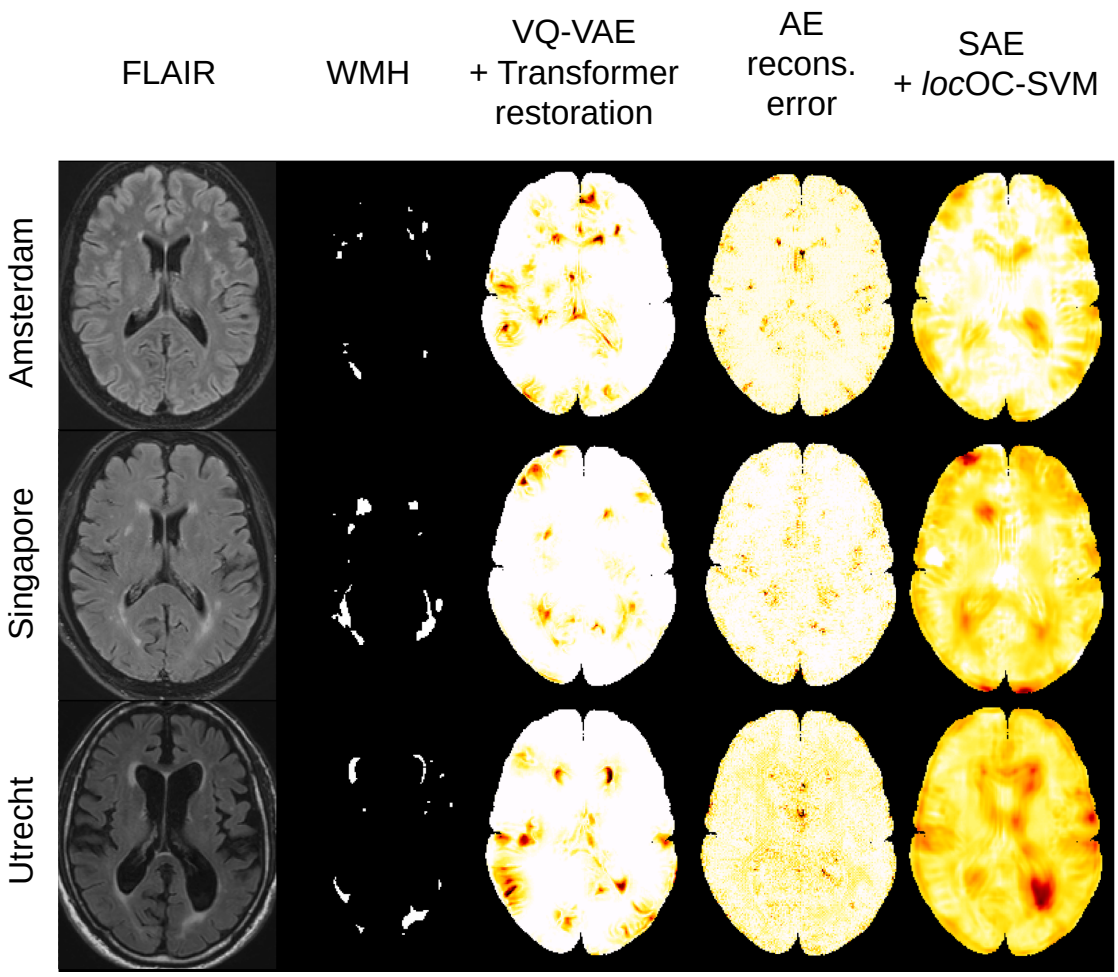

Figure III.8: Showcase of the studied methods anomaly maps on three slices from three patients from one of each hospital, Baur et al. (2021b), Pinaya et al. (2022b) and ours, redder mean anomaly score higher. T1 + FLAIR is used as input for the models but only FLAIR is shown, only ground truth of the WMH is shown.

The reported $\lceil Dice \rceil$ and $AU\ PR$ in Pinaya et al. (2022b) ($\lceil Dice \rceil = 0.269$, $AU\ PR = 0.158$) and Baur et al. (2021b) ($\lceil Dice \rceil = 0.45$ and $AU\ PR = 0.37$) are much higher than the values reported here. However, they do not compare since they were achieved with training on FLAIR data only (versus T1+FLAIR here), with different training databases, and different testing databases (non-complete WMH for Baur et al. 2021b and 4 central slices only for Pinaya et al. (2022b)). The goal here was to propose a fair comparison between the three models.

III.2.5 Conclusion and perspectives

We saw that overall, the performances of the models, especially the AE recons. error and the SAE+*loc*OC-SVM, are quite disappointing because near chance level, despite the task being seemingly easy (hyperintense lesions detection). However, the detailed per-hospital analysis showed heterogeneous performances among the hospitals, which might be due to a domain shift between the hospitals and the training set. The VQ-VAE + Transformer method gave promising results and the qualitative aspect of the score maps of the SAE+*loc*OC-SVM seemed promising too.

WMH (T1+FLAIR) Amsterdam	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM
AU ROC	0.76 ± 0.10	0.62 ± 0.08	0.62 ± 0.15
AU ROC 30	0.54 ± 0.16	0.34 ± 0.10	0.25 ± 0.16
AU PR	0.084 ± 0.103	0.047 ± 0.041	0.015 ± 0.018
AU PRO	0.65 ± 0.06	0.47 ± 0.06	0.41 ± 0.16
AU PRO 30	0.35 ± 0.09	0.15 ± 0.05	0.084 ± 0.127
$\lceil Dice \rceil$	0.13 ± 0.11	0.10 ± 0.06	0.03 ± 0.04
WMH (T1+FLAIR) Singapore	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM
AU ROC	0.73 ± 0.11	0.46 ± 0.03	0.51 ± 0.20
AU ROC 30	0.44 ± 0.15	0.13 ± 0.02	0.19 ± 0.19
AU PR	0.074 ± 0.071	0.018 ± 0.014	0.034 ± 0.045
AU PRO	0.54 ± 0.07	0.45 ± 0.04	0.47 ± 0.20
AU PRO 30	0.17 ± 0.07	0.10 ± 0.04	0.12 ± 0.17
$\lceil Dice \rceil$	0.14 ± 0.11	0.04 ± 0.03	0.06 ± 0.07
WMH (T1+FLAIR) Utrecht	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM
AU ROC	0.58 ± 0.11	0.49 ± 0.06	0.45 ± 0.16
AU ROC 30	0.22 ± 0.12	0.13 ± 0.04	0.12 ± 0.11
AU PR	0.038 ± 0.042	0.019 ± 0.016	0.020 ± 0.017
AU PRO	0.44 ± 0.04	0.58 ± 0.08	0.41 ± 0.14
AU PRO 30	0.07 ± 0.02	0.20 ± 0.07	0.06 ± 0.06
$\lceil Dice \rceil$ *	0.07 ± 0.07	0.04 ± 0.04	0.05 ± 0.04

Table III.6: Mean (± std) metric on every patient from the Amsterdam, Singapore, and Utrecht hospitals for each method. *AU PR* for a random classifier would be 0.003±0.004 for Amsterdam, 0.008±0.006 for Singapore, and 0.008±0.007 for Utrecht. We recall that *AU ROC 30* and *AU PRO 30* for a random classifier would be 0.15.

We further investigate the performance of the SAE+*loc*OCSVM model in the next chapter, and propose a novel support estimation learning framework, with the aim of improving the performance. We can thus see these experiments as a first baseline, and a comparison with state-of-the-art methods on a public neuroimaging database.

III.3 Anomaly detection for *de novo* Parkinson patient classification and characterization

We have seen the performances of the proposed anomaly detection model on a common computer vision anomaly detection task and on a less common medical imaging anomaly detection task. We now wish to see if anomaly detection models could be of any use on a new task which is to discriminate healthy from *de novo* Parkinson patients based on MRI scans. We use the anomaly detectors as a proxy to measure the degree of abnormality of each subject and evaluate the studied methods with classical classification metrics. This study is a proof of concept aiming to evaluate if the proposed UAD model can detect subtle anomalies in different subcortical brain regions that are known to be involved in the pathological mechanism of Parkinson's disease. If the model proves to be sensitive enough, this opens the way to a finer characterization, e.g. questioning which structures are mostly impacted, and how these structures evolve throughout time (assuming we can access longitudinal data, which is the case for the PPMI database), do

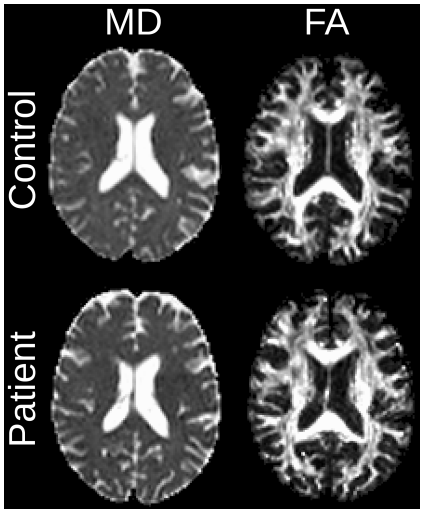

Figure III.9: Showcase of MD and FA diffusion MRI for a random control and a random patient of the PPMI database. There seems to be no visible difference between the two subjects.

all patients evolve the same way (presumably not), etc.

Part of this work has been presented at ISBI 2023 Muñoz-Ramírez, Pinon et al. (2021).

III.3.1 Parkinson's Progression Markers Initiative database

We briefly introduced in section I.2.2.ii the Parkinson's Progression Markers Initiative (PPMI) database. It is a database composed of 3805 patients and controls (as of August 2023), containing healthy controls, recently diagnosed Parkinson patients (*de novo*), and patients that are at risk due to genetic variants or other biomarkers.

From this massive database, 57 healthy control scans and 129 *de novo* Parkinson patients were extracted[1]. This selection was made so that every control and patient had a scanner from the same MRI manufacturer, and had a Diffusion Tensor Imaging (DTI) MR sequence. From DTI, two quantitative maps were extracted: Mean Diffusivity (MD) and Fractional Anisotropy (FA), roughly corresponding to the mean movement of water protons and the asymmetry of such movement.

The MD and FA maps were extracted from DTI with MRTrix3.0, normalized in intensity to the range $[0, 1]$ with the 1% and 99% quantile, and non-linearly registered to the MNI atlas (registration pipeline detailed in appendix B.1), to obtain maps of size $121\times145\times121$ with a voxel size of $1.5\times1.5\times1.5$ mm^3.

To showcase the complexity of the task, we show two slices of MD and FA modality for a random control and a random Parkinson patient on figure III.9. For the naked eye (and even for radiologists), it is very difficult to discriminate between control and patient. Only image-level labels, indicating if a subject is healthy or Parkinsonian, are available.

III.3.2 Methods

For this study, we compare three models: the SAE+*loc*OC-SVM, that we studied in section III.1 and III.2, and compare it to SAE reconstruction error and to image-level auto-encoder reconstruction error, we detail the implementations hereafter.

[1] This extraction was first done in Muñoz-Ramírez et al. (2020), a work that we then continued in collaboration.

III.3.2.i Siamese auto-encoder + *localized* one class SVM

As in section III.1 and III.2, we consider for this study a patch-based siamese auto-encoder + one-class SVM per voxel model, as presented in chapter II section II.3. For chronological reasons, the architecture of the auto-encoder used is the one described in II.3.1 (ConvSiamAlaverdyan) and not the one optimized in III.1.2 (ConvSiamNew1).

III.3.2.ii Patch-based auto-encoder reconstruction error

This method is simply the difference between the mean squared error between the input patch and the reconstructed patch from the siamese auto-encoder presented in II.3.1. Note that each patch then has a reconstruction error value for each of its voxels, but only the central voxel score is kept for the final score map. Patches that are brought closer in the latent space (by the cosine loss) are patches from the same localization but from different healthy controls (see figure II.3 from the previous chapter).

III.3.2.iii Image-level auto-encoder reconstruction error

For this study, we also wished to compare to a reconstruction error computed with an auto-encoder that is not patch-based, and not siamese. The architecture studied has been described in Muñoz-Ramírez et al. (2020). This work is the result of a collaboration with a Grenoble research team from GIN and LJK laboratory. This auto-encoder could not have a siamese counterpart, as there are no patches, thus no different localization (pairing any image from any healthy control together would just bring everything closer in the latent space as there would be no differentiation based on the localization).

This comparison with image-based reconstruction error could allow assessing the benefits of the siamese constraint of the proposed architecture, and also to compare reconstruction error baselines between image-based and patch-based models. The image-level auto-encoder and the siamese patch-based auto-encoder, are presented in figure III.10.

Note that it could not be possible to have the same AE + OC-SVM method on image-based auto-encoders, the feature map obtained in the latent space of an image-based AE, if run one way or another through a OC-SVM to obtain anomaly scores per voxel, would have to be up-sampled to the image size, losing the finer details that we wish to detect in such tasks.

III.3.3 Experiments

As the end task is to classify healthy controls and Parkinson's patients, we have to separate the controls into training and testing sets. We use a 10-fold cross-validation procedure with bootstrapping as advised in Poldrack et al. (2020), leading to each fold containing 41 training controls and 15 test controls. Special care was taken to maintain the same age and sex ratio in the training and testing folds.

The first step of the experiment is to train each unsupervised anomaly detection method, on the train controls only: the auto-encoders learn to reconstruct healthy controls, and the SVM learns the healthy boundary in the latent space. Then, we generated the anomaly score maps on the train controls and assumed that these maps should have less than s=2% anomalies, thus giving a threshold $t_s = t_{2\%}$ that serves to binarize score maps.

Anomaly score maps are then inferred on test controls and patients, binarized with the threshold, and then, the number of detected anomalies are used to classify patients from healthy by varying a threshold above which a subject is considered pathological, thus generating a *ROC* curve. The best achievable geometrical mean (g-mean) between sensitivity and specificity is then used as a performance metric (g-mean = $\sqrt{\text{Sensitivity} \times \text{Specificity}}$).

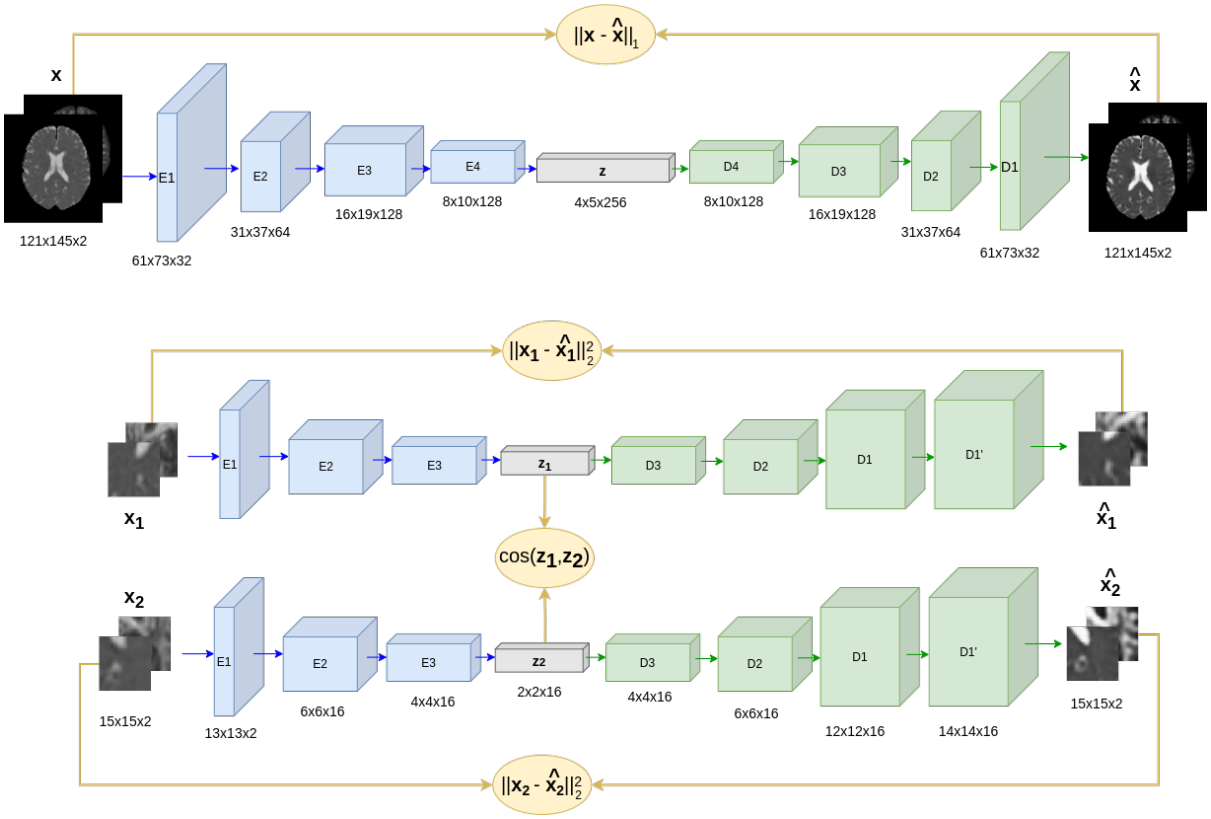

Figure III.10: The two auto-encoder architectures studied, on top the image-level auto-encoder, on the bottom the patch-based siamese auto-encoder. In blue are the encoders' convolutional blocks and in green are the decoders' convolutional blocks.

This analysis is also carried out on each of the brain regions delimited by two atlases, meaning that the healthy versus Parkinson will be classified only with the voxel of said regions. First, the Neuromorphometrics atlas (Bakker et al. 2015), that delimits 8 macro-regions of the brain: subcortical structures, white matter and the 5 gray matter lobes (frontal, temporal, parietal, occipital, cingulate/Insular). Second, the MNI PD25 atlas (Xiao et al. 2015) was specifically designed for PD patients' exploration. It contains 8 regions: substantia nigra (SN), red nucleus (RN), subthalamic nucleus (STN), globus pallidus interna and externa (GPi, GPe), thalamus, putamen and caudate nucleus.

The goal of carrying out a morphological analysis is to investigate whether state-of-the-art machine learning techniques would find such or such structure responsible for *de novo* Parkinson's disease. If abnormalities found in specific areas allow to discriminate between healthy and patients, this could mean that they are responsible for the disease in the early stages of the pathology.

III.3.4 Results and discussion

Figure III.11 presents some reconstructions obtained with the patch-based siamese auto-encoder[1], image-level auto-encoder, as well as the original images. The siamese auto-encoder seems to capture the contrast more precisely than the image-level auto-encoder. Both achieve to reconstruct

[1]Similarly to the reconstruction error obtained, the reconstruction presented is obtained by taking only the central voxel reconstructed for each patch at a time. We found that this gives a less blurry map than reconstructing patches and averaging. Reconstructing side-by-side patches gave 'tiled maps'.

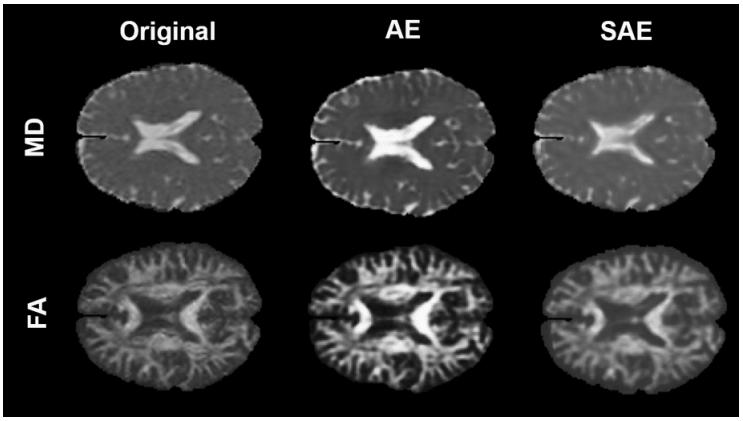

Figure III.11: Showcase of the two MR modalities (MD and FA), for a random control, and its reconstruction by the image-level AE and the siamese patch-based AE.

very small details in the image. Note that this allows to verify that the auto-encoders were trained correctly and that they can compress and reconstruct the images, but this does not give any insight into the potential ability to detect lesions, as they could be able to reconstruct very well fine subtle structures that they did not see during training (as commented in paragraph II.2.2:Reconstruction methods), this also doesn't predict anything about the SVM capabilities.

As there are 10 folds, we obtain 10 values of g-means, that we represent as boxplots. Thus there is one box plot for each brain structure. We present these results in figure III.12. We also present the percentage of anomalies found by the reconstruction error of the siamese model on controls and patients figure III.13.

By the quantitative results we see on figure III.12, we can see that for the three methods, for almost every structure, the mean g-means are above chance (g-mean=0.5), meaning that it is possible to discriminate Parkinson from healthy, and for almost any structures. However, most structures and methods show great variance, indicating great variability of the method depending on the training population.

For this task, the two reconstruction error methods seem on-par on most structures, with a slight advantage for the patch-based auto-encoder. However, the patch-based auto-encoder + localized one-class SVM seems to underperform compared to the reconstruction methods, and this is for any of the presented structures.

We hypothesize that the number of training controls (41) is not sufficient for the one-class SVM, especially considering that the dimension of the latent space is 64.

Other authors such as Correia et al. (2020), report a mean accuracy score for a selection of white matter regions of 0.61 when using supervised SVM to classify between Parkinson's and controls. Schuff et al. (2015) obtained an $AU\,ROC$ of 0.59 for the rostral segment of the SN when using linear models for the classification. While these metrics are not directly comparable, this comparison still reflects that the presented methods are roughly at state-of-the-art performances.

As we used DTI data, we did not search for structural atrophy or lesion load but rather for degradation of white matter properties in the early stages of Parkinson's disease that could appear everywhere in the brain. This partly explains why the white matter obtains the slightly highest g-mean scores.

Note that at this early stage of the disease (1-2 on the Hoehn and Yahr scale[1]) the patients have no tremor or uncontrolled movements compared to healthy controls. This rules out the

[1]The Hoehn and Yahr scale Hoehn and Yahr (1967) is a commonly used scale for classifying Parkinson's Disease progress, which ranks from 1 to 5.

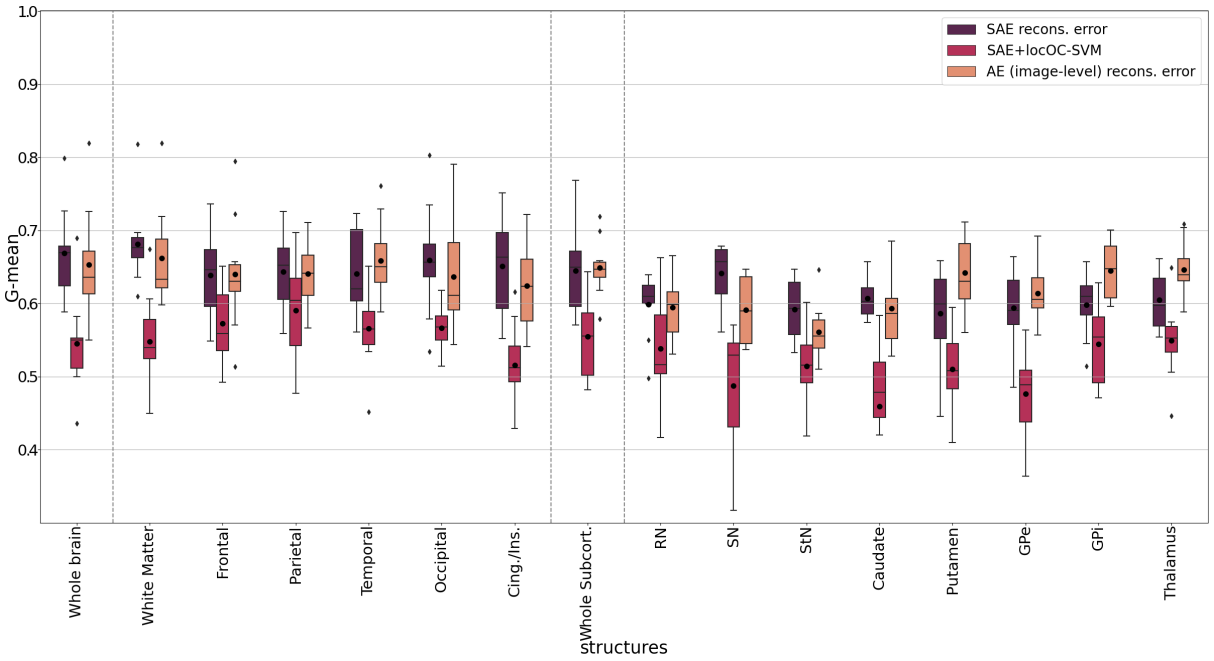

Figure III.12: g-mean scores of the studied methods for the whole brain and several anatomical structures. The vertical dashed lines separate macro and micro brain structures.

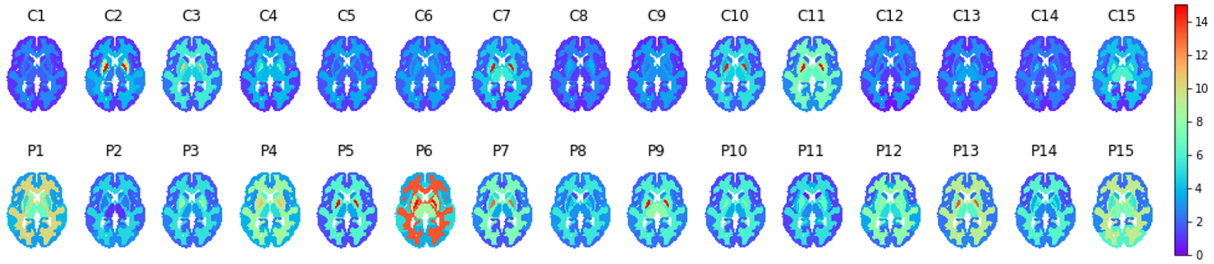

Figure III.13: The percentage of abnormal voxels found by the SAE reconstruction error in the anatomical macro-structures. Top: the test controls of fold 0. Bottom: 15 randomly selected Parkinson's patients

possibility that movement could cause motion blurriness that would have allowed easy classification. The choice of the threshold $s = 2\%$ was not found to have great influence over the results when chosen among the range $]0\%, 10\%]$, in earlier experiments.

III.3.5 Conclusion and perspectives

We saw in this section that unsupervised anomaly detection could be used for the classification of Parkinson's versus healthy. The presented results show above above-chance level for every method. However, the performances are still far from ideal, but the task is very complex, and this disease is still not understood completely. We found that the patch-based approach seemed to have on-par performances when looking at the reconstruction error. While the image-level auto-encoder benefits from a more straightforward implementation, the patch-based auto-encoder can be efficiently trained on smaller databases (as there would be more patches than images). Moreover, the latent space features of this kind of model contain local information that can be used to classify healthy and pathological individuals at the voxel level and produce full-resolution anomaly maps.

We also found that the support estimation method proposed seemed to under-perform on this task compared to the reconstruction methods and that this might be due to the dimension of the latent space that could be reduced (with the architecture proposed in section III.1.2.i for example). A straightforward extension of this study could be to incorporate more MR modalities into the models, and to increase the size of the dataset by adding exams from different MRI vendors. We could also use harmonization procedures as a preprocessing step (e.g. Dewey et al. 2019) to reduce the domain shift between scans from different MRI vendors or with different MRI parameters. We continue the analysis of other methods on the PPMI database in section IV.1.3.

III.4 Conclusion

In this first contribution chapter, we have proposed to extend significantly the evaluation of the model proposed by Alaverdyan et al. (2020).

This was first done on an industrial computer vision database, that is widely used in the anomaly detection community, this allowed comparing the SAE+OC-SVM to other methods based on reconstruction or restoration, to position the method in relation to the state of the art. This also allowed improving the architecture of the auto-encoder that we use in the following chapters. We introduced rigorous evaluation through multiple metrics and qualitative examples that we continue to use in the following chapters.

We then tried to tackle the anomaly detection problem in medical images, by looking at the WMH challenge database. This allowed us to compare the SAE+*loc*OC-SVM to the two other reported state-of-the-art methods on this dataset. As this dataset is public and can be challenging due to the small sizes of the lesions, we re-use it in the following chapter.

We then tried to apply the SAE+*loc*OC-SVM and SAE recons. on a different task: *de novo* Parkinson versus healthy classification. On this complex task, as there are no pixel-level labels, nor certainty about which zone could be lesioned/altered, we showed promising results as we were partly able to detect Parkinson's patients. We wish to extend this study in the following chapters.

Overall, in this chapter, we mainly addressed the shortcomings of the previous studies on the evaluation. We tested the model on three different, public, and challenging databases, with a wide range of metrics. In the following chapters, we propose more methodological contributions to strengthen the performances of support estimation methods, which we believe are relevant for unsupervised anomaly detection in medical imaging.

IV | Patient-specific and robust anomaly detection

We have seen in chapter III that the method proposed by (SAE+*loc*OC-SVM, Alaverdyan et al. 2020), i.e. representation learning through a patch-based auto-encoder and support estimation with one class SVM, seemed to be outclassed by state-of-the-art methods based on density estimation (for MVTecAD, section III.1), restoration (for WMH, section III.2) and reconstruction (for PPMI, section III.3). We saw however that the performances obtained on MVTecAD were close to the state-of-the-art, and the proposed model has several other advantages, such as being lightweight, and support-based[1].

We thus propose in this chapter to improve the existing support estimation methods, by making them patient-specific and more robust. In section IV.1, we first propose a different learning strategy for the one class SVM, that allows learning a frontier that is specific to each patient and that removes the dependence on the limited size of the training set. We show that it allows reaching state-of-the-art performances on the WMH database. In section IV.2, we then propose to study methods that allow to convert distance to probability, with the aim of doing ensemble learning or uniformization, allowing for more robust anomaly detection.

IV.1 Patient-specific anomaly detection

IV.1.1 Inference-time one class SVM

In this section, we propose to explicit more formally the method proposed by Alaverdyan et al. (2020), in section IV.1.1.i, we do so to highlight the differences with the method we propose in section IV.1.1.ii.

IV.1.1.i Details of the reference pipeline

We recall that in section III.2 we used the method proposed by Alaverdyan et al. (2020), presented in section II.3. This method is composed of a patch-based siamese auto-encoder, trained on a set of healthy patches from healthy controls, and then one class SVMs (one per voxel) are used to estimate the support of the healthy distribution of each localization latent representation.

In a more formal way, an auto-encoder (encoder $\mathbf{E}$ and decoder $\mathbf{D}$) is trained on a set of N_{H} healthy images $\mathcal{X} = (\mathcal{X}^h)_{1\leq h\leq N_{\mathrm{H}}}$, indexed with the subscript h for healthy. From these images, patches $\mathbf{x}$ are extracted, from M' localizations of the M coordinate of the brain, with $M' < M$. This produces the set of patches $(\mathbf{x}_i^h)_{1\leq i\leq M', 1\leq h\leq N_{\mathrm{H}}}$. Each patch $\mathbf{x}_i^h$ is then run through the encoder to obtain a latent representation of the patch of coordinate i and patient h: $\mathbf{z}_i^h = \mathbf{D}(\mathbf{x}_i^h)$. Then, M one class SVM are trained, on the set of N_{H} latent representations $(\mathbf{z}_i^h)_{1\leq h\leq N_{\mathrm{H}}}$, leading to M decisions function f_i.

[1] We advocated in section I.1.5.i that density estimation methods solve a more complex problem then they should, and that reconstruction/restoration methods lack interpretability and regularity, as we have no guarantee on the smoothness of the 'normal manifold' learned.

At inference, the N_{P} patient images[1] $(\mathcal{X}^h)_{1\leq p\leq N_{\text{P}}} = (\mathbf{x}_i^h)_{1\leq i\leq M, 1\leq p\leq N_{\text{P}}}$ are used, patch by patch, each patch going through the encoder and the decision function, giving an anomaly score for each voxel of each patient, thus one score map $\mathcal{S}^p$ per patient, i.e. $(s_i^p)_{1\leq i\leq M, 1\leq p\leq N_{\text{P}}} = (f_i(\mathbf{D}(\mathbf{z}_i^p)))_{1\leq i\leq M, 1\leq p\leq N_{\text{P}}}$.

Algorithm SAE+*loc*OC-SVM and figure IV.1 summarizes this procedure.

SAE+*loc*OC-SVM: *Training of the siamese auto-encoder* (Step 1)

Input: Subsample of healthy patches: $(\mathbf{x}_i^h)_{1\leq i\leq M', 1\leq h\leq N_{\text{H}}}$
Output: Trained encoder: E ▷ Decoder is not used in the following

for each epoch **do**
  for every batch of patch $(\mathbf{x}_i^h)_{1\leq i\leq b, 1\leq h\leq N_{\text{H}}}$ **do** ▷ Batch size b
    Train the auto-encoder with the following loss: ▷ Patch from the same location
    $L(\mathbf{x_1}, \mathbf{x_2}) = \sum_{t=1}^{2} ||\mathbf{x}_i^t - \hat{\mathbf{x}}_i^t||_2^2 - \alpha \cdot cos(\mathbf{z}_i^1, \mathbf{z}_i^2)$ ▷ but different patients
  end for
end for

SAE+*loc*OC-SVM: *Training of M One class SVM* (Step 2)

Input: Healthy patches: $(\mathbf{x}_i^h)_{1\leq i\leq M, 1\leq h\leq N_{\text{H}}}$
Input: Encoder: E
Output: M one class SVM decision functions: $(f_i)_{1\leq i\leq M}$

for each voxel localization i **do**
  Encode the patches: $(\mathbf{z}_i^h = \mathbf{D}(\mathbf{x}_i^h))_{1\leq h\leq N_{\text{H}}}$
  Train the one class SVM with the N_{H} latent representations $(\mathbf{z}_i^h)_{1\leq h\leq N_{\text{H}}}$
end for

SAE+*loc*OC-SVM: *Inference* (Step 3)

Input: Encoder: E
Input: M one class SVM decision functions: $(f_i)_{1\leq i\leq M}$
Output: Patient score maps $(\mathcal{S}_i^p)_{1\leq i\leq M, 1\leq p\leq N_{\text{P}}}$

for each patient p **do**
  for each voxel localization i **do**
    Encode the patch: $\mathbf{z}_i^p = \mathbf{D}(\mathbf{x}_i^p)$
    Obtain the score associated to this patch: $s_i^p = f_i(\mathbf{z}_i^p)$
  end for
end for

SAE+*loc*OC-SVM: Algorithm of the method proposed by Alaverdyan et al. (2020) (without post-processing), which we call siamese auto-encoder + *localized* one class SVM.

M' was taken to be 4% of M in the authors' original work, such that the coverage of the brain is sufficient, and the total training time is not hindered by taking all the possible patches. When training the SAE, the patches from the batch are paired randomly with each other (same localization, different patients). Note that when we attribute the anomaly score of a patch to its central voxel, other techniques could be utilized, such as attributing the score to every patch's voxel and averaging the multiple scores obtained per voxel.

[1] Note that patient images are sure to have some pathological patches, but not every patch is pathological. It is quite the opposite: most of the patient's patches are healthy

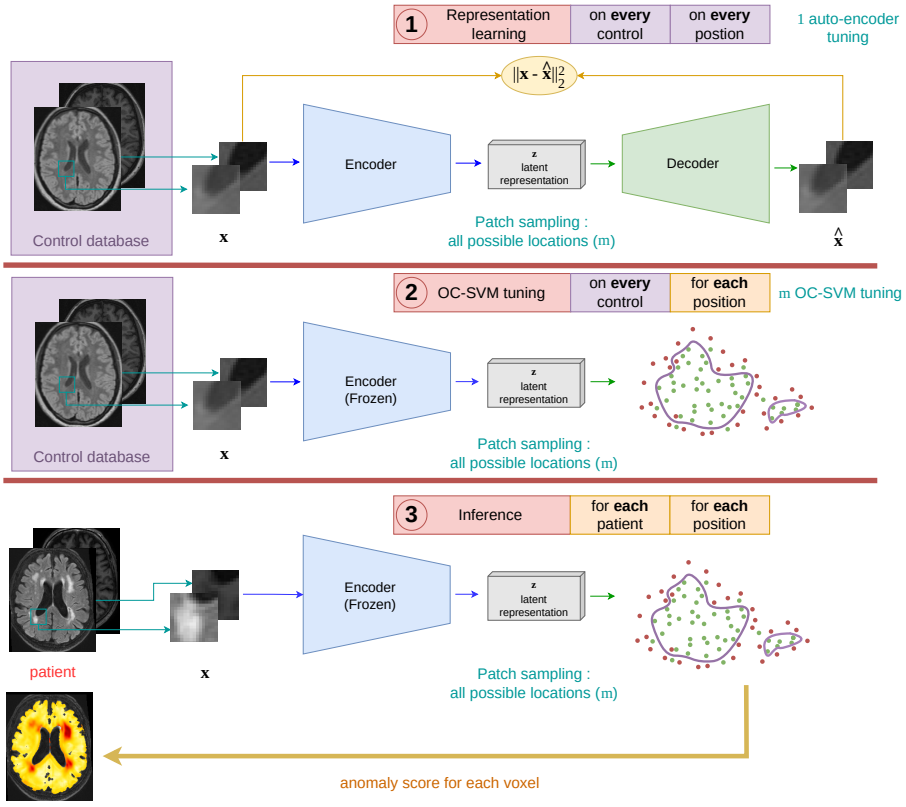

Figure IV.1: Synthetic view of SAE+*loc*OC-SVM, consisting of patch-based siamese auto-encoder training and localized one class SVM.

IV.1.1.ii Patient-specific pipeline: inference-time one class SVM

As we presented in section II.3.5.i, there exist several limits to the reference pipeline, some of them that we recall here: sensitivity to the registration due to the voxel-wise approach, sensitivity to the training set size, no patient-specific characteristics[1] and long optimization.

We propose to address these 4 limits (among others) by proposing the following modification to the pipeline (Step 2 and 3): instead of learning the support of the healthy distribution of each voxel surrounding patch, we propose to learn the support of the healthy distribution of patches randomly drawn in the patient. This implies several things, that address the limits in the same order, first, the frontier learned is not tied to a specific localization, second, as the patches are randomly drawn from the patient, we can increase the training set size, third, the support learned is specific to each patient, fourth as there are significantly less patients than localizations, the optimization time will be shorter. The drawbacks implied by this modification might be a weaker sensitivity due to the complexity of having to learn a unique support for every patch localization.

More formally, the pipeline is the following: the first step of auto-encoder training (step 1) is the same as in the pipeline SAE+*loc*OC-SVM. Then, for each patient image $\mathcal{X}^p$, a subset of n patches are drawn from this patients, from random localizations: $(\mathbf{x}_i^p)_{1\leq i\leq n}$, with $n << M$, encoded into $(\mathbf{z}_i^p)_{1\leq i\leq n}$, and used to train a one-class SVM, with decision function f^p, that is specific to the patient (indexed by p), contrary to the previous pipeline where the decision function was specific to the localization (indexed by i). This decision function is then used on

[1]Such characteristics might be a different MRI machine used or different MRI settings used for the image (hence we would call this 'image-specific') or relative to the specific anatomy of the patient (patient-specific).

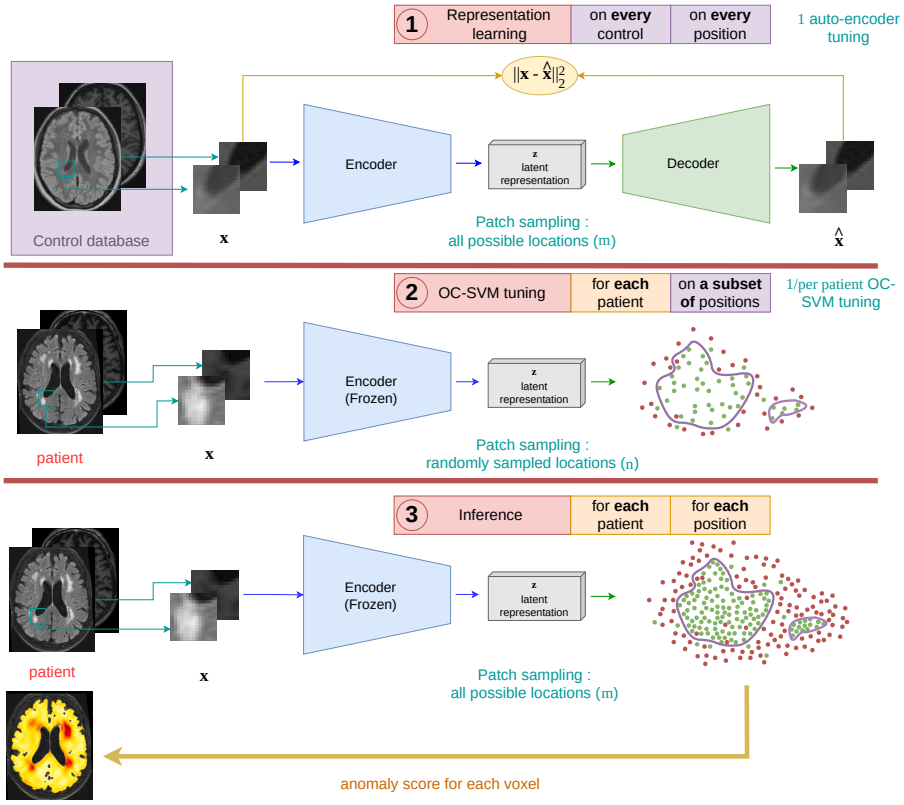

Figure IV.2: Synthetic view of SAE+*ps*OC-SVM, consisting of patch-based siamese auto-encoder training and patient-specific one class SVM.

each voxel localization to obtain the anomaly score map $\mathcal{S}^p$ for this patient. The process is then repeated for every patient image.

Algorithm SAE+*ps*OC-SVM and figure IV.2 summarizes this procedure.

Theory-based rationale: the ν property

There is one important point that we must address: by using patches from the patients to learn the 'healthy frontier', instead of patches from the controls, we will inevitably sample pathological patches, that could influence the learned frontier. One important property of one class SVM is derived in Schölkopf et al. (2001) called the ν property: the hyperparameter ν is an upper bound of the fraction of 'outliers' (and when the number of training samples becomes large, ν is equal to the fraction of 'outliers'). Outliers are points in the training set that are outside the learned frontier, this means that we can have ν% of outliers in the training set without influencing the learned frontier.

Common values of ν are generally pretty low, for instance, $\nu = 0.03$, but one key assumption of anomaly detection (the *concentration assumption*, see I.1.3) is that the anomalies are scarce, thus justifying our approach. For instance, in the WMH dataset, anomalies represent 0.65% voxel of the total volume, which is significantly below 3%.

Additional comments

We also called this new *patient-specific* pipeline *inference-time one class SVM*, indeed, the SVM is not trained on controls, and is only trained at inference, on each new coming patient. This setup is named *outlier detection* in Pedregosa et al. (2011b), where the data is contaminated

SAE+*ps*OC-SVM: *Training of the siamese auto-encoder* (Step 1)

Input: Healthy patches: $(\mathbf{x}_i^h)_{1\leq i\leq M, 1\leq h\leq N_{\text{H}}}$
Output: Trained encoder: E

for each epoch **do** ▷ This step is the same as in algorithm SAE+*loc*OC-SVM
  for every batch of patch $(\mathbf{x}_i^h)_{1\leq i\leq b, 1\leq h\leq N_{\text{H}}}$ **do**
    Train the auto-encoder with the following loss:
    $L(\mathbf{x_1}, \mathbf{x_2}) = \sum_{t=1}^{2} ||\mathbf{x}_i^t - \hat{\mathbf{x}}_i^t||_2^2 - \alpha \cdot cos(\mathbf{z}_i^1, \mathbf{z}_i^2)$
  end for
end for

SAE+*ps*OC-SVM: *Training and inference of* N_{P} *One class SVM* (Step 2)

Input: Encoder: E
Output: Patient score maps $(\mathcal{S}_i^p)_{1\leq i\leq M, 1\leq p\leq N_{\text{P}}}$

for each patient p **do**
  Sample patches from n random localization $(\mathbf{x}_i^p)_{1\leq i\leq n}$
  Encode the patches: $(\mathbf{z}_i^p = \mathbf{D}(\mathbf{x}_i^p))_{1\leq i\leq n}$
  Train the one class SVM with the n latent representations $(\mathbf{z}_i^p)_{1\leq i\leq n}$ to obtain f^p
  for each voxel localization i **do**
    Encode the patch: $\mathbf{z}_i^p = \mathbf{D}(\mathbf{x}_i^p)$
    Obtain the score associated to this patch: $s_i^p = f^p(\mathbf{z}_i^p)$
  end for
end for

SAE+*ps*OC-SVM: Algorithm of the method proposed in this section, which we call siamese auto-encoder + *patient specific* one class SVM.

by outliers and the goal is to identify them. In such a setup, there is no distinction between 'training' and 'testing', the data is presented as is and one must find the outliers in the data. The only difference here is that we subsample this data to train the SVM, but all data are 'test'.

Figure IV.1 and IV.2 present synthetic diagrams of SAE+*loc*OC-SVM and SAE+*ps*OC-SVM.

IV.1.2 Application to hyperintensities detection

Now that we have detailed the method of our contribution, we wish to evaluate the performances of the proposed model. The evaluation is done on the WMH dataset presented in section III.2.1 and we compare the performances to the proposition of Alaverdyan et al. (2020) (localized one class SVM), Pinaya et al. (2022b) (restoration) and Baur et al. (2021b) (reconstruction), as in section III.2.4.
This work has been presented at MIDL 2023 Pinon et al. (2023c).

IV.1.2.i Experiments

As in III.2.3, all models were trained on 75 healthy controls (60 training and 15 validation) from the database partially published in Mérida et al. (2021), which is presented in section III.2.3.i, and tested on the 60 patients of the WMH challenge.

The siamese patch-based auto-encoder has the same architecture as presented in section III.2.2.i. It was trained with 8 750 000 patches of size 15×15×2 (250 000 patches per subject)

for 30 epochs with Adam optimizer Kingma and Ba (2015) with default hyperparameters. The best model selection is based on validation loss and the training batch size is 1000. The one class SVM, for both methods, was used with $\nu = 0.03$ and γ (RBF kernel width) was set such that $\frac{1}{\gamma}$ was equal to the product of the variance and the dimension of the $\mathbf{z}_i^p$. For the patient-specific one class SVM, we sampled $n = 500$ patches (sampling ratio $\frac{n}{m} \simeq 0.02\%$)[1].

We found in the early experiment, that the patient-specific approach seemed to generate a high number of false positives in the cerebrospinal fluid (CSF), in the cerebral ventricles, and near the border of the cortex. To mitigate this effect, we used the FMRIB's Automated Segmentation Tool (FAST, Zhang et al. 2001a) to segment the white and the grey matter, allowing us to exclude some of the CSF from the anomaly maps. We added these post-processed maps as an additional method. Details about this post-processing can be found in appendix B.2.

The experiment, parameters used, and thus results for Baur et al. (2021b) and Pinaya et al. (2022b) were the same as in section III.2.2.iii and III.2.2.ii respectively.

IV.1.2.ii Results and discussion

Table IV.1 present the quantitative evaluation of all the metrics (mean±std[2]) on the 60 patients of the three different hospitals. Figure IV.3 presents some qualitative evaluation, where we see anomaly score maps of the different methods. This figure is the same as III.8 with the added SAE+*ps*OC-SVM plus this very method with CSF segmentation.

We performed non-parametric statistical tests to strengthen the claim that one method might be superior to another (i.e. their mean score on a specific metric). A Kruskal-Wallis test (Kruskal and Wallis 1952) was performed to assess if any method is superior to the others. If it is statistically significant that a model is superior to the others (with $p > 0.01$), a Dunn's test (Dunn 1964) is performed between the best model and the others (pairwise comparisons with Bonferroni correction) to see if the best model is truly superior (with $p > 0.01$). These two tests are non-parametric and thus do not assume any distribution shape of the means.

WMH (T1+FLAIR) 3 hospitals	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	SAE +*ps*OC-SVM	SAE +*ps*OC-SVM + CSF seg
AU ROC	0.69±0.13	0.53±0.09	0.52±0.19	**0.80**±0.09	**0.81**±0.10
AU ROC 30	0.40±0.20	0.20±0.12	0.19±0.16	**0.48**±0.20	**0.59**±0.17
AU PR	0.065±0.079	0.028±0.030	0.023±0.031	**0.084**±0.099	**0.165**±0.168
AU PRO	0.55±0.10	0.50±0.08	0.43±0.17	0.71±0.11	**0.80**±0.07
AU PRO 30	0.19±0.13	0.15±0.07	0.09±0.13	0.33±0.18	**0.48**±0.13
$\lceil Dice \rceil$	0.11±0.10	0.06±0.05	0.05±0.05	**0.14**±0.13	**0.22**±0.17

Table IV.1: Mean (± std) metric on every patient from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.007±0.006. In bold are shown the best model and those for which the statistical difference with the best model for Dunn's test is not significant (p-value ≥ 0.01).

[1] We found in preliminary experiments that the results obtained while varying n (from 300 to 1500) were roughly similar

[2] Recall that the mean and standard deviation are computed among the patients and not among some cross-validation of the model. As the patients have large differences in terms of lesion size, MRI parameters, pathologies, etc., it is not surprising that the variance of the performance can be quite large.

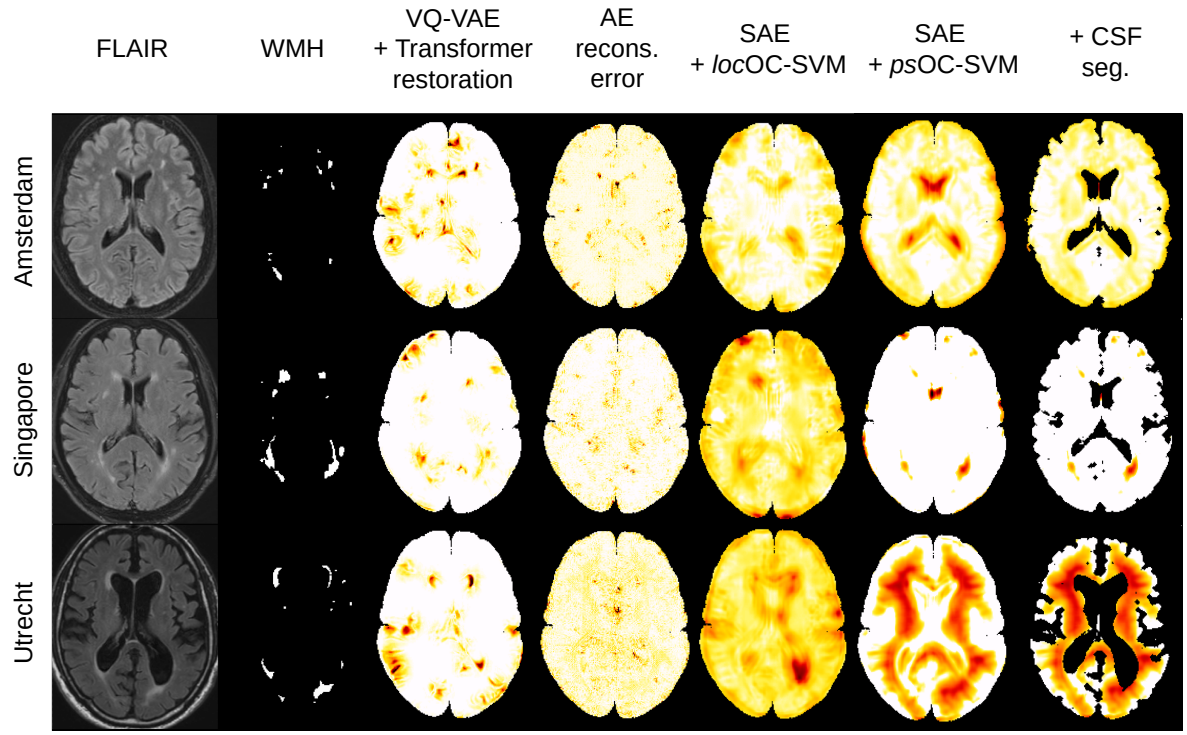

Figure IV.3: Showcase of anomaly maps generated from the five studied methods. Each line illustrates a patient case of each of the three hospitals. The redder the more anomalous. T1 + FLAIR is used as an input for the models but only FLAIR is shown, only ground truth of the WMH are shown. Dynamic range is adjusted for each model to best showcase the anomalies.

Quantitative analysis

Results reported in table IV.1, on the 3 hospitals, show SAE+*ps*OC-SVM outperforms the three methods from Pinaya et al. (2022b), Baur et al. (2021b), and Alaverdyan et al. (2020) (with confidence $p < 0.01$), and that is the case for every metrics proposed. We emphasize that the *AU PRO* and *AU PRO 30*[1] are particularly relevant for this task as they indicate the ability to detect very small lesions. Statistical tests only indicate the superiority of adding a CSF segmentation on the *PRO* metrics.

When looking at the detailed performances per hospital, in table IV.2 we see that every model has a drop in performance when evaluated on the Utrecht hospital. Patients who have been imaged in different centers vary in scanner constructor (GE for Amsterdam, Siemens for Singapore and the control database, and Philipps for Utrecht), MRI parameters, and also average lesional load (3495, 26123 and 29296 mm^3 for Amsterdam, Singapore and Utrecht respectively). All these factors of variations are complex, and it is difficult to understand which one, or which combination of such factors, causes the drop in performances. However, this finding strengthens the need for a more heterogeneous control database and/or the need for domain adaptation, particularly in medical imaging[2].

We see that the CSF segmentation step seems to have very little influence on the metrics, except for *AUPR* and especially for Amsterdam and Utrecht. This seems to indicate that the

[1]Recall that the *AU ROC 30* and *AU PRO 30* for a random classifier would be 0.15, and the AU PR would be 0.007 when considering the 3 hospitals together.

[2]Still, the domain adaptation for medical imaging is a literature niche in itself and is beyond the scope of this thesis.

most anomalous detections are very often in the ventricles or on the cortex border and that removing these regions helps most anomalous detections to be true positives.

We see that the VQ-VAE+Transformer method shows on-par performances on Amsterdam for all metrics, for all but *PRO* metrics on Singapore, and a significant drop on Utrecht, a finding that was missed by the first analysis on the three centers. We believe that such a per-hospital analysis is often overlooked in the literature, and it is of great interest. We also see, quantitatively that on this dataset and for these MR modalities, the proposed method (SAE+*ps*OC-SVM) outclasses the previously studied method (SAE+*loc*OC-SVM).

Qualitative analysis

When looking at the qualitative results of figure IV.3, we see that the seemingly easy task is still difficult for many models: lesions are often missed, and often coarsely found (with large spillover). Moreover, a lot of false positives are generated, particularly in the ventricles (see additional visualization figure IV.4), for our method, initially justifying the need for a CSF segmentation, but that did not conclude in superiority quantitatively[1].

What is also particularly striking, especially with SAE+*ps*OC-SVM, is the anomaly score map ranges variation among hospitals: figure IV.3, last column, we see that anomaly maps are significantly more 'anomalous' on average for Utrecht than Amsterdam, and the same goes when comparing Amsterdam to Singapore. This also strengthens our previous comment on domain adaptation, whether from hospital to hospital or centers to centers (see additional differences between patients from the same hospital in figure IV.4), as the produced anomaly maps vary greatly. This also initiates the need for score map normalization, which we will develop in section IV.2. Figure III.7 shows additional control and patients T1 and FLAIR MRI, showcasing the difference in contrast of all the different centers.

Comparison with the literature

In their original study, Pinaya et al. (2022b) reported a $\lceil Dice \rceil$ of 0.269 and an $AU\ PR$ of 0.158. Baur et al. (2021b) reported $\lceil Dice \rceil$ of 0.45 and $AU\ PR$ of 0.37. These performances were achieved with models trained on FLAIR only, whereas we provided here each model with T1 and FLAIR inputs. This could suggest that the T1 modality only adds confusion and no performance gain, especially since the hyperintensities are clearly visible on FLAIR (and not T1). Additionally, the training database used in Pinaya et al. (2022b) was composed of 15,000 ($>>$ 75) pseudo-healthy FLAIR MRI volumes from the UK biobank database Sudlow et al. (2015). Also, the training and the testing were done on the 4 central slices of the volume only, thus limiting recordings of potential false detections. The training database used in Baur et al. (2021b) was composed of 109 healthy controls ($>$75), and reported the performances on 51 out of the 60 WMH patients.

The performance gap between the reported values and the experiment done here could also be due to the added data augmentation strategy used both in Pinaya et al. (2022b) and Baur et al. (2021b), that included intensity scaling and contrast adjustments.

The main purpose here was to provide a fair comparison between the different models, by training and testing these models on the same dataset as well as using the same evaluation metrics.

[1]The segmentation was fast and easy to obtain, which resulted in a coarse segmentation. We computed that around 8% of the lesions were actually outside the segmented brain, meaning that we cap our performances at 92% of sensitivity when using this post-processing. Thus what is gained in specificity can be lost in sensitivity.

WMH (T1+FLAIR) Amsterdam	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	SAE +*ps*OC-SVM	SAE +*ps*OC-SVM + CSF seg
AU ROC	**0.76**±0.10	0.62±0.08	0.62±0.15	**0.83**±0.10	**0.83**±0.07
AU ROC 30	**0.54** ±0.16	0.34±0.10	0.25±0.16	**0.55**±0.21	**0.65**±0.14
AU PR	**0.084**±0.103	**0.047**±0.041	0.015±0.018	**0.099**±0.130	**0.193**±0.204
AU PRO	**0.65**±0.06	0.47±0.06	0.41±0.16	**0.67**±0.12	**0.77**±0.07
AU PRO 30	**0.35**±0.09	0.15±0.05	0.17±0.09	**0.27**±0.20	**0.43**±0.15
$\lceil Dice \rceil$	**0.13**±0.11	**0.10**±0.06	0.03±0.04	**0.14**±0.15	**0.25**±0.19
WMH (T1+FLAIR) Singapore	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	SAE +*ps*OC-SVM	SAE +*ps*OC-SVM + CSF seg
AU ROC	**0.73**±0.11	0.46±0.03	0.51±0.20	**0.81**±0.09	**0.84**±0.09
AU ROC 30	**0.44**±0.15	0.13±0.02	0.19±0.19	**0.49**±0.20	**0.64**±0.18
AU PR	**0.074**±0.071	0.018±0.014	0.034±0.045	**0.090**±0.085	**0.212**±0.160
AU PRO	0.54±0.07	0.45±0.04	0.47±0.20	**0.75**±0.09	**0.84**±0.05
AU PRO 30	0.17±0.07	0.10±0.04	0.12±0.17	**0.37**±0.16	**0.55**±0.09
$\lceil Dice \rceil$	**0.14**±0.11	0.04±0.03	0.06±0.07	**0.16**±0.12	**0.27**±0.17
WMH (T1+FLAIR) Utrecht	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	SAE +*ps*OC-SVM	SAE +*ps*OC-SVM + CSF seg
AU ROC	0.58±0.11	0.49±0.06	0.45±0.16	**0.76**±0.08	**0.75**±0.10
AU ROC 30	0.22±0.12	0.13±0.04	0.12±0.11	**0.39**±0.13	**0.49**±0.12
AU PR *	0.038±0.042	0.019±0.016	0.020±0.017	0.062±0.070	0.091±0.094
AU PRO	0.44±0.04	0.58±0.08	0.41±0.14	**0.71**±0.11	**0.78**±0.05
AU PRO 30	0.07±0.02	0.20±0.07	0.06±0.06	**0.35**±0.15	**0.46**±0.10
$\lceil Dice \rceil$ *	0.07±0.07	0.04±0.04	0.05±0.04	0.11±0.10	0.14±0.12

Table IV.2: Mean (± std) metric on every patient from the Amsterdam, Singapore, and Utrecht hospitals for each method. *AU PR* for a random classifier would be 0.003±0.004 for Amsterdam, 0.008±0.006 for Singapore, and 0.008±0.007 for Utrecht. We recall that *AU ROC 30* and *AU PRO 30* for a random classifier would be 0.15. In bold are shown the best model and those for which the statistical difference with the best model for Dunn's test is not significant (p-value ≥ 0.01).
*: Non-significant Kruskal–Wallis test (no best model with p-value ≥ 0.01)

Additional study: influence of patch size

As the patch size was an arbitrarily chosen hyperparameter of our method, we wanted to study the influence of such a parameter. We report in table IV.3 the performances of SAE+*ps*OC-SVM, without CSF segmentation, for 9×9, 21×21 and 27×27, to complete the already reported results with patch size 15×15.

Note that for the 9x9 experiment, we had to tweak the auto-encoder by removing the max-pooling and upsampling blocks. We find almost no significant difference between the different studied patch sizes, indicating that this hyperparameter is not crucial in the success of SAE+*ps*OC-SVM.

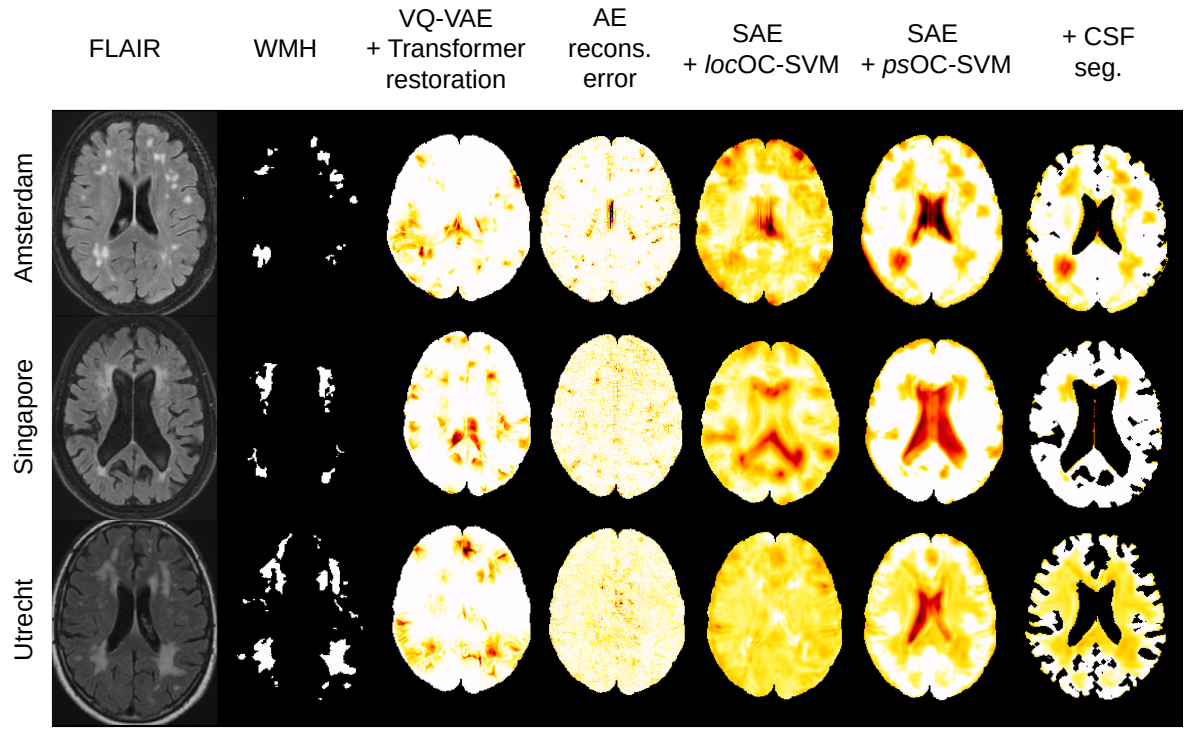

Figure IV.4: Additional showcase of the studied methods anomaly maps on three slices from three patients (AM114, SIN63 and UT37) from one of each hospital (as in figure IV.3), Baur et al. (2021b), Pinaya et al. (2022b) and ours, redder means anomaly score higher. T1 + FLAIR is used as an input for the models but only FLAIR is shown, only ground truth of the WMH is shown.

WMH (T1+FLAIR) 3 hospitals	SAE +*ps*OC-SVM 9×9	SAE +*ps*OC-SVM 15×15	SAE +*ps*OC-SVM 21×21	SAE +*ps*OC-SVM 27×27
AU ROC *	0.81±0.14	0.80±0.09	0.78±0.09	0.75±0.12
AU ROC 30 *	0.53±0.25	0.48±0.20	0.48±0.16	0.41±0.19
AU PRC *	0.150±0.171	0.084±0.099	0.094±0.098	0.056±0.062
AU PRO	**0.70**±0.14	**0.71**±0.11	**0.71**±0.11	0.61±0.13
AU PRO 30	**0.34**±0.20	**0.33**±0.18	**0.35**±0.16	0.22±0.15
⌈ Dice ⌉ *	0.20±0.18	0.14±0.13	0.15±0.13	0.10±0.09

Table IV.3: Mean (± std) metric on every patient from the 3 different hospitals for each patch size. In bold are shown the best model and those for which the statistical difference with the best model for Dunn's test is not significant (p-value ≥ 0.01).
*: Non-significant Kruskal–Wallis test (no best model with p-value ≥ 0.01)

IV.1.3 Application to Parkinson *de novo* classification

We wish to the analysis of the PPMI database that was initiated in section III.3. The goal is to evaluate the performance of the proposed patient-specific method to classify healthy and Parkinson's patients, as proof of concept aiming at evaluating if the UAD model can detect subcortical anomalies in the brain regions known to be involved in Parkinson's disease. In this section, we also wish to extend the comparison of the developed models with a density estimation model. This work is the result of a collaboration with a Grenoble research team from GIN and LJK laboratory. We also extend the comparison to fully supervised models.

The database used, pre-processing, post-processing, and metrics evaluated are the same as in section III.3. This work has been presented at an international conference Pinon et al. (2023a).

IV.1.3.i Methods

We compare five methods for this study: two fully supervised methods, described hereafter, and three unsupervised methods, each based on the patch-based siamese auto-encoder described in section III.1.2.i (architecture ConvSiamNew1). Figure IV.6 presents the three UAD methods.

Reconstruction method

This is simply the mean squared error, voxel-wise (by taking the reconstruction of the central voxel of each patch), between the input patch and its reconstruction. We already studied this method for the MVTecAD dataset in section III.1.3.i, and applied it to PPMI in section III.3.2.ii. We call this method SAE recons.

Support estimation method

We study the SAE+*ps*OC-SVM method, presented section IV.1.1.ii.

Density estimation method

The main idea is to use the SAE to encode patches from the image, and to estimate the density of the healthy patches with a probability distribution generalizing the mixture of Gaussians.

For SAE+*ps*OC-SVM, we trained one SVM per voxel, on every healthy control. For SAE+*ps*OC-SVM we trained one SVM per patient, on any voxel. Here, we train the density estimation on all voxels of every healthy control. This choice is justified by the fact that the shape of the distribution can be very complex, as we see hereafter, and that it is a mixture of K components. Thus the goal is to learn a global distribution of the healthy control patch distribution, that is not specialized to one localization as for SAE+*loc*OC-SVM or specific to each patient as for SAE+*ps*OC-SVM. One advantage of such a strategy is to increase considerably the number of training samples (as many as brain localization $\times$ number of controls) but at the cost of having a model that doesn't take into account the localization of the patch.

The different healthy controls patches latent representations $\mathbf{z}_i^h = \mathbf{D}(\mathbf{x}_i^h)$ and their associated probability distribution $p(\mathbf{z})$ is estimated by the mixture distribution $\hat{p}(\mathbf{z}; \mathbf{\Theta})$ of parameters $\mathbf{\Theta}$. It takes the form:

$$\hat{p}(\mathbf{z}; \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k f(\mathbf{z}; \Phi_k) \qquad \text{with} \sum_{k=1}^{K} \pi_k = 1$$

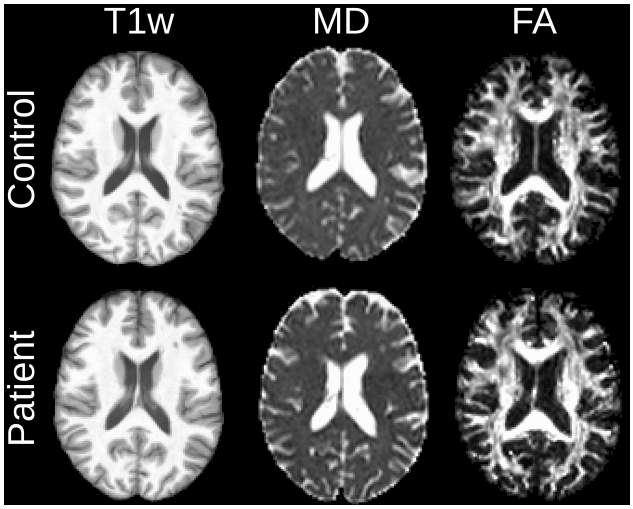

Figure IV.5: Showcase of the three modalities used for experiments in this section, T1, MD, and FA, from a random control and patient extracted from the PPMI database.

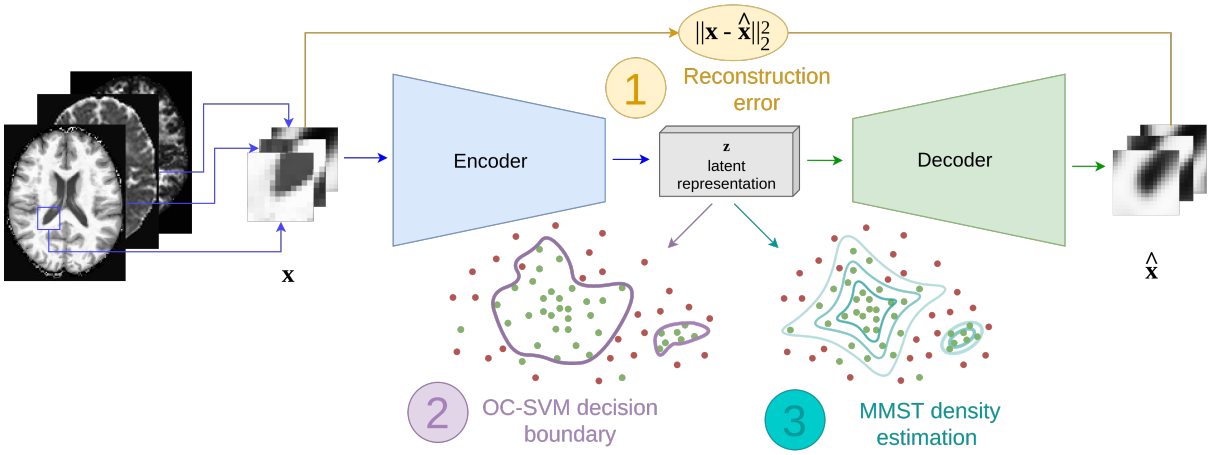

Figure IV.6: Diagram of the three UAD methods compared in this section. The patch-based auto-encoder is used to compute the reconstruction error and to extract latent representation that is used to learn the decision boundary for the SVM and probability density for the mixture of $\mathcal{MST}$.

Where π_k is the mixture weight (or proportion) of the component k, $\mathbf{\Theta}$ the whole set of parameters of each probability distribution $f(\mathbf{x}; \Phi_k)$. Forbes and Wraith (2014) proposed a generalization of the multivariate t-distribution, called $\mathcal{MST}$ for multiple scale t-distributions. The standard univariate scale variables are replaced with a D-dimensional scale variable $(W_d)_{1 \leq d \leq D} \in \mathrm{R}^D$. This notably allows for more complex shapes, beyond elliptical distributions as in the Gaussian case.

The estimation of the parameters of the distribution, theoretically feasible with the EM algorithm Dempster et al. (1977), is too time and memory-costly in practice when the amount of data is large. An online version of the EM algorithm has thus been developed in Oudoumanessah et al. (2022), and notably applied in Oudoumanessah et al. (2023) to scale this estimation to a large amount of data. We refer to these two publications for additional details on the method.

Fully supervised methods

In section III.3 we compared three unsupervised anomaly detection methods and used the detected anomalies to classify between healthy and Parkinson's. We were curious about the potential performances of methods that were fully supervised. To this end, we trained two networks (3D ResNet with 18 layers Tran et al. 2018 and DenseNet-264 Huang et al. 2019), to directly classify healthy controls from patients. Thus each network takes as input the whole 3D volume and a dense layer was added at the end to obtain a one-dimensional output indicating with a binary label whether healthy or Parkinson.

On one hand, supervised methods could be expected to have a significant advantage over unsupervised methods, but on the other hand, the task is very complex, with very few inductive biases, thus the task may be too complex for a one-step method.

IV.1.3.ii Experiments

To summarize, we compare two supervised deep learning models, trained directly for classification to three unsupervised models, trained to detect anomalies, and then post-processed as described in section III.3.3, where the anomalies are used for the classification. As in section III.3.3, we perform this analysis per anatomical structure, to infer if any of such structures could be responsible for Parkinson's disease. The per-structure analysis is not possible with the supervised models as take as input the whole 3D volume.

To extend the study we presented in section III.3.3, we add the T1 modality, available to all methods as an additional input channel (concatenated to MD and FA inputs), this addition decreased the number of controls from 57 to 54 and the number of patients from 129 to 124 (as the T1 scan was not available for these patients). The three unsupervised methods are based on the siamese patch-based auto-encoder presented in section III.1.2, whether used for reconstruction error, support estimation, or density estimation. Figure IV.6 represents the three studied methods. As the image-level auto-encoder did not outperform the patch-based auto-encoder (in the study presented section III.3), we did not consider this architecture for this experiment.

As the supervised methods need to be trained on controls and patients, we separate the control database into training and testing (as in section III.3.3) and the patients into training and testing. Even if the unsupervised methods don't need training patients, we still remove them from the testing set, to have the same testing set for every method. This splitting, done on 10-fold cross-validation with bootstrapping, gave [39, 41] train controls and [13, 15] test controls, [36, 40] train patients and [82, 86] test patients. The true number depends on the exact fold. Special care was put into balancing the age and sex distribution of each fold.

Every hyperparameter used are the same as in section IV.1.2.i. For the mixture of $\mathcal{MST}$, $K = 9$ was chosen.

IV.1.3.iii Results and discussion

We present figure IV.7 the boxplots of the g-means obtained. Figure IV.8 presents a 3D visualization of the obtained anomalies for the unsupervised methods on subcortical structures. In this qualitative example, we see that each unsupervised method seems to have its own range of number of anomalies (e.g. one class SVM has more anomalies than reconstruction error), but the three methods seem to detect more anomalies for the patients.

In the quantitative analysis, we see that the three unsupervised models achieve a median g-mean score of around 0.65 on the whole brain, whereas the supervised models achieve a median g-mean score in the range [0.55, 0.6]. This result validates the use of unsupervised anomaly

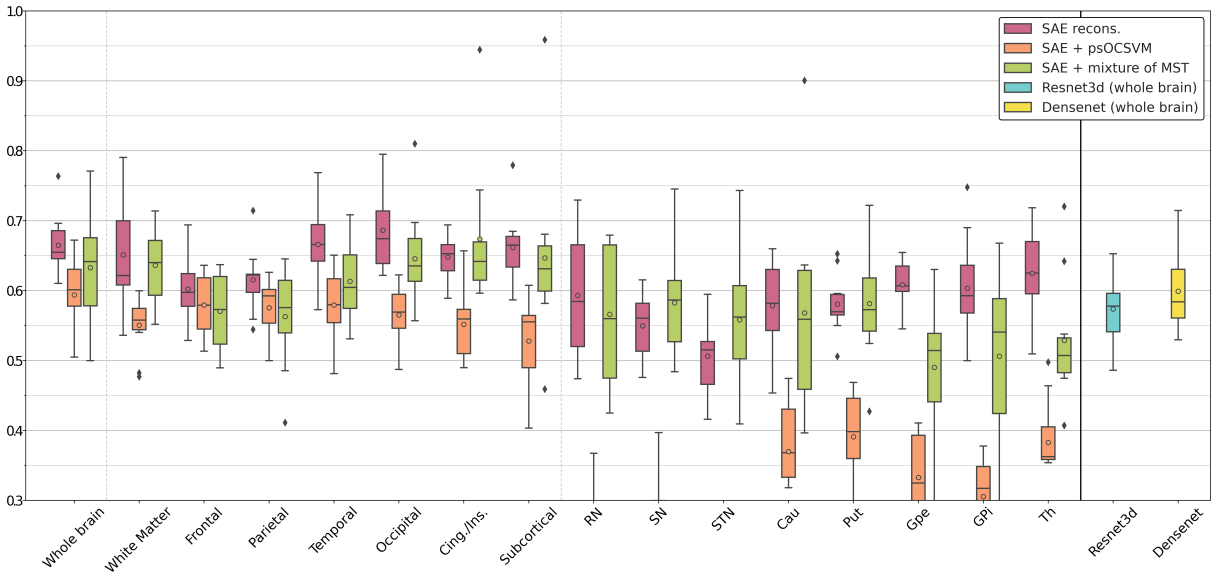

Figure IV.7: g-mean scores of the studied methods for the whole brain and several anatomical structures including the 8 subcortical structures from the MNI PD25 atlas: substantia nigra (SN), red nucleus (RN), subthalamic nucleus (STN), globus pallidus interna and externa (GPi, GPe), thalamus, putamen and caudate nucleus. Resnet3D and Densenet are computed on whole brain.

detections for Parkinson's versus healthy classification. We hypothesize that this task is too complex, the lesions too subtle to be handled by large 3D classification models.

The median g-means of the unsupervised methods are higher on the macro-structures than on the micro-structures, especially for the SVM method. The discrimination of Parkinson based only on subcortical structures seems barely feasible (g-means just above 0.5, with a large standard deviation), this conclusion is also reported in Prasuhn et al. (2020). The overall conclusion on performances per structure is roughly the same as in section III.3, thus we hypothesize that the T1 modality addition does not change the ability of the UAD models to find anomalies able to discriminate Parkinson patients from controls.

Overall the reconstruction error seems to produce higher g-means than SVM and $\mathcal{MST}$, but the variance is very high among the cross-validation folds (and for every structure) thus we highly doubt that these differences are statistically significant. We also find that using a patch-based encoder, as a feature extractor to feed a $\mathcal{MST}$ model, gives promising results as it allows capturing some spatial context, which was lacking in a previous study their (Arnaud et al. 2018a).

We used here SAE+*ps*OC-SVM, i.e. the patient-specific approach. The SVM method used in the previous Parkinson experiment (section III.3), was SAE+*loc*OC-SVM. The performances of the patient-specific approach is higher (approximately +0.1 g-mean on the whole brain), but the experiments don't compare, as there was an addition of T1 modality in this section and a reduction of around 30% of the number of patients. Additional experiments would be necessary to compare the patient-specific/localized approach.

In this section, we were interested in comparing three unsupervised anomaly detection methods on a Parkinson classification task: a reconstruction method, a support estimation method, and a density estimation method. All three were applied downstream of the patch-based siamese network. We saw no clear superiority between the methods, but once again saw that this classification task was possible. We saw that the classification based on the anomalies detected

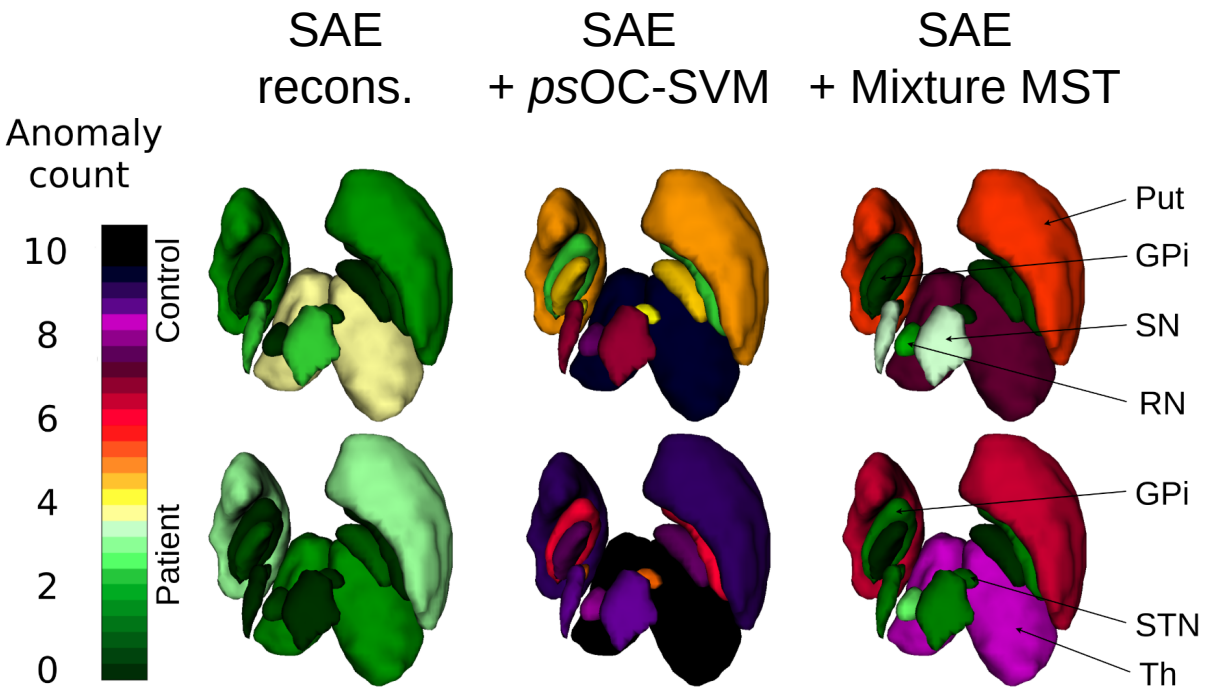

Figure IV.8: 3D representation of the number of abnormal voxels per brain subcortical structures only.

gave a greater g-mean than fully supervised CNN, showing the great complexity of the task. Further analysis could be done to stratify the evaluation concerning the severity of Parkinson's disease, to see if patients at a more advanced stage of the disease get higher anomaly scores. The performances could also be greatly increased by adding other MR modalities such as T2/T2*, which captures the iron load and could allow us to detect the reduction of dopaminergic neurons, which is reported in the early stage of the disease but not visible in diffusion MRI.

IV.1.4 Conclusion and perspectives

In this section, we proposed a new pipeline (algorithm SAE+*ps*OC-SVM), that changes the training samples used of the one class SVM, allowing to building of a patient-specific boundary at inference.

We first tested this method on the WMH dataset, which has visible lesions (hyperintense) but that can be very small. We demonstrated the superiority of the SAE+*ps*OC-SVM, against the localized SVM approach (SAE+*loc*OC-SVM), and against the two state-of-the-art methods on the WMH dataset.

We then extended the *de novo* Parkinson classification study, by comparing the proposed patient-specific algorithm to the classical reconstruction error and a density estimation method. These three candidates gave similar performances, on a task that we believe is fairly complex, as there is no lesion visible (and no pixel-level ground truth). We outperformed two fully supervised CNNs for this task, proving the complexity of said task, and proving that looking at anomalies detected by UAD models can be a strong inductive bias for Parkinson's versus healthy classification.

The patient-specific proposed method is also faster to optimize[1], removes the dependency on

[1] As many SVM as the number of subjects to test (60 for the WMH experiment) versus as many SVM as the

the size of the training set (as we can sample many patches in the brain), allows some kind of independence from the registration (in the outlier detection phase at least), and allows multiple SVM with different training sets being tuned (which we could then fuse, as we will see in section IV.2.2.i).

Despite many of the advantages of the SAE+*ps*OC-SVM, we believe it also has its limits. Compared to a model that is specific to brain localization, we believe the patient-specific boundary learned will be simpler, as it has to represent every possible localization, and could not encompass complex anatomical differences. Figure IV.9 gives a visual depiction of this hypothesis. We believe that on the WMH dataset, the fact that the lesions were hyperintense greatly helped this model's performance. As we will see in V, when not having these visual clues, this model might underperform in terms of sensitivity when compared to a more localized one.

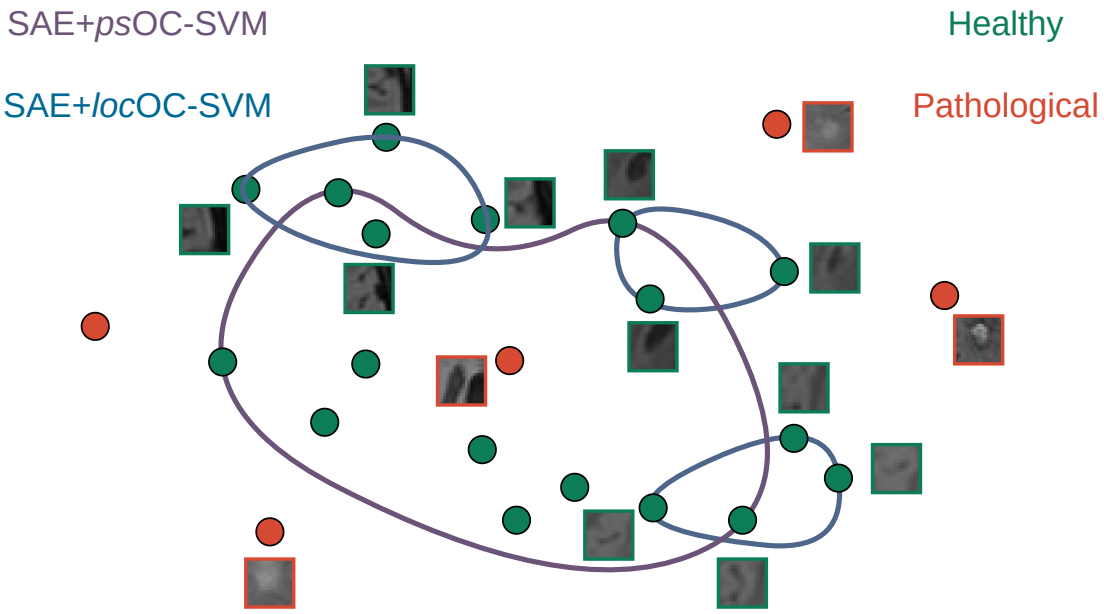

Figure IV.9: Diagram of the hypothesized support boundaries estimated by the algorithms SAE+*loc*OC-SVM and SAE+*ps*OC-SVM. The SAE+*loc*OC-SVM algorithm estimates multiple supports, one per localization, allowing for sparse but precise support estimates. On the other hand, the SAE+*ps*OC-SVM algorithm estimates a single support for every localization, which could oversimplify the estimate. In this depiction, SAE+*ps*OC-SVM accurately detects hyperintense lesions but fails to identify a non-hyperintense pathological area located in the center of its large frontier. Conversely, the sparsity of the estimates of SAE+*loc*OC-SVM may result in poor generalizability.

IV.2 Robust anomaly detection

We have detailed in section IV.1 a first contribution based on the training of the one class SVM, that allows faster optimization, less dependence on the training set size, the quality of the registration, and has patient-specific characteristics.

We saw however that the produced anomaly score maps can vary greatly in range from center to center (or even patient to patient). Furthermore, generated score maps, whether from

number of voxels in the brain ($\simeq$1.5million). The 60 SVM are each trained with 500 sampled patches, and the $\simeq$1.5million are trained with $\simeq$60 controls. A detailed computation of the complexity is derived in appendix D.

the per-pixel or per-patient approach, have no guarantee of being homogeneous[1]. This lack of homogeneity (or uniformization or calibration) reduces the ability to merge score maps obtained from various modalities, methods, or different subsampled patches of the same image. This, in turn, hinders the robustness of such methods. It also reduces the interpretability of the output score maps.

In this section, we first introduce two methods that allow probabilistic outputs for SVM (section IV.2.1), and then use these methods to perform ensemble learning (section IV.2.2.i) and score map uniformization (section IV.2.2.ii).

IV.2.1 Probabilistic outputs for one class support vector machines

We first introduce a method that converts one class SVM outputs to probability *a posteriori*, meaning that it only acts as an additional step performed after the one class SVM is trained.

In the second part, we introduce a more complex method, that involves changing the optimization problem of the SVM, but that offers more theoretical guarantees that the probability output is well calibrated.

IV.2.1.i Binning of one class SVM

In Que and Lin (2023)[2] the authors introduce two methods to recover probabilities *a posteriori* from one class SVM outputs, which we present in the following one of these.

Binning by decision values

This method is pretty simple and consists of binning the different decision function values of the training set. Three assumptions are made: the most negative decision function value has a probability of 0 of being normal, the most positive value has a probability of 1 of being normal, and the probability of points lying exactly on the support frontier is 0.5.

Then, assume $b = 5$ one-side bins: there will be 5 bins per side plus the central bin so 11 bins in total, the probabilities[3] of these bins of score m_j $(j = 1, \dots, 2b + 1)$ will be $p(m_j) = \frac{j}{2b}$:

$$p(m_0) = 0, \quad p(m_1) = 0.1, \quad \dots, \quad p(m_5) = 0.5, \quad \dots, \quad p(m_9) = 0.9, \quad p(m_{10}) = 1$$

Here the m_j represents the center of a bin, and the m_j are computed as follows, to ensure the bins represent the density of the decision functions:

- m_j, $j = 0, \dots, b - 1 = 4$ are the ranked $5 \times j$ percentile of the negative scores
- m_j, $j = b + 1 = 6, \dots, 2b = 10$ are the ranked $5 \times j$ percentile of the positive scores
- $m_{b=5} = 0$, such that a score of 0 will lead to a probability 0.5 of being normal

[1] When using one-class SVM, for instance, the distance to one hyperplane obtained by a one-class SVM, has no guarantee of being comparable to a distance from another hyperplane, obtained with different samples (thus different support vectors). This is the case for nearly every support estimation or reconstruction method.

[2] This article, available as a pre-print, was written by the authors of LIBSVM (Chang and Lin 2011), a popular SVM library, which source code is used for SVM related computations in Scikit-learn (Pedregosa et al. 2011b). In a personal correspondence with the authors, we were informed that the article was under submission in a journal.

[3] These probabilities, e.g. $p(f_i)$ are the probabilities of being normal given the anomaly score, i.e. $p(\text{normal}|f_i)$.

Equivalently, by defining δ_j, the bins edges as:

$$\delta_i = \frac{m_j + m_{j+1}}{2}, \quad j = 0, \dots, 2b - 1 = 9$$

For a given anomaly score f_i, searching for the closest m_j amount to find in which of these intervals the score falls:

f_i	$]-\infty, \delta_0[$	$[\delta_0, \delta_1[$	$[\delta_1, \delta_2[$	$\dots$	$[\delta_9, +\infty[$
$p(f_i)$	0	0.1	0.2	$\dots$	1

The advantages of such a method are that it is non-parametric, and based on the density of the training anomaly scores.
One major drawback is that it does not account for the value of ν. Indeed: ν indicates that the frontier should contain $1-\nu$ of the probability mass of the normal data distribution. It is proven in Vert et al. (2006) that the one class SVM is a consistent estimator of the maximum volume set of probability mass $1 - \nu$ (for $n \to \infty$), meaning that ν is an estimate of the probability density function of points lying at the decision boundary if n is large enough. This means that this method should consider $p(m_b) = \nu$ instead of $p(m_b) = 0.5$.
Another drawback is that this calibration is obtained based only on a fitting of a single one-class SVM, with a single value of ν, whereas multiple estimates could provide more information if combined rightfully.

IV.2.1.ii Concentric SVDD

In El Azami et al. (2017), the authors propose to modify the optimization of the SVM problem, to obtain scores that can be converted to probability with more theoretical guarantees, that take advantage of estimating multiple frontiers for different ν. We detail the method hereafter.

Reminder on Support Vector Data Description

We have seen in section I.3.1.iii:Support vector data description (SVDD) a description of the SVDD Tax and Duin (2004) algorithm. This algorithm, very similar to the one class SVM, finds the ball $(\mathbf{c}, R)$ enclosing the data in a feature space obtained by a transformation $\mathbf{\Phi}(.)$ (see figure IV.10 for a visual example without transformation). If using the RBF kernel, the SVDD and the one class SVM algorithm can be proven rigorously equivalent Schölkopf and Smola (2002).

The SVDD primal problem, with kernel feature space transformation, is presented below in equation IV.1.

$$\begin{aligned} \min_{R,\mathbf{c},\xi} \quad & R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & ||\mathbf{\Phi}(\mathbf{z}_i) - \mathbf{c}||^2 \geq R^2 + \xi_i \quad i \in [1, n] \\ & \xi_i \geq 0 \qquad\qquad\qquad\qquad i \in [1, n] \end{aligned} \tag{IV.1}$$

where slack variables ξ_i are introduced to relax the optimization problem, and the hyper-parameter ν controls this fraction of points not enclosed in the ball.

By using the Lagrange multiplier method, we can show that the dual problem is the following:

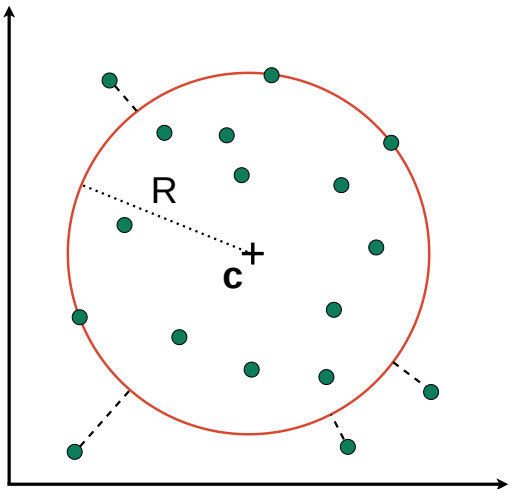

Figure IV.10: Depiction of a 2-dimensional linear (no feature mapping) SVDD, with training data points $\mathbf{x}_i$ in green, fitted center $\mathbf{c}$ and radius R. Note that some points are inside the ball (inlier), some are exactly on the boundary (essential support vectors) and some are outside the boundary (outliers or non-essential support vectors).

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

The decision function for SVDD then takes the form of:

$$f(\mathbf{z}) = R^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\mathbf{z}_i,\mathbf{z}_j) + 2\sum_{i=1}^{n}k(\mathbf{z}_i,\mathbf{z}) - k(\mathbf{z},\mathbf{z}).$$

We see here that indeed when using the RBF kernel, i.e. $k(\mathbf{z}_i,\mathbf{z}_j) = e^{-\gamma||\mathbf{z}_i-\mathbf{z}_j||^2}$, the second term of the optimization problem will be constant, and we can prove that the decision function is the same as for one class SVM.

Concentric Support Vector Data Description

The concentric SVDD (cSVDD) was proposed in El Azami et al. (2017), as a way to train multiple SVDD with different ν, but within the same optimization problem, such that all estimated support boundaries are nested, and every ball has the same center in the feature space. This enables the use of probability calibration, which is based on the consistency property of SVDD as a density-level set estimator of probability.

The goal of cSVDD is to estimate q SVDD, each associated with a specific ν_j, such that $\nu_1 < \nu_2 < \cdots < \nu_q$, meaning each will be associated with a radius R_j, but all sharing the same center $\mathbf{c}$ (hence the name 'concentric'). Sharing the same center allows to have nested radii $R_q < \cdots < R_2 < R_1$ (proven in El Azami et al. 2017).

The primal convex optimization problem is the following:

$$\begin{aligned} \min_{\mathbf{c},\mathbf{R},\underline{\underline{\boldsymbol{\xi}}}} \quad & \sum_{j=1}^{q} R_j^2 + \sum_{j=1}^{q} \frac{1}{\nu_j n} \sum_{i=1}^{n} \xi_{ji} \\ \text{subject to} \quad & ||\Phi(\mathbf{z}_i) - \mathbf{c}||^2 \leq R_j^2 + \xi_{ji} \qquad i \in [1,n] \quad j \in [1,q] \\ & \xi_{ji} \geq 0 \qquad\qquad\qquad\qquad\quad i \in [1,n] \quad j \in [1,q] \end{aligned} \tag{IV.3}$$

The cSVDD objective function is a sum over all q SVDD objective functions, thus all the SVDD are jointly optimized. Again by using the Lagrange multiplier method we arrive at the dual problem (derivation can be found in appendix A.1 or in El Azami et al. (2017)):

$$\begin{aligned} \min_{\underline{\underline{\boldsymbol{\alpha}}}} \quad & \frac{1}{q} \sum_{j=1}^{q} \sum_{k=1}^{q} \sum_{i=1}^{n} \sum_{l=1}^{n} \alpha_{ji} \alpha_{kl} k(\mathbf{z}_i, \mathbf{z}_l) - \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} k(\mathbf{z}_i, \mathbf{z}_i) \\ \text{subject to} \quad & 0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n} \qquad\qquad j \in [1,q] \quad i \in [1,n] \\ & \sum_{i=1}^{n} \alpha_{ji} = 1 \qquad\qquad\qquad\quad j \in [1,q] \end{aligned} \tag{IV.4}$$

We recognize the same structure as for SVDD, with additional sums on q. The decision function associated with each ν_j is the following:

$$f_j(\mathbf{z}) = f_{\mathbf{c}}(\mathbf{z}) - R_j^2$$

with $f_{\mathbf{c}}(\mathbf{z}) = ||\boldsymbol{\Phi}(\mathbf{z}) - \mathbf{c}||^2$ the distance to the center:

$$f_{\mathbf{c}}(\mathbf{z}) = k(\mathbf{z},\mathbf{z}) - \frac{2}{q} \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} k(\mathbf{z}_i, \mathbf{z}) + \frac{1}{q^2} \sum_{j=1}^{q} \sum_{k=1}^{q} \sum_{i=1}^{n} \sum_{l=1}^{n} \alpha_{ji} \alpha_{kl} k(\mathbf{z}_i, \mathbf{z}_l)$$

The first term is equal to 1 in the case of the RBF kernel, and the third term does not depend on $\mathbf{z}$ and can be pre-computed.

Additional comments on the nested property of cSVDD

It has been shown in Vert et al. (2006) that one class SVM with RBF kernels (or equivalently SVDD) are consistent estimators of density level sets of probability mass $1 - \nu$. One could say that it would suffice to train q SVDD independently, with the associated ν_j, to obtain q level-sets of probability mass $1 - \nu_j$, and then do the probability calibration on these.

However, it has been proven in Lee and Scott (2007) that the generated level sets are *not* nested, meaning calibration to probability would not be possible, as the probability as a function of the decision score would not be monotone.

Conversion to probabilistic output

As it is done in El Azami et al. (2017), we now have q nested density level set estimators.

The support vectors associated with a specific ν_j are as follows (see Schölkopf and Smola 2002):

$$\mathbf{SV}_j = \{\mathbf{z}_l \mid 0 < \alpha_{jl} < \frac{1}{\nu_j n}\}$$

We know that for any support vector $\mathbf{sv}_{jl} \in \mathbf{SV}_j$, its associated probability of being normal should be equal to ν_j, thus we should set, for every j:

$$p(\mathbf{sv}_{jl}) = \nu_j \qquad j \in [1, q] \qquad \forall l$$

As the distance to the center is the same for every support vector having the same j, it does not matter which support vector we take for such a fitting (index l). However, for numerical stability, the fitting is done with every available support vector.

A simple sigmoid fitting is then done on the ordered pairs $(f_{\mathbf{c}}(\mathbf{sv}_{jl}), \nu_j)$, on at least q points, to obtain a function mapping a distance to the center to a probability:

$$p(\mathbf{z}) = \frac{1}{1 + e^{-Af_{\mathbf{c}}(\mathbf{z})+B}}$$

with A and B the slope and intercept of the sigmoid, respectively.

A more complex fitting, with for instance more complex functions could be done, but having such strong guarantees on the density level sets, if q is high enough, we believe the calibration should be good enough already, and that it is not necessary to add another layer of complexity to the calibration.

IV.2.2 Application to WMH detection

As in section IV.1.2, we wish to apply the presented probability calibration methods to the task of anomaly detection in the form of hyperintense lesions and other pathologies on the WMH dataset. We first use the probabilistic outputs generated for ensemble learning in the case of the patient-specific method, and in a second part calibrate score maps in the case of the localized method.

The control database used is the same as presented in section III.2.3.i, composed of 60 healthy controls, with T1 and FLAIR images. The testing database is the same as presented in section III.2.1, 60 patients from 3 different hospitals, with white matter hyperintensities and other pathologies.

IV.2.2.i Probabilistic outputs for ensemble learning

There is one interesting property about SAE+*ps*OC-SVM that we did not explore, it is that as we do subsampling of all the possible patches for a given patient (we sample n patches out of M possible localizations) we could train K one class SVM, that would have different characteristics, as they are trained on different samples. However, there is no guarantee that the obtained distance can be comparable to one from another, as the supporting hyperplane will have different support vectors.

Having introduced in the previous section probability calibration methods, we can now use the possibility of having K one class SVM, to fuse them, once calibrated into probabilities. The probability calibration is also facilitated by the higher number of patches we can sample, indeed, the lower bound on the number of points for sigmoid fitting with cSVDD is q, but if n is too small, the estimated level sets will be the same (will have the same support vectors), so a necessary condition to increase q is to have large n (as an example El Azami et al. (2017) recommend $n > 20$ when $q = 9$).

Experiments

For this study, we compare 4 methods:

- SAE+*ps*OC-SVM.

- SAE+*ps*OC-SVM, sampled $K = 5$ times (thus producing 5 anomaly score maps), converted to probability with the binning method and averaged over the 5 probability maps.
- SAE+*ps*OC-SVM with cSVDD ($q = 9$) in place of the one class SVM, which we will call SAE+*ps*cSVDD.
- SAE+*ps*cSVDD, sampled $K = 5$ times, converted to probability with the sigmoid, and averaged.

This allows us to see the benefits of the binning or cSVDD + sigmoid method for ensemble learning. These two methods are compared to the baseline of the patient-specific method and with SAE+*ps*cSVDD (that might already alter the results).

We use $K = 5$, $n = 500$ patches sampled, and the same patch-based auto-encoder (see section III.1.2) for representation learning. The database used for testing and inference-time fitting of the SVM/cSVDD is the WMH database, with 60 patients. We use $q = 9$ for the cSVDD (thus 9 level sets estimated) and $b = 20$ for the binning method (thus 41 bins). We do not evaluate calibrated score maps on their own, as the performances would be the same as the base maps since the calibration does not change the ordering of the scores[1].

Results and discussion

We present in table IV.4 the quantitative results obtained, and in figure IV.11, qualitative examples of the generated score maps, calibrated maps, and averaged calibrated map. As for the previously presented score maps, the full range of scores is not presented, it is truncated manually for the probability maps (to [0, 0.1] or [0, 0.4]), to showcase with more detail the outliers.

WMH (T1+FLAIR) 3 hospitals	SAE +*ps*OC-SVM	SAE +5 *ps*OC-SVM calibrated with binning and averaged	SAE +*ps*cSVDD	SAE +5 *ps*cSVDD calibrated with sigmoid and averaged
AU ROC	0.80±0.09	0.75±0.10	0.75±0.11	0.75±0.11
AU ROC 30	0.48±0.20	0.44±0.15	0.45±0.16	0.45±0.17
AU PR	0.084±0.099	0.071±0.078	0.081±0.083	0.081±0.083
AU PRO	0.71±0.11	0.53±0.09	0.55±0.10	0.55±0.11
AU PRO 30	0.33±0.18	0.14±0.09	0.15±0.10	0.15±0.10
$\lceil Dice \rceil$	0.14±0.13	0.13±0.10	0.14±0.11	0.14±0.11

Table IV.4: Mean (± std) metric on every patient from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.007±0.006

On the quantitative results, we see that neither the binning method, nor the cSVDD+sigmoid method seem to improve the performances, and this is valid for every metric. We observe a small drop in performances, on all metrics except *AU PR*, and a more significant drop on the *PRO* metrics. One hypothesis is that the generated score maps are not diverse enough, thus the calibration and combination of the score maps don't boost the performances. Worse: the calibration does not change the ordering of the scores but can lead to different scores being pulled in the same bin (for the binning method), thus creating 'staircase' calibration functions that might decrease the performances. We see indeed that for SAE+*ps*OC-SVM, on figure IV.11, the 5 maps do not have a great variability. We also see in this figure that after calibration the maps have several 'plateaus' due to the calibration.

[1] Recall that for the patient-specific method, there is one SVM/cSVDD for a whole patient, thus applying any monotone function (as for calibration) will not change the ordering of the scores.

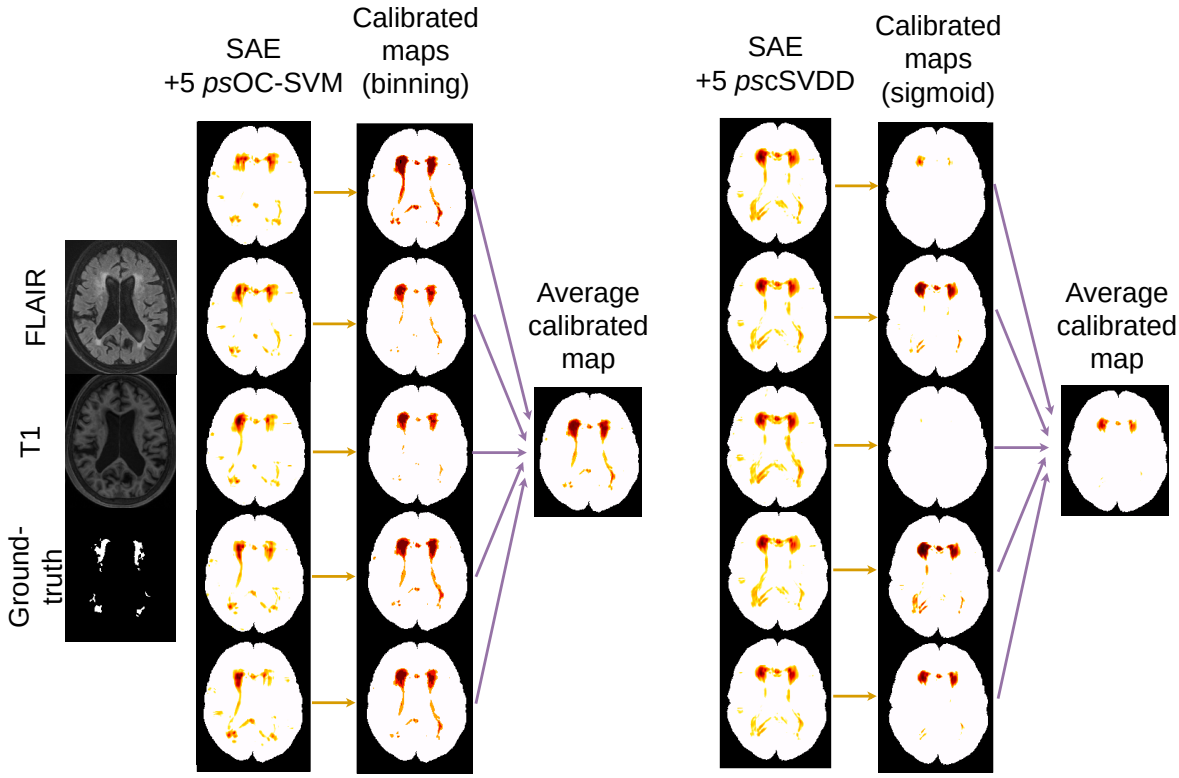

Figure IV.11: Showcase of the two calibration methods studied for a random Singapore patient (the same as in the middle row of figure IV.4, SIN63), on the left, calibration by binning, on the right, calibration with cSVDD + sigmoid. Orange arrows represent calibration, and purple arrows represent averaging.

For cSVDD, this comment is the same as the variability does not seem high enough to produce interesting averaging. Although, with the calibration with sigmoid we do not see the 'plateaus', for cSVDD though, the drop in performance seems to occur directly when detecting anomalies with the cSVDD, as there is no significant difference of performance before and after calibration. Interestingly enough, we see on the right of figure IV.11 that some calibrated maps do not have anomalies in the lower part of the image, thus when averaged, this produces a 'cleaning' effect for this part of the brain, though this does not translate into a performance gain. The drop in the *PRO* metric seems to indicate that some small lesion detections are lost when calibrating, this might be due to such a 'cleaning' effect on the small lesions.

When looking at the per-hospital results (table IV.5) for Singapore, we see that the performances remain equal when doing ensemble learning as opposed to only 1 OC-SVM/cSVDD for all metrics (even *PRO*) except *AU PR* increases. On Utrecht, all metrics improve (*ROC*, *PR* and *PRO)*. On Amsterdam, only the *PRO* decrease. This seems to indicate that on Amsterdam, the calibration and fusion seem to decrease the finding of small lesions, on Singapore this method ensures most anomalous scores are true lesions[1], and improve every metric on Utrecht, thus the finding of big and small lesions. It is not clear from the figure IV.12, an additional visual result, this time for an Amsterdam patient, that the small lesions would be lost when calibrating. However, we see again a phenomenon that we already witnessed in section IV.1.2.ii: the ventricles are detected as anomalous. As the patients studied are older than the control

[1]The 'cleaning' effect that we witness on the figure IV.11 might be responsible for the *AU PR* gain. Note that this improvement is not found when using binning.

WMH (T1+FLAIR) Amsterdam	SAE +*ps*OC-SVM	SAE +5 *ps*OC-SVM calibrated with binning and averaged	SAE +*ps*cSVDD	SAE +5 *ps*cSVDD calibrated with sigmoid and averaged
AU ROC	0.76±0.10	0.75±0.13	0.73±0.14	0.73±0.14
AU ROC 30	0.54±0.16	0.48±0.20	0.46±0.20	0.46±0.20
AU PR	0.084±0.103	0.088±0.103	0.086±0.102	0.089±0.103
AU PRO	0.65±0.06	0.49±0.08	0.46±0.09	0.46±0.09
AU PRO 30	0.35±0.09	0.12±0.07	0.11±0.07	0.11±0.07
$\lceil Dice \rceil$	0.13±0.11	0.15±0.13	0.15±0.13	0.16±0.13
WMH (T1+FLAIR) Singapore	SAE +*ps*OC-SVM	SAE +5 *ps*OC-SVM calibrated with binning and averaged	SAE +*ps*cSVDD	SAE +5 *ps*cSVDD calibrated with sigmoid and averaged
AU ROC	0.73±0.11	0.77±0.09	0.78±0.09	0.79±0.10
AU ROC 30	0.44±0.15	0.45±0.13	0.48±0.14	0.49±0.14
AU PR	0.074±0.071	0.079±0.073	0.099±0.080	0.105±0.084
AU PRO	0.54±0.07	0.52±0.05	0.56±0.05	0.57±0.05
AU PRO 30	0.17±0.07	0.10±0.06	0.12±0.06	0.12±0.07
$\lceil Dice \rceil$	0.14±0.11	0.13±0.10	0.15±0.13	0.16±0.11
WMH (T1+FLAIR) Utrecht	SAE +*ps*OC-SVM	SAE +5 *ps*OC-SVM calibrated with binning and averaged	SAE +*ps*cSVDD	SAE +5 *ps*cSVDD calibrated with sigmoid and averaged
AU ROC	0.58±0.11	0.74±0.05	0.74±0.07	0.74±0.07
AU ROC 30	0.22±0.12	0.40±0.11	0.40±0.13	0.40±0.13
AU PR	0.038±0.042	0.045±0.034	0.052±0.041	0.052±0.039
AU PRO	0.44±0.04	0.59±0.09	0.63±0.09	0.63±0.09
AU PRO 30	0.07±0.02	0.19±0.10	0.23±0.12	0.23±0.11
$\lceil Dice \rceil$ *	0.07±0.07	0.09±0.06	0.10±0.06	0.10±0.06

Table IV.5: Mean (± std) metric on every patient from the Amsterdam, Singapore, and Utrecht hospitals for each method. *AU PR* for a random classifier would be 0.003±0.004 for Amsterdam, 0.008±0.006 for Singapore, and 0.008±0.007 for Utrecht. We recall that *AU ROC 30* and *AU PRO 30* for a random classifier would be 0.15.

database, enlarged ventricles could be viewed as a deviation from the normality. However, we do not find this deviation in all the patients. An additional post-processing removing the ventricles from the anomaly maps as done in section IV.1.2.i could still lead to a performance gain.

For now, even though the calibration with cSVDD is more theoretically grounded than SVM binning, this doesn't seem to translate into performance gains, with the notable exception of *AU PR* on Singapore. Additional experiments varying the number of bins (to $b = 50$) and the number of level sets estimated (to $q = 100$) did not change the conclusions of this study, which seem to indicate robustness for these parameters. Variance among the metrics did not improve with calibration and is still pretty high concerning the mean metric, indicating that we should carefully consider the findings of these studies.

IV.2.2.ii Probalistic outputs for score map uniformization

We have seen that the probability calibration could allow us to aggregate multiple maps when using the patient-specific algorithm SAE+*ps*OC-SVM. We now present another use-case of probability calibration, based this time on the localized SVM of algorithm SAE+*loc*OC-SVM.
With this algorithm, M separate one class SVM (one per brain localization) are trained on N_{H}

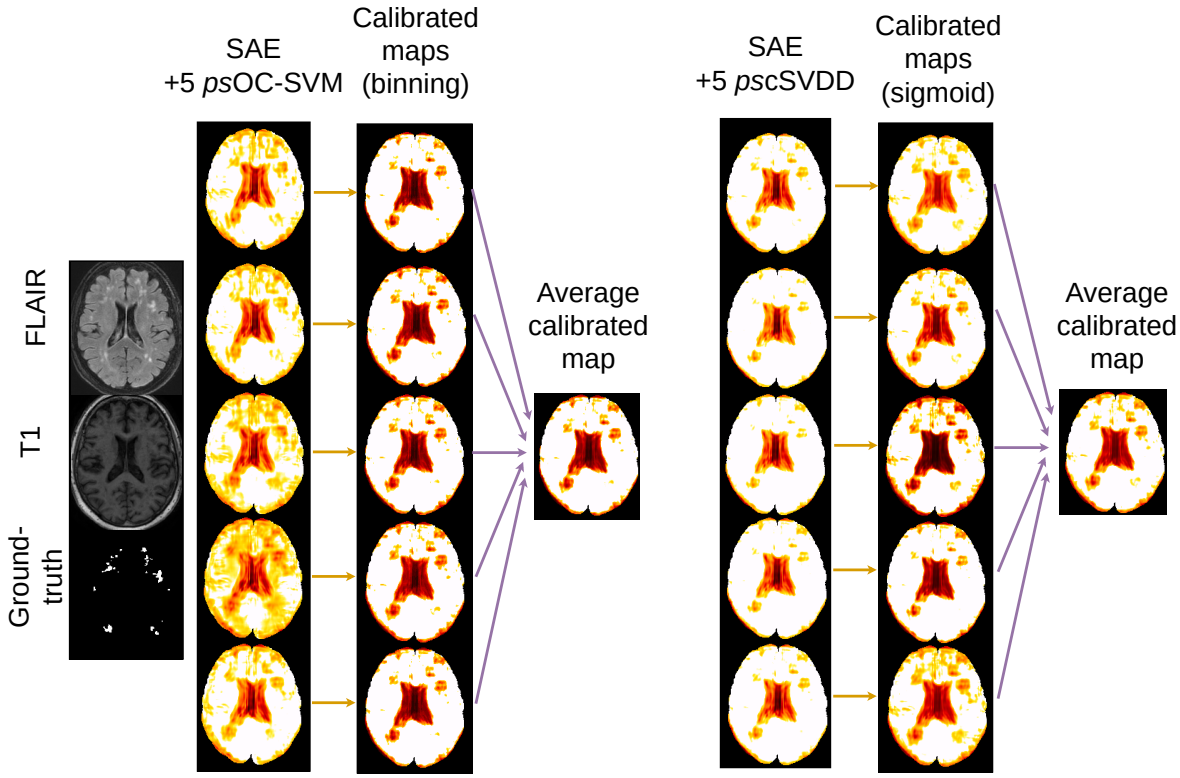

Figure IV.12: Additional showcase of the two calibration methods studied for a random Amsterdam patient (the same as in the top row of figure IV.4, AM114), on the left, calibration by binning, on the right, calibration with cSVDD + sigmoid. Orange arrows represent calibration, and purple arrows represent averaging.

healthy controls samples, as such, the obtained score map has no guarantee of being homogeneous/uniform, in the sense that from voxel to voxel, the obtained scores are not comparable because they are obtained with different support vectors, thus different hyperplanes and margin. Probability calibration, as presented in section IV.2.1, is a solution to this problem, and allows having comparable scores and thus uniform maps.

Experiments

For this study, we use the probability calibration with cSVDD, as binning did not show any superiority in the previous section (IV.2.2.i) and as the cSVDD approach is more theoretically grounded. The method used is the same as in algorithm SAE+*loc*OC-SVM, but instead of training a one-class SVM, we train a cSVDD (thus called SAE+*loc*cSVDD), and convert the anomaly scores to probabilities with the sigmoid calibration, thus obtaining a uniform score map. We thus call the 'raw' method, without calibration. We compare the calibrated (also called uniformized in this case) map to the 'raw' SAE+*loc*cSVDD map, and to SAE+*loc*SVM.

The other parameters of the experiment are the same as presented in section III.2.3.

Results and discussion

We present in table IV.6 the quantitative results obtained for this task, and in figure IV.13 examples of uniformized score maps versus raw cSVDD score maps and raw SVM score maps.

The quantitative performances of table IV.6 indicate that there is an improvement of all metrics when using cSVDD instead of SVM for anomaly detection, however, the calibration

WMH (T1+FLAIR) 3 hospitals	SAE +*loc*OC-SVM	SAE +*loc*cSVDD	SAE +*loc*cSVDD calibrated
AU ROC	0.52±0.19	0.59±0.18	0.60±0.16
AU ROC 30	0.19±0.16	0.29±0.21	0.28±0.20
AU PR	0.023±0.031	0.045±0.061	0.032±0.036
AU PRO	0.43±0.17	0.49±0.11	0.49±0.11
AU PRO 30	0.09±0.13	0.13±0.08	0.14±0.08
⌈*Dice*⌉	0.05±0.05	0.09±0.10	0.07±0.07

Table IV.6: Mean (± std) metric on every patient from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.007±0.006

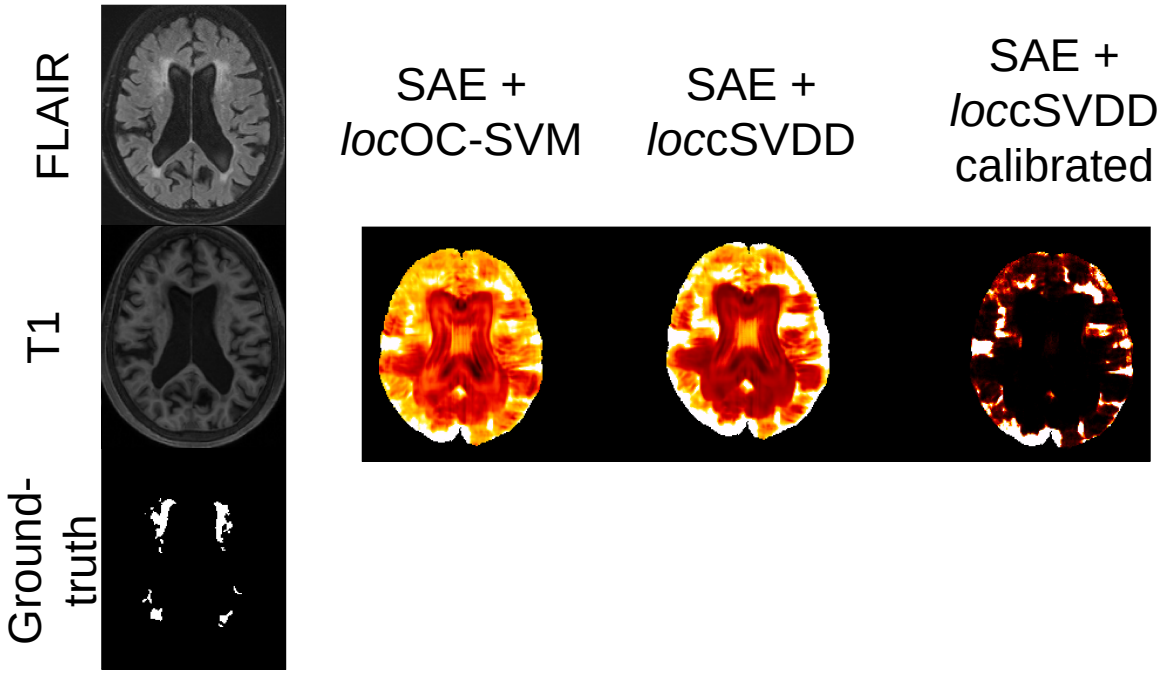

Figure IV.13: Showcase of the uniformized score map obtained with probability calibration, for a random Singapore patient (the same as in the middle row of figure IV.4, SIN63).

using sigmoid leads to a drop of *AU PR*. Per-hospital results shown in table IV.7 show no similar performances for all metrics on Singapore, with SVM, cSVDD, or calibrated cSVDD, a slight improvement on all metrics for Utrecht, and improvement of all metrics when using cSVDD instead of SVM, with a drop in *AU PR* for calibrated cSVDD. Figure IV.13 shows a qualitative example of such a procedure, it seems in this example that the fitted sigmoids have a too steep slope, as the obtained scores are often extreme (very normal or very abnormal). As the calibration is done on every voxel, this is done only on $N_{\mathrm{H}} = 60$ points, which could be not sufficient to obtain a robust calibration.

At this point, it is not clear with quantitative and qualitative results, if such a calibration provides more lesion detections. Despite some improvements in the quantitative metrics, this improvement is not shared for every hospital and is not seen in the qualitative example. Moreover, when the metrics show improvement (take Utrecht for instance), the improvement seems to be when changing from SVM to cSVDD and not from cSVDD to calibrated cSVDD. This could suggest that the cSVDD procedure, by estimating jointly q nested level sets, already enforces a

calibration of some kind[1].

WMH (T1+FLAIR) Amsterdam	SAE +*loc*OC-SVM	SAE +*loc*cSVDD	SAE +*loc*cSVDD calibrated
AU ROC	0.62±0.15	0.76±0.12	0.76±0.12
AU ROC 30	0.25±0.16	0.52±0.18	0.50±0.17
AU PR	0.015±0.018	0.081±0.089	0.048±0.051
AU PRO	0.41±0.16	0.51±0.08	0.52±0.09
AU PRO 30	0.17±0.09	0.18±0.09	0.19±0.09
$\lceil Dice \rceil$	0.03±0.04	0.16±0.13	0.10±0.09
WMH (T1+FLAIR) Singapore	SAE +*loc*OC-SVM	SAE +*loc*cSVDD	SAE +*loc*cSVDD calibrated
AU ROC	0.51±0.20	0.48±0.14	0.50±0.13
AU ROC 30	0.19±0.16	0.18±0.12	0.18±0.10
AU PR	0.034±0.045	0.030±0.106	0.027±0.026
AU PRO	0.47±0.20	0.39±0.11	0.41±0.11
AU PRO 30	0.12±0.17	0.07±0.05	0.09±0.05
$\lceil Dice \rceil$	0.06±0.07	0.07±0.05	0.06±0.05
WMH (T1+FLAIR) Utrecht	SAE +*loc*OC-SVM	SAE +*loc*cSVDD	SAE +*loc*cSVDD calibrated
AU ROC	0.45±0.16	0.54±0.12	0.53±0.09
AU ROC 30	0.12±0.11	0.18±0.10	0.16±0.06
AU PR	0.020±0.017	0.023±0.020	0.022±0.018
AU PRO	0.41±0.14	0.55±0.08	0.53±0.06
AU PRO 30	0.06±0.06	0.16±0.06	0.15±0.05
$\lceil Dice \rceil$	0.05±0.04	0.05±0.04	0.05±0.04

Table IV.7: Mean (± std) metric on every patient from the Amsterdam, Singapore, and Utrecht hospitals for each method. *AU PR* for a random classifier would be 0.003±0.004 for Amsterdam, 0.008±0.006 for Singapore, and 0.008±0.007 for Utrecht. We recall that *AU ROC 30* and *AU PRO 30* for a random classifier would be 0.15.

As the computation of around 1.5 million cSVDD (one per brain localization) is computationally heavy, the hyperparameter $q = 12$ was tightly chosen as a compromise between reasonable computational time (high q) and numerical solvability (low q: with low n yield unstable support vectors). A higher q, and most importantly higher $n = N_{\text{H}}$ (number of controls) would ensure a better calibration, and maybe more solid conclusions about this score map uniformization method.

IV.2.3 Conclusion and perspectives

We have seen in this section two methods from the literature that allow calibration of anomaly score maps derived from support estimation methods, one that is straightforward based on the binning of the SVM scores, and the other, which requires changing the optimization problem but offers theoretical guarantees on the calibration. We have applied these probability calibration methods to the problem of anomaly detection in neuroimaging, which to the best of our knowledge was never done before. We proposed to use the calibration for two different tasks: fusion

[1] As each level is a consistent estimator of the level-set of probability mass $1 - \nu$, take the extreme case where the number of samples n and the number of estimated level sets q would tend to infinity: then for each level-set j associated with ν_j, we would be sure that there are $1 - \nu_j$ samples inside the j^{th} support. As there is an infinite number of level sets and as they are all nested, this ensures perfect calibration, but up to a multiplicative factor, which is not guaranteed to be the same across each voxel.

of multiple score maps and uniformization of score maps, which we argue improves the robustness of the proposed methods. These two tasks were each downstream task of two algorithms presented before (algorithm SAE+*loc*OC-SVM and SAE+*ps*OC-SVM).

For the fusion of score maps, it seems for now that the produced variability of score maps is not high enough to make this method interesting to boost the performances of lesion detection on the WMH dataset. Fusion of such score maps could be more useful in the context of multi-modality, where one model per modality (T1, FLAIR, etc.) would be trained (instead of early fusion by channel concatenation), and score maps from different specialized models would be fused.

For the score map uniformization, it seems that fitting a cSVDD instead of a one-class SVM, per voxel, seems to improve the performances, even though it is not clear from visual results. The calibration, per se, does not seem to improve the detection of lesions, which could suggest a 'calibration effect' by the cSVDD. Experiments with more healthy controls have to be carried on to strengthen the conclusions of this study.

A straightforward perspective of this section would be to do uniformization and fusion, to combine the two presented methods. By fitting K cSVDD per voxel, each on a subset of the N_{H} controls, we could have uniform and aggregated maps. One weakness of such an approach is that we would need even more healthy controls to subdivide the already small training set. Another perspective would be, instead of q evenly spaced level sets, to work on a non-uniform distribution of these level sets, to study the best configuration possible. We hypothesize that the best configuration could be one where smaller ν_j are estimated, as they provide an estimation of the 'true' support, which is the object of interest at the end. Said differently, we are more interested in accurate probability estimates of samples of low probability (outliers) than samples of high probability (inliers).

Related to that matter, there is one essential question that we need to address before closing this section: if interested in probability calibration of support estimation methods, why not directly estimate the probabilities with density estimation methods? First of all, as said hereabove, we are more interested in accurate probabilities of outliers than inliers, as they are the object of interest. Because in the end, even after the probability calibration, we are interested in saying if a sample is an outlier or normal (thus thresholding this probability estimate). Secondly, the studied methods (SVM/cSVDD) are robust to outliers, even with a relatively low number of samples, which is a property difficult to achieve with a low number of samples for density estimation.

IV.3 Conclusion

In this second contribution chapter, we proposed a new learning framework for support estimation methods, that we applied to hyperintense lesion detection and Parkinson's disease classification. We found above-state-of-the-art performances on the WMH challenge dataset and outperformed supervised methods for Parkinson's classification.

This proposed framework allowed for relaxing the constraint on the number of training controls, which allowed us to use calibration methods to fuse different score maps derived from different subsamples of patches. At this stage, the different models lack enough variability to make this fusion interesting in terms of performance. The studied calibration method was also used to uniformize per-voxel anomaly score maps, which seem to improve the detection of lesions on this dataset, although this result should be treated with caution.

The proposed patient-specific framework allowed us to relax the constraint on the number of training controls, but also supposedly on the quality of the registration, as for the outlier

detection step, there is no need to have one model per voxel. A natural extension of the proposed experiments would be to train a patch-based auto-encoder (non-siamese, which needs pairing between registered subjects) on non-registered subjects, and then use the SAE+*ps*OC-SVM. This would allow to have no registration step in the full pipeline. This is beneficial as the registration is a problem and a literature niche on its own, and it is especially difficult to have accurate non-linear registration when presented with a wide variety of modalities, MR vendors, patient age, etc. We also found that our models in general tend to generate anomalies where the registration errors are supposed to be the highest (near the cortex border), which would support the need for registration-free models for medical imaging.

In this chapter, we also found, throughout the extensive validation on three different hospitals, that the results, both qualitative and quantitative, could vary significantly from one center to another. This strengthens the need for domain adaptation techniques, especially in medical imaging, where the inter-patient, inter-center variability is so high. We also found that sometimes the quantitative results did not match the qualitative ones. We wish, in the following chapter to delve into more in-depth analysis, as we think that even by looking at 6 different quantitative metrics, some blind spots can still appear.

V | Structured latent space for anomaly detection

We have extended the evaluation of the model on three public databases in chapter III and proposed methodological contributions in chapter IV to improve the outlier detection step. We now wish to propose methods that improve the representation learning step, notably by coupling the representation learning and the outlier detection step into a unified framework.

We first propose, in section V.1, to take a step back and look at some evaluation flaws we need to address when evaluating hyperintense lesion detections and motivate the need for evaluation on a more challenging task. We then explore existing methods that structure the latent space of the auto-encoder in section V.2, to try to improve the performances of the proposed methods. We then propose a novel framework for coupling auto-encoder and one class SVM in section V.3, that allows end-to-end learning, and thus latent representation adapted to the downstream task of support estimation. In section V.4, we conclude this chapter by delving into a more in-depth analysis of the obtained anomaly score maps and latent spaces, to strengthen our findings.

V.1 Rationale for a more structured latent space

V.1.1 Hyperintensity detectors on FLAIR MRI

As already discussed in section II.2.1, one of the identified blindspots of the current literature in anomaly detection for neuroimaging is that most of the methods are evaluated on hyperintense lesions. For instance Meissen et al. (2021b) and Meissen et al. (2021a) have proved, by numerous experiments, that current state-of-the-art methods were *hyper-intensity detectors*, that were not taking into account the texture of the anomalies. They explicitly compare against Baur et al. (2021b) and Pinaya et al. (2022b) (as we did in chapter III and IV) and show that simply thresholding FLAIR images, after histogram equalization, gives better detection performances than these methods. As the performances were evaluated partially on hyperintense detection in chapter III and IV, we wish to repeat this experiment, on the WMH challenge dataset. The anomaly map considered is simply minus the FLAIR map so that most hyperintense voxels have the most negative scores. Evaluating performances based on the different metrics (e.g. $AU\ ROC$) then amounts to thresholding this map (we did not utilize histogram equalization). Figure V.1 showcases an example of this process, where we clearly see that by thresholding the FLAIR map, we obtain a binary map that is very close to the ground truth, although not perfect.

Evaluating performance based on the different metrics (eg AUC..) amounts to.

We recall (section III.2.1) that the WMH challenge dataset has two types of reported lesions: white matter hyperintensities and other pathologies. In III and IV, we pulled together these two types of lesions as anomalies to detect. To identify clearly the effect of hyperintensity detection, we separate the evaluation for hyperintensity (presented in table V.1) and other pathologies (presented in table V.2).

By looking at the quantitative performances obtained when looking at hyperintensities, and the qualitative example, it is clear that only thresholding the FLAIR map suffices and that this method outperforms our proposed method (SAE+psOC-SVM), SAE+locOC-SVM and the methods from Pinaya et al. (2022b) and Baur et al. (2021b) as demonstrated by Meissen et al.

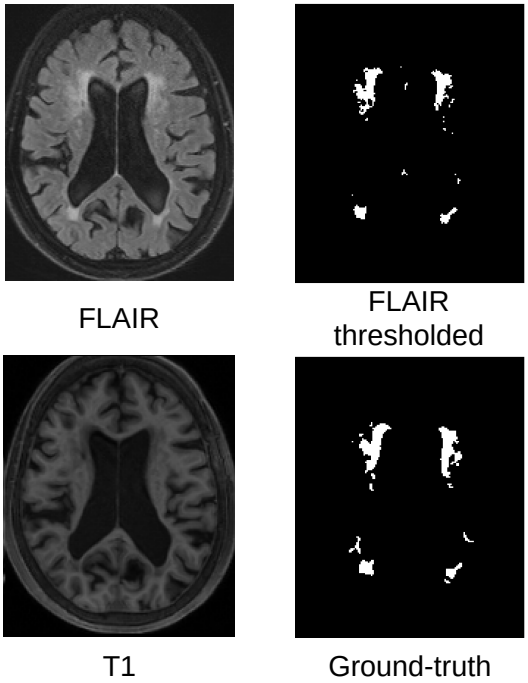

Figure V.1: Showcase of the process of thresholding the FLAIR map to obtain a binary anomaly map. Patient SIN63, same patient and slice as middle row of figure IV.4.

WMH (T1+FLAIR) 3 hospitals **hyperintensities**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	Siamese AE +*ps*OC-SVM	- FLAIR thresholding
AU ROC	0.69±0.14	0.54±0.10	0.55±0.20	0.81±0.09	**0.96**±0.03
AU ROC 30	0.40±0.20	0.21±0.12	0.25±0.22	0.54±0.17	**0.90**±0.08
AU PR	0.061±0.078	0.027±0.030	0.028±0.035	0.091±0.081	**0.494**±0.276
AU PRO	0.56±0.10	0.50±0.08	0.42±0.12	0.64±0.12	**0.92**±0.04
AU PRO 30	0.21±0.13	0.14±0.06	0.11±0.08	0.26±0.14	**0.77**±0.10
$\lceil Dice \rceil$	0.11±0.10	0.06±0.05	0.07±0.07	0.16±0.10	**0.53**±0.22

Table V.1: Mean metric on every patient from the 3 different hospitals for each method (evaluation on hyperintensities only). *AU PR* for a random classifier would be 0.006±0.006. In bold is shown the best model.

(2021b). Nonetheless, Meissen et al. (2021a) argued that these hyperintensity detections were caused by the use of reconstruction error methods, whereas we show here that the support estimation method also suffers from this drawback.

We believe that this task (WMH detection on FLAIR) is still relevant, as it allows comparison of the models and can establish the best-performing model. However, these models turn out to be of no use in clinical practice, as they are outperformed by a simple procedure.

This finding, however, is only true when provided with the FLAIR images, and when evaluating on hyperintensities. As we see in table V.2, the thresholding method is not able to obtain such performances when looking at other pathologies than hyperintensities. Histogram presented in figure V.2 summarizes these results: on FLAIR, voxel intensities of the other pathological lesions strongly overlap with the healthy voxels ones, meaning that these other pathological lesions are not characterized by hyperintense FLAIR signal, whereas *hyperintensities*, by definition, are more intense than the healthy voxels.

Also note that for hyperintensities detections, the FLAIR images being much more informa-

WMH (T1+FLAIR) 3 hospitals **other pathologies**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	Siamese AE +*ps*OC-SVM	- FLAIR thresholding
AU ROC	0.65±0.19	0.56±0.09	0.63±0.20	0.62±0.14	0.64±0.13
AU ROC 30	0.32±0.22	0.21±0.11	0.32±0.26	0.22±0.21	0.39±0.17
AU PR	0.014±0.026	0.005±0.007	0.010±0.014	0.009±0.016	0.053±0.11
AU PRO	0.23±0.32	0.21±0.28	0.23±0.32	0.13±0.25	0.24±0.32
AU PRO 30	0.11±0.19	0.08±0.13	0.11±0.21	0.05±0.12	0.15±0.22
$\lceil Dice \rceil$	0.03±0.05	0.01±0.02	0.02±0.03	0.02±0.04	0.10±0.14

Table V.2: Mean metric on every patient from the 3 different hospitals for each method (evaluation on other pathologies only). *AU PR* for a random classifier would be 0.001±0.002.

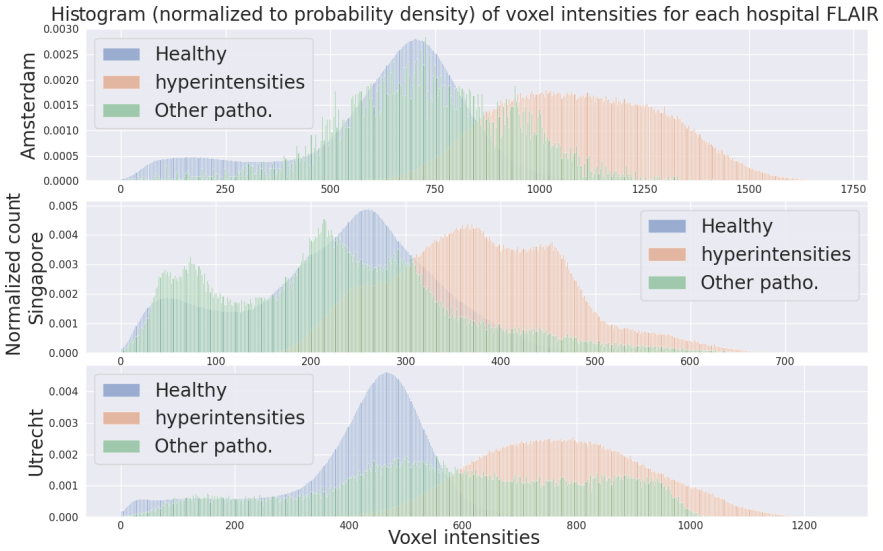

Figure V.2: Histogram of the different voxel classes (healthy, hyperintensities/WMH, other pathologies) on the FLAIR images, for each hospital of the WMH dataset.

tive than the T1 images, we believe that concatenating T1 and FLAIR as input could generate more confusion and decrease the performances than using FLAIR images only. Also, by comparing performance reported by the different metrics for the combined detection of hyperintense and other pathological lesions (table IV.1) and for the detection of hyperintense lesions only (table V.1), we conclude we conclude that the hyperintensities largely dominate the evaluation, as the performances are very similar. This can be easily explained by the fact that hyperintensities represent 98% of the number of total lesions and 90% of the total lesional volume.

We believe that these findings justify proposing a more challenging task to compare the performances of the different anomaly detection models, where trivial methods would not outperform the state-of-the-art models.

V.1.2 Poor sensitivity on T1 MRI

We proved in the previous section that hyperintense lesion detection on FLAIR was too trivial to correctly evaluate models that would translate to clinical practice and, as such, we wish to move to a more challenging task.

We now propose to investigate the task of detecting the so-called *hyperintense* (on FLAIR) and other pathological lesions on T1 images only. As we can see in figure V.1 and figure V.3, this is much more challenging for the 'hyperintensities' using T1 images, as they indeed appear to

have a similar intensity profile as the healthy voxels, and thus would have to be detected because of texture, or context to the surrounding voxels, making them subtle lesions. The goal of this chapter, in terms of performances, would be two-fold: to improve the detection performances of the other pathologies, and to detect the 'hyperintensities'[1], both using T1 images only.

Table V.3 sets the performance baseline for the detection of 'hyperintensities' and other pathologies in T1+FLAIR images, for the same models as presented in the previous chapter (see table IV.1). Performance of the 'thresholding method' is also reported, this time, without inversion, as the 'hyperintensities' appear slightly hypointense on T1 images. Figure V.4 showcases anomaly maps obtained with the T1 input only.

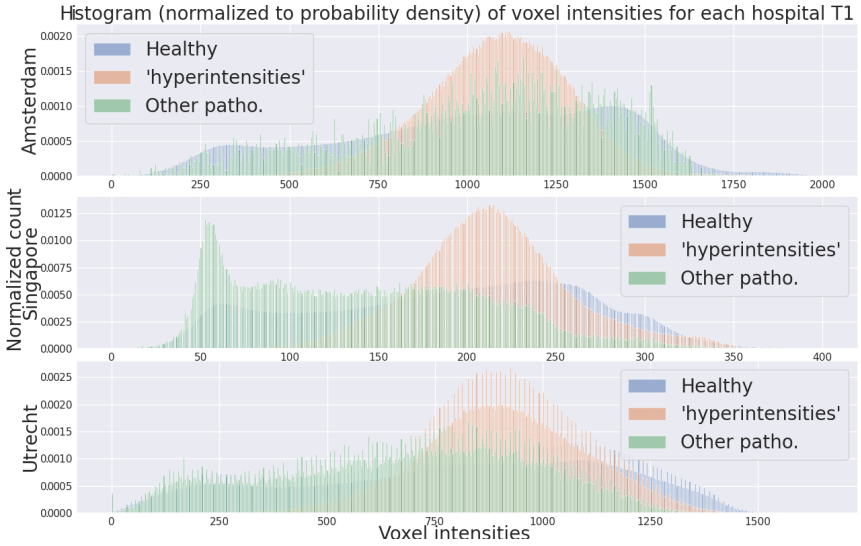

Figure V.3: Histogram of the different voxel classes (healthy, 'hyperintensities'/WMH, other pathologies) on the T1 images, for each hospital of the WMH dataset. We observe that contrary to the histogram presented in figure V.2, the 'hyperintensities' are now blended with the healthy voxels.

[1]We call these 'hyperintensities' on T1: subtle lesions or WMH or 'hyperintensities', using quotation marks.

WMH (T1 only) 3 hospitals **'hyperintensities'**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	Siamese AE +*ps*OC-SVM	T1 tresholding
AU ROC	0.57±0.09	0.48±0.04	0.41±0.16	0.53±0.13	0.41±0.05
AU ROC 30	0.22±0.09	0.10±0.03	0.13±0.13	0.17±0.11	0.01±0.01
AU PR	0.022±0.023	0.013±0.013	0.017±0.018	0.015±0.012	0.012±0.012
AU PRO	0.47±0.05	0.50±0.06	0.51±0.20	0.52±0.15	0.31±0.05
AU PRO 30	0.10±0.04	0.11±0.05	0.18±0.17	0.13±0.10	0.00±0.01
⌈*Dice*⌉	0.05±0.05	0.03±0.03	0.04±0.04	0.04±0.03	0.03±0.03
WMH (T1 only) 3 hospitals **other pathologies**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	Siamese AE +*ps*OC-SVM	T1 tresholding
AU ROC	0.71±0.10	0.50±0.05	0.60±0.15	0.50±0.15	0.55±0.11
AU ROC 30	0.40±0.18	0.14±0.05	0.26±0.22	0.17±0.17	0.14±0.10
AU PR	0.018±0.031	0.004±0.007	0.017±0.025	0.005±0.008	0.006±0.010
AU PRO	0.25±0.34	0.18±0.24	0.22±0.30	0.08±0.17	0.20±0.27
AU PRO 30	0.14±0.21	0.05±0.07	0.09±0.17	0.02±0.07	0.05±0.08
⌈*Dice*⌉	0.04±0.06	0.01±0.02	0.04±0.06	0.01±0.02	0.01±0.02

Table V.3: Mean metric on every patient from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.006±0.006 for 'hyperintensities' and 0.001±0.002 for other pathologies.

WMH (T1 only) 3 hospitals **'hyperintensities'**	Atlas difference after histogram matching	WMH (T1 only) 3 hospitals **other pathologies**	Atlas difference after histogram matching
AU ROC	0.62±0.06	*AU ROC*	0.62±0.13
AU ROC 30	0.15±0.08	*AU ROC 30*	0.27±0.16
AU PR	0.020±0.021	*AU PR*	0.010±0.015
AU PRO	0.52±0.08	*AU PRO*	0.22±0.30
AU PRO 30	0.06±0.04	*AU PRO 30*	0.10±0.16
⌈ *Dice* ⌉	0.05±0.05	⌈ *Dice* ⌉	0.03±0.03

Table V.4: Performances of another simple baseline, obtained by voxel-wise mean squared error with the MNI152 T1 Atlas, after histogram matching.

We see in the table V.3, for the performances of the 'hyperintensities' detection on T1, a large drop of performances, for all methods and all metrics, compared to the hyperintensities detection on T1+FLAIR reported in table V.1, indicating the difficulty of this task. Still on table V.3, for the other pathologies, we also observed a drop in performances (compared to V.2 for T1+FLAIR) but that is much smaller. Qualitatively, in figure V.4, we see almost no correct detections. Anomaly score maps are either saturated with false positive (SAE+*loc*OC-SVM) or false negative (SAE+*ps*OC-SVM detections and Baur et al. 2021b).

We present in table V.4 another baseline, that outperforms the thresholding but is still very simple to implement. For each patient image, its histogram intensity is matched to a reference T1 Atlas (ICBM 2009a Nonlinear Symmetric Fonov et al. 2009), and then the mean squared error between the matched image and the template is used as an anomaly score. The best method from V.4 barely surpasses this simple baseline.

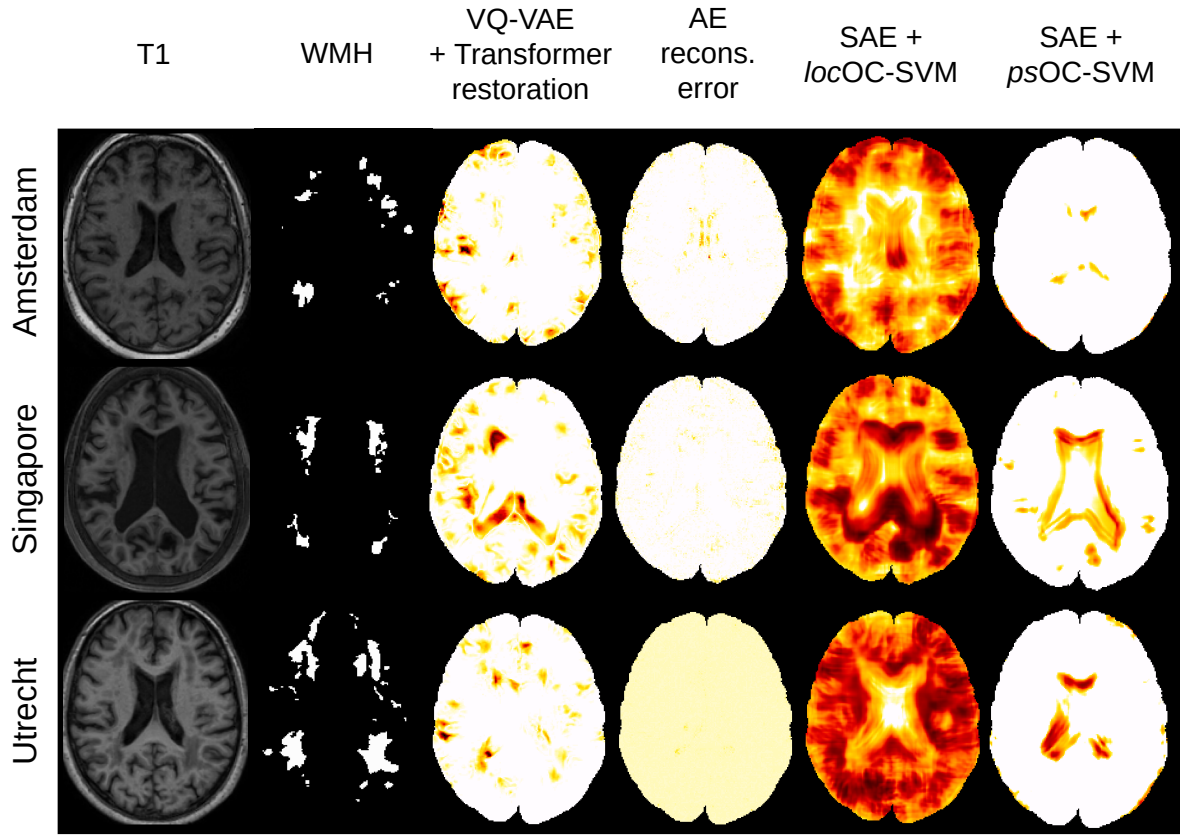

Figure V.4: Showcase of anomaly maps obtained with the models on T1 images only. Selected slices and patients are the same as in figure IV.4 (AM114, SIN63, and UT37). Baur et al. (2021b), Pinaya et al. (2022b) and ours, redder means anomaly score higher. Only T1 is used as an input for the models but only FLAIR is shown.

V.2 Improved representations

Now that we have set up a baseline for the studied models on the T1, we propose to improve the structure of the latent representation space, with the goal of improving the sensitivity or specificity of the proposed method. Section V.2.1 and V.2.2 propose two simple modifications to structure the latent space: variational regularization and positional encoding. We study in section V.2.3 the application to subtle lesion detection and compare it to the previously set-up baseline (section V.1.2).

V.2.1 Measurable latent spaces

In chapter III and IV, we used 'simple' auto-encoders for the representation learning step, with no constraint on the mapping the encoder has to make in the latent space. We introduced in paragraph I.3.1.i:Variational auto-encoders a commonly used tweak for auto-encoder: Variational auto-encoders, see figure I.13. Recall that VAE, instead of deterministically mapping the patch $\mathbf{x}$ to a fixed $\mathbf{z}$, maps a patch $\mathbf{x}$ to a mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$. Random sampling then occurs to obtain $\mathbf{z}$ from the multivariate normal distribution obtained with $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, i.e. $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}))$.

The distribution of all latent representations $\mathbf{z}$ is pushed toward the standard normal distribution, through a KL divergence term to $\mathcal{N}(\mathbf{0}, diag(\mathbf{1}))$. This has the notable effect of producing

a latent space that is denser, and thus continuous, as all 'sampling balls' $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ are brought close to the center of the space, and samples are decoded each time with a dose of randomness: this ensures that points that are close in the latent space will have similar decoding (the decoder, as the encoder, is deterministic). Figure V.5 represents this concept.

We believe this property of continuity of the latent space will enforce that this space would be more 'measurable', in the sense that without this regularization, two points that are close in the latent space have no guarantee of being decoded into a similar patch, and thus the measured distance would not be representative of patch differences. As one class SVM works with the RBF kernel, that measures distance with the L2 norm, forcing the 'measurability' of the latent space could improve the ability of the SVM to draw the classification boundary.

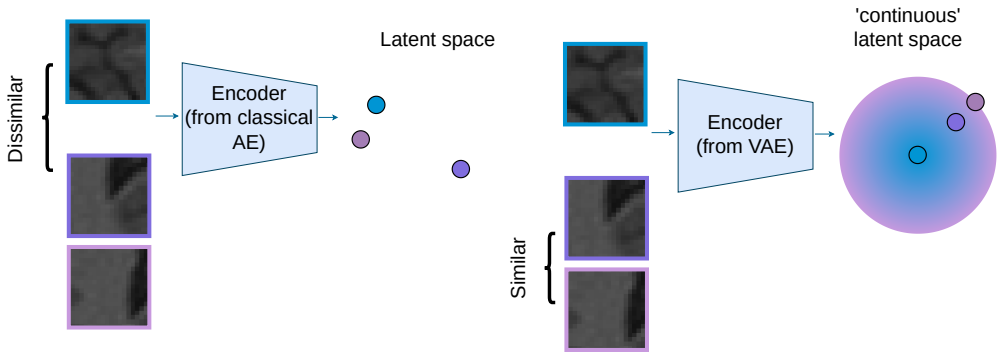

Figure V.5: Diagram of the latent space of a 'classical' auto-encoder (top) and variational auto-encoder (bottom). With the classical AE, there is not guarantee that patches that are projected to close localizations are similar in the image space, whereas this is enforced with the VAE.

These intuitions must be tempered in two ways. First, the variational regularization, because of the sampling process that introduces stochasticity for the latent representation, will make each patch less precisely localized in the latent space. This could make the estimation of the classification boundary more difficult (as evidence, it is well-known that the stochasticity of the latent representation produces blurred decoded images.). Secondly, the auto-encoder previously described in this manuscript is a siamese auto-encoder (SAE), where patches from the same localization are brought closer in the latent space. This additional constraint also structures the latent space, as it constrains close latent representations to share similarly looking patches. While not making the space more 'measurable', this still structures the latent space.

V.2.2 Localization aware latent spaces

As we train the auto-encoder by sampling patches from different localization in the brain, one other way would be to encode the coordinates of each patch in the latent space. As a result, an outlier patch that resembles one healthy patch, but is located in the wrong area of the brain could be detected as an outlier. Encoding the coordinate, more generally, gives more information about the patch and thus we believe this could increase either sensitivity or specificity if this additional information is not too complicated to handle with the existing information contained in the patch's voxels.

To implement this positional encoding, we chose the approach considered by Liu et al. (2018), which incorporates the localization as additional channels for the encoder. As we work with 3-dimensional MRI volumes, 3 channels are added, and the intensity of each voxel, for each channel, will represent the position of said voxel along this dimension, normalized to [0, 1]. For

MRI volume of size (max_x, max_y, max_z), a voxel of position (x, y, z) will have $\frac{x}{max_x}$, $\frac{y}{max_y}$, and $\frac{z}{max_z}$ as coordinate channel inputs. Figure V.6 summarizes this concept.

In Liu et al. (2018), they use the encoder for classification only (no decoder). To adapt this framework to our task, we force the decoder to reconstruct these additional channels (as the MRI channels). These additional channels that are encoded in the latent representation $\mathbf{z}$, must be thus retrieved after decoding, ensuring that the positional information is present in the latent representation.

Again, this positional encoding, though different from the siamese constraint, shares some similarities with the siamese auto-encoder: patches that are at the same position are enforced to be close in the latent space.

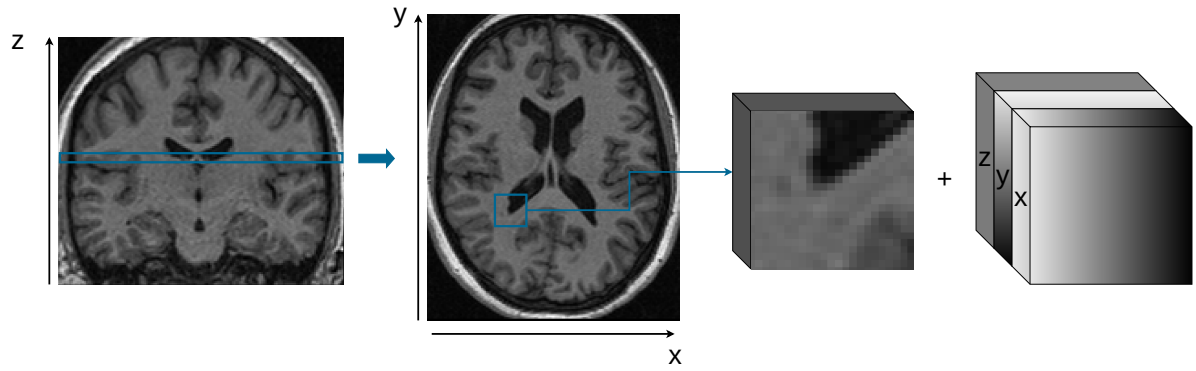

Figure V.6: Positional encoding implemented in this section (proposed by Liu et al. (2018)). Channels are added, each representing a voxel coordinate as intensity.

V.2.3 Application to subtle lesions detection

We investigate the performances of the variational regularization and the positional encoding on the task of detecting the subtle lesions ('hyperintensities') and other pathologies from the T1 images only.

V.2.3.i Experiments

The control database is again the CERMEP database (presented section III.2.3.i), with only T1 images used this time. Evaluation is performed separately for 'hyperintensities' and other pathologies. For the models, we investigate every combination with and without variational regularization (AE/VAE), with and without positional encoding (∅/Posi. enc.). We call the SAE with variational constraint VSAE. We investigate these architecture changes for SAE+*loc*OC-SVM and SAE+*ps*OC-SVM. The KL divergence weighting is put to 1 (the same weight as the reconstruction error). Other hyperparameters are the same as described in III.2.3. As the CSF segmentation did not allow a significant performance gain (see section IV.1.2.ii), we did not reproduce this post-processing step.

V.2.3.ii Results and discussion

Quantitative results are presented figure V.5, as well as qualitative results in figures V.9, V.7, V.8 and V.10.

We present in table V.5 the quantitative performances of these combinations of models, for 'hyperintentisites' and other pathologies. Figure V.7, V.8, V.9 and V.10 present qualitative

examples.

3 hospitals **'hyperintensities'**	Siamese AE							
	+*loc*OC-SVM				+*ps*OC-SVM			
	AE		VAE		AE		VAE	
	∅	Posi. enc.	∅	Posi. enc.	∅	Posi. enc.	∅	Posi. enc.
AU ROC	0.41±0.16	0.42±0.14	0.38±0.12	0.60±0.07	0.53±0.13	0.53±0.11	0.45±0.13	0.44±0.08
AU ROC 30	0.13±0.13	0.13±0.12	0.10±0.12	0.18±0.08	0.17±0.11	0.13±0.10	0.09±0.08	0.09±0.05
AU PR	0.017±0.018	0.014±0.014	0.016±0.019	0.019±0.018	0.015±0.012	0.016±0.015	0.012±0.012	0.013±0.013
AU PRO	0.51±0.20	0.38±0.12	0.35±0.13	0.58±0.06	0.52±0.15	0.56±0.13	0.49±0.14	0.43±0.06
AU PRO 30	0.18±0.17	0.07±0.05	0.06±0.08	0.17±0.06	0.13±0.10	0.17±0.12	0.10±0.11	0.09±0.04
⌈*Dice*⌉	0.04±0.04	0.04±0.03	0.04±0.05	0.04±0.04	0.04±0.03	0.04±0.03	0.03±0.03	0.03±0.03
3 hospitals **other pathologies**	Siamese AE							
	+*loc*OC-SVM				+*ps*OC-SVM			
	AE		VAE		AE		VAE	
	∅	Posi. enc.	∅	Posi. enc.	∅	Posi. enc.	∅	Posi. enc.
AU ROC	0.60±0.15	0.59±0.14	0.61±0.18	0.71±0.14	0.50±0.15	0.50±0.18	0.45±0.14	0.44±0.23
AU ROC 30	0.26±0.22	0.27±0.20	0.35±0.22	0.37±0.19	0.17±0.17	0.12±0.12	0.12±0.13	0.13±0.20
AU PR	0.017±0.025	0.017±0.029	0.027±0.045	0.009±0.011	0.005±0.008	0.004±0.006	0.005±0.008	0.005±0.008
AU PRO	0.22±0.30	0.21±0.29	0.22±0.31	0.26±0.35	0.08±0.17	0.09±0.19	0.16±0.22	0.16±0.25
AU PRO 30	0.09±0.17	0.09±0.17	0.12±0.21	0.14±0.21	0.02±0.07	0.02±0.05	0.04±0.09	0.05±0.14
⌈*Dice*⌉	0.04±0.06	0.04±0.06	0.07±0.07	0.02±0.03	0.01±0.02	0.01±0.02	0.01±0.02	0.01±0.02

Table V.5: Metrics average over every patient from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.006±0.006 for 'hyperintensities' and 0.001±0.002 for other pathologies.

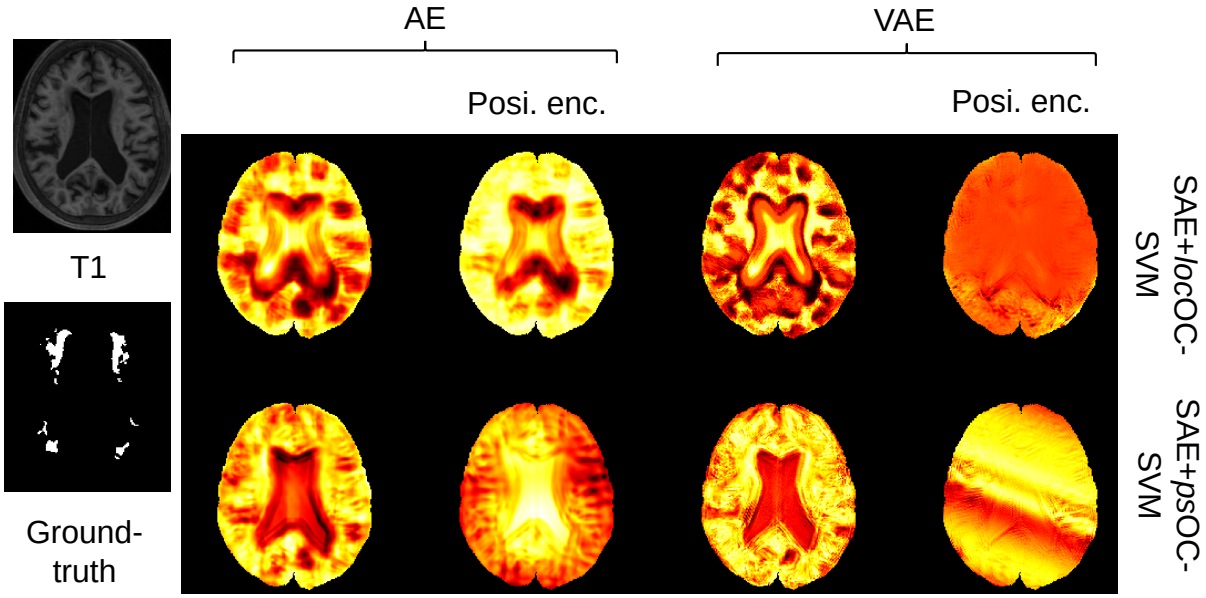

Figure V.7: Showcase of the methods presented in section V.2.3.i for patient SIN63 (same patient as middle row of figure IV.4).

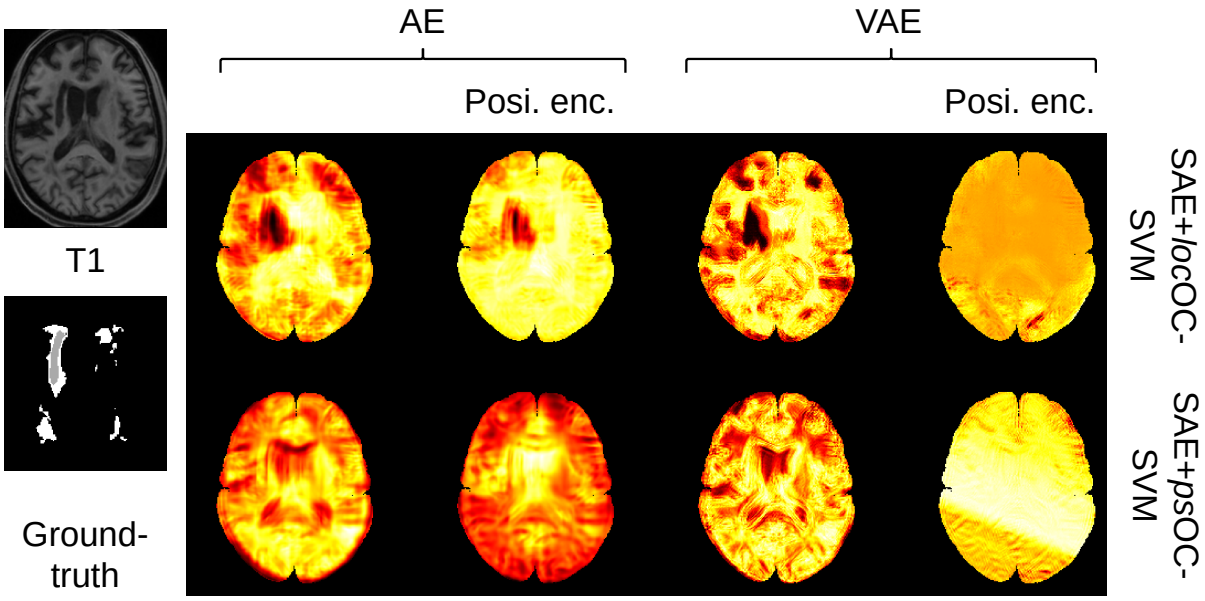

Figure V.8: Showcase of methods presented in section V.2.3.i for patient SIN67.

We see in table V.5, without going into too much detail about every metric, that most combinations do not exceed baseline performance reported in table V.4. Two notable exceptions are the configurations VSAE +*loc*OCSVM and particularly VSAE + positional encoding +*loc*OCSVM. This last method achieves better *AU PRO* and *AU PRO 30* than Pinaya et al. (2022b) on 'hyperintensities' but worse *AU PR* on other pathologies; other metrics reported are approximately similar.

When dealing with performances that are quite low, we believe it is particularly informative to look at the qualitative results. On figure V.7, we see that methods with VAE and positional encoding give degenerate anomaly score maps. We believe that the addition of the siamese, variational, and positional encoding constraints is just too much for the encoder to handle. We see the impact of positional encoding in the intensity of the score maps (score varying along x and y, generating diagonal stripes). The same result is found in the other qualitative examples in figures V.7, V.8 and V.10.

Still, in figure V.7, we see that the *patient specific* models fail to find any of the true anomalies and produce score maps that are very homogeneous, and thus uninformative. We believe that figure V.8 particularly illustrates this point: the patient, along with subtle white matter lesions, has a large lacuna beside the left ventricle. *Patient-specific* models completely fail to localize this anomaly, whereas we see that the *localized* models detect this lacuna almost perfectly.

Figure V.9 and figure V.10 confirm these findings: the *patient-specific* models fail to provide any informative anomaly score maps, whereas *localized* models, although producing a lot of false positive detections, manage to find some subtle lesions (i.e. 'hyperintensities') and to find the other pathologies (the same finding echoed in the table V.5, where *localized* models achieved higher performances for the other pathologies than for subtle lesions). All the results are shown here without any kind of post-processing or thresholding. Adding such processing would make the anomaly score maps more relevant, surely improve performance, and make it easier to read for a clinician. However, the goal here was to compare the 'raw capabilities' of the models, to assess their potential, by looking at the raw score maps.

The careful reader would have noticed that combining a positional encoding with a *localized*

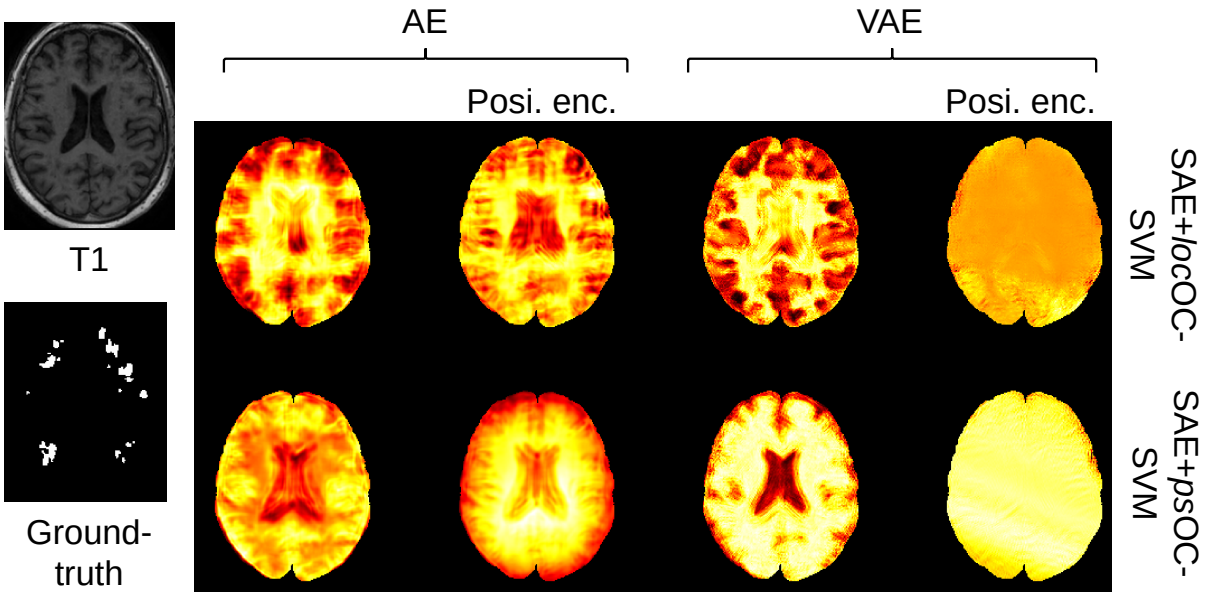

Figure V.9: Showcase of the methods presented section V.2.3.i for the patient AM114 (same patient as top row of figure IV.4).

OC-SVM seems to create redundancy on the positional constraint. We argue that these two mechanisms are not equivalent, as the positional encoding acts on the latent representation, whereas the *localized* OC-SVM operates on the way the outlier detection is done. Additionally, redundancy is not necessarily negative, as it can emphasize important concepts.

V.2.4 Conclusions and perspectives

We introduce in this section two ways to give additional structure to the latent space, with the goal of improving the learned representation space, which could, in the end, improve the final detection performances of the models. We believe that this objective has not been fulfilled, as the reported performances are worse when adding these additional encoding constraints than without. One track that we haven't explored in this section is how these additional constraints combine with the siamese constraint (that is used for the models in this section). It would be of great interest to repeat these experiments without this constraint, to make the comparison complete.

Another track that we did not explore, and particularly highlighted by the degenerated maps, would be to add a weighting to the reconstruction of the coordinates during training, that would be inferior to the weighting of the image, thus producing the encoding without interfering too much with the image reconstruction.

We believe that this series of experiments has shown that for subtle lesions detection, the SAE+*ps*OC-SVM model provided uninformative maps, whereas the SAE+*loc*OC-SVM, despite a poor specificity, was much more promising in terms of sensitivity. We also found that it was particularly important to look at the qualitative results, as seemingly satisfactory quantitative performances can hide uninformative visual score maps.

For these reasons, in the following section, we propose an alternative approach to structuring the latent space. We focus solely on the SAE+*loc*OC-SVM model, as it was shown more effective at identifying subtle lesions.

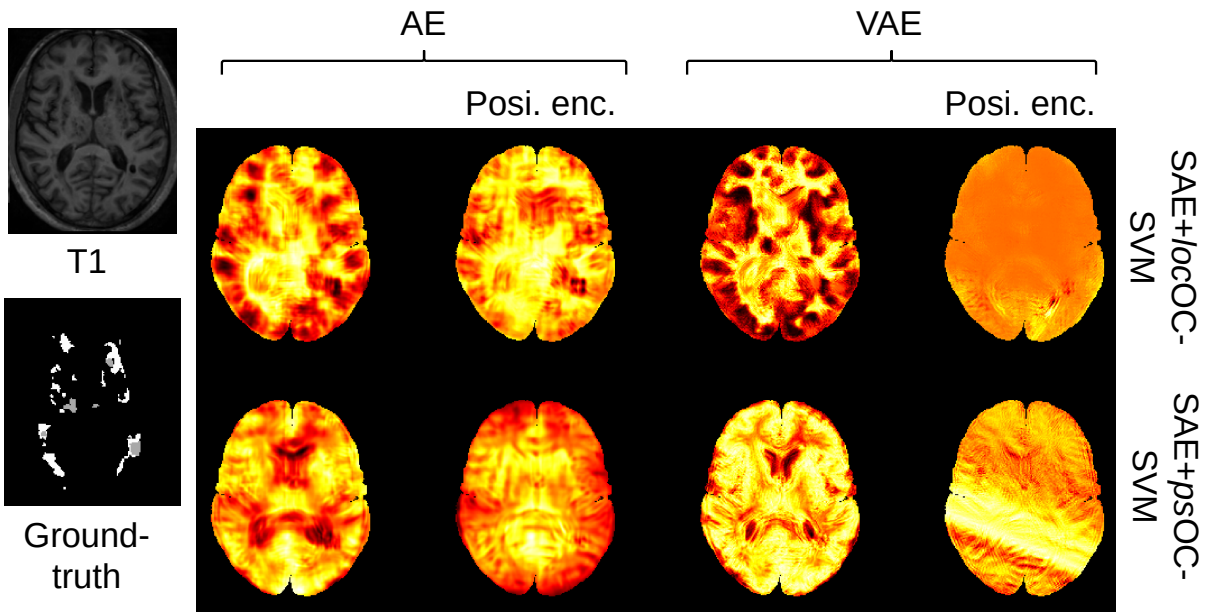

Figure V.10: Showcase of the methods presented section V.2.3.i for the patient UT4.

V.3 End-to-end support estimation

Throughout this thesis, we made extensive use of a framework composed of two steps: representation learning and outlier detection (whether with SAE+*loc*OC-SVM or SAE+*ps*OC-SVM). These two steps are totally decoupled, meaning there is no guarantee that the learned representations are optimal for the outlier detection step. In this section, we propose a novel method for fusing patch-based auto-encoder representation learning and one-class SVM. We name this method JeanZAD for Jean Zay Anomaly Detection, to pay tribute to the former French minister of national education Jean Zay[1].

We introduce this novel framework in section V.3.1, then, use this for subtle lesion detections, and compare it to the previously presented methods.

V.3.1 Fusion of auto-encoder and one class SVM

We recall the classical auto-encoder loss, for a batch of patches[2] $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$:

$$L_{AE}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2$$

where $\hat{\mathbf{x}}_i$ are obtained by decoding the latent variable $\mathbf{z}_i$, itself obtained by encoding $\mathbf{x}_i$. In other terms, with $\mathbf{E}$ the encoder and $\mathbf{D}$ the decoder, $\hat{\mathbf{x}}_i = \mathbf{D}(\mathbf{z}_i) = \mathbf{D}(\mathbf{E}(\mathbf{x}_i))$. $||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2$ is the reconstruction term between the input $\mathbf{x}_i$ and output of the auto-encoder.

[1] We thought that the notations like SAE+*loc*OC-SVM were already overloaded. Also, the supercomputer that we used for the computations for this thesis is named Jean Zay, referring to this famous minister whose ashes lie in the Panthéon. (`http://www.idris.fr/jean-zay/jean-zay-presentation.html`).

[2] Here, we present the batch of patches as any patches, without taking into account their localization or from which subject they come from, as this is not important for the following. The reader can understand the $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ as any sequence of patches, (co-localized if need be, from the same patient if need be).

The batch of patches, after encoding, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is split into two: one part is used to solve the one class SVM problem ($\mathbf{z}^{\mathbf{SVM}}{}_i$) and the other are used later on for the final loss ($\mathbf{z}^{\mathbf{Loss}}{}_i = \mathbf{z}^{\mathbf{L}}{}_i$):

$$\mathbf{z}^{\mathbf{SVM}}{}_i = \mathbf{z}_i \text{ with } i = \{1, \ldots, \frac{n}{2}\} \qquad \mathbf{z}^{\mathbf{L}}{}_i = \mathbf{z}_i \text{ with } i = \{\frac{n}{2}, \ldots, n\}$$

At each batch, we solve the one class SVM problem for the $\mathbf{z}^{\mathbf{SVM}}{}_i$:

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

which gives the optimal $\boldsymbol{\alpha}$: $\boldsymbol{\alpha}^*$. We recall that the decision function (positive for the estimated support of the $\mathbf{z}^{\mathbf{SVM}}{}_i$ and negative elsewhere) can then be written as:

$$f(\mathbf{z}) = \sum_{j=1}^{\frac{n}{2}} \alpha_j^* k(\mathbf{z}^{\mathbf{SVM}}{}_j, \mathbf{z}) - \rho^*$$

for any new sample $\mathbf{z}$. ρ^* is obtained through the same process[1] as described in section II.3.2. Note that $\boldsymbol{\alpha}^*$ and ρ^* are functions of the $\mathbf{z}^{\mathbf{SVM}}{}_i$.

We then propose to use the following loss for JeanZAD:

$$L_{JZAD}(\mathbf{x}) = \underbrace{\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2}_{\text{reconstruction term}} + \lambda \underbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -f(\mathbf{z}^{\mathbf{L}}{}_i))}_{\text{penalization of misclassified } \mathbf{z}^{\mathbf{L}}{}_i}$$

The second term (weighted by λ) penalize only the misclassified $\mathbf{z}^{\mathbf{L}}{}_i$, as the correctly classified $\mathbf{z}^{\mathbf{L}}{}_i$ will have positive decision function, and thus $\max(0, -f(\mathbf{z}^{\mathbf{L}}{}_i))$ will be 0. Misclassified $\mathbf{z}^{\mathbf{L}}{}_i$ will have penalization proportional to their euclidean distance to the estimated hyperplane.

The interest of separating the latent representation vectors into two parts $\mathbf{z}^{\mathbf{SVM}}{}_i$ and $\mathbf{z}^{\mathbf{L}}{}_i$ appears here: as the SVM frontier is estimated on the $\mathbf{z}^{\mathbf{SVM}}{}_i$, most of them are correctly classified[2]. This justifies the use of another set of latent vectors $\mathbf{z}^{\mathbf{L}}$. Penalizing samples not used for the support estimation amounts to penalize bad generalization to unseen samples. We can develop L_{JZAD} with the expression of f:

$$L_{JZAD}(\mathbf{x}) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 + \lambda \sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k(\mathbf{z}^{\mathbf{SVM}}{}_j, \mathbf{z}^{\mathbf{L}}{}_i) - \rho^*)$$

Recall that $\boldsymbol{\alpha}^*$ and ρ^* are functions of the $\mathbf{z}^{\mathbf{SVM}}{}_i$. If we separate the second term into what depends on $\mathbf{z}^{\mathbf{SVM}}{}_i$ and what depends on $\mathbf{z}^{\mathbf{L}}{}_i$, using the stopgradient operator sg[.] and $\beta_1 + \beta_2 = 1$, we can write L_{JZAD} as:

[1] We average the ρ obtained for every support vector for numerical stability.

[2] More precisely, the ν-property guarantees that no more than $\nu \times \frac{n}{2}$ of the $\mathbf{z}^{\mathbf{SVM}}{}_i$ would be misclassified.

$$L_{JZAD}(\mathbf{x}) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 + \lambda\beta_1 \overbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k(\mathbf{z^{SVM}}_j, \text{sg}[\mathbf{z^L}_i]) - \rho^*)}^{\text{Gradient flow only through the } \mathbf{z^{SVM}}_i} + \lambda\beta_2 \underbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \text{sg}[\alpha_j^*] k(\text{sg}[\mathbf{z^{SVM}}_j], \mathbf{z^L}_i) - \text{sg}[\rho^*])}_{\text{Gradient flow only through the } \mathbf{z^L}_i} \tag{V.2}$$

The value of the loss is the same with this derivation, however, the influence of $\mathbf{z^{SVM}}_i$ and $\mathbf{z^L}_i$ is now separated. We argue that the β_1 term, which gradient will flow through the $\mathbf{z^{SVM}}_i$, will influence the frontier of the SVM, as it will move samples in directions such that it includes the misclassified $\mathbf{z^L}_i$ in the frontier, we call this term an *expander*. The β_2 term, which gradient will flow through the $\mathbf{z^L}_i$, will influence the misclassified $\mathbf{z^L}_i$, as it will move the samples in directions such that they enter the boundary drawn by the $\mathbf{z^L}_i$, we call this term a *compactor*. Figure V.11 gives a visual intuition of this idea.

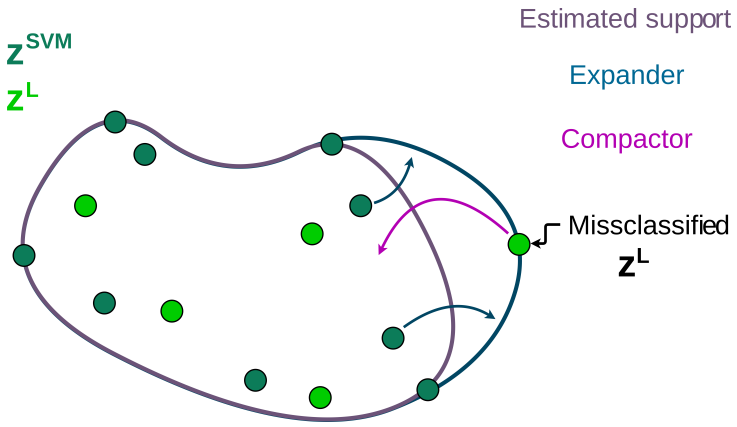

Figure V.11: At iteration N, some support (purple) is estimated using the $\mathbf{z^{SVM}}$. The *compactor* term acts on the misclassified $\mathbf{z^L}$ and pushes them inside the estimated support. The *expander* term acts on the $\mathbf{z^{SVM}}$ and pushes them outside the estimated support to include the misclassified $\mathbf{z^L}$.

β_1 and β_2 can be tuned to control the proportion of compaction/expansion wanted. The intuition would suggest that for a more diverse latent representation of the samples, which would allow easier classification, it would be better to have more expansion. We found in early experiments, by comparing ($\beta_1 = 0$, $\beta_2 = 1$), ($\beta_1 = 1$, $\beta_2 = 0$) and ($\beta_1 = \beta_2 = \frac{1}{2}$) that indeed the full *expander* gave the most promising results. One perspective to this framework would be to adaptatively tune the $\beta_{1,2}$ along the training, for instance, to extend the frontier at the beginning and then fix it. In the remainder of the manuscript, we use $\beta_1 = 1$ and $\beta_2 = 0$, i.e. the full *expander* setup.

V.3.2 Application to subtle lesions detection

V.3.2.i Experiments

The JeanZAD framework is designed to improve the final latent representation space with the targeted one-class classification as the final target. The learned representation can then be used

either with *loc*OC-SVM or *ps*OC-SVM. In section V.2, we concluded that the SAE+*loc*OC-SVM algorithm was the most promising model for the task of subtle anomaly detection. Also, preliminary experiments did not show conclusive results for JeanZAD when combined with *ps*OC-SVM. As a consequence, in the following experiments, we combine the latent representation learned with the JeanZAD model with the *loc*OC-SVM.

λ was set to 0.1 to balance the loss with the reconstruction error term, however, further experiments would be needed to adjust this parameter. For this section, to wrap up a part of the contributions made in this thesis, we recopy the results of SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and the two state-of-the-art methods re-implemented by Pinaya et al. (2022b) and Baur et al. (2021b). As not to overload the JeanZAD framework with an additional constraint, we drop the siamese constraint, meaning we consider only a simple auto-encoder AE[1].

The models are all trained on the same control database: the CERMEP database and tested on the WMH dataset. Hyperparameters that are not specified in this section are set to the same value as in section III.2.3.

V.3.2.ii Results and discussion

We present the quantitative performances in table V.6. Qualitative examples are presented figure in V.12, figure V.13, figure V.14 and figure V.15.

WMH (T1 only) 3 hospitals **'hyperintensities'**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	Siamese AE +*ps*OC-SVM	JZAD *expander* (*locOC-SVM*)
AU ROC	0.57±0.09	0.48±0.04	0.41±0.16	0.53±0.13	0.64±0.12
AU ROC 30	0.22±0.09	0.10±0.03	0.13±0.13	0.17±0.11	0.23±0.20
AU PR	0.022±0.023	0.013±0.013	0.017±0.018	0.015±0.012	0.040±0.054
AU PRO	0.47±0.05	0.50±0.06	0.51±0.20	0.52±0.15	0.57±0.12
AU PRO 30	0.10±0.04	0.11±0.05	0.18±0.17	0.13±0.10	0.14±0.14
$\lceil Dice \rceil$	0.05±0.05	0.03±0.03	0.04±0.04	0.04±0.03	0.08±0.09
WMH (T1 only) 3 hospitals **other pathologies**	VQ-VAE + Transformer restoration	AE recons. error	SAE +*loc*OC-SVM	Siamese AE +*ps*OC-SVM	JZAD *expander* (*locOC-SVM*)
AU ROC	0.71±0.10	0.50±0.05	0.60±0.15	0.50±0.15	0.75±0.09
AU ROC 30	0.40±0.18	0.14±0.05	0.26±0.22	0.17±0.17	0.41±0.18
AU PR	0.018±0.031	0.004±0.007	0.017±0.025	0.005±0.008	0.032±0.071
AU PRO	0.25±0.34	0.18±0.24	0.22±0.30	0.08±0.17	0.15±0.30
AU PRO 30	0.14±0.21	0.05±0.07	0.09±0.17	0.02±0.07	0.09±0.19
$\lceil Dice \rceil$	0.04±0.06	0.01±0.02	0.04±0.06	0.01±0.02	0.06±0.10

Table V.6: Mean metric over all patients from the 3 different hospitals for each method. *AU PR* for a random classifier would be 0.006±0.006 for 'hyperintensities' and 0.001±0.002 for other pathologies.

In table V.6, we see that JeanZAD slightly outperforms VQ-VAE+Transformer restoration (Pinaya et al. 2022b) for *AU ROC* (+0.07), *AU PRO 30* (+0.04), Best Dice (+0.03). This difference in performances increases furthermore for *AU PRO* (+0.10) and *AU PR* (+0.018). This seems to indicate better detection of the lesions, especially for the detections with the highest anomaly score (*AU PR*). JeanZAD outperforms SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and Baur et al. (2021b), for every metric.

[1] As described in the method section V.3.1, where we didn't add the siamese constraint in the loss (equation V.2).

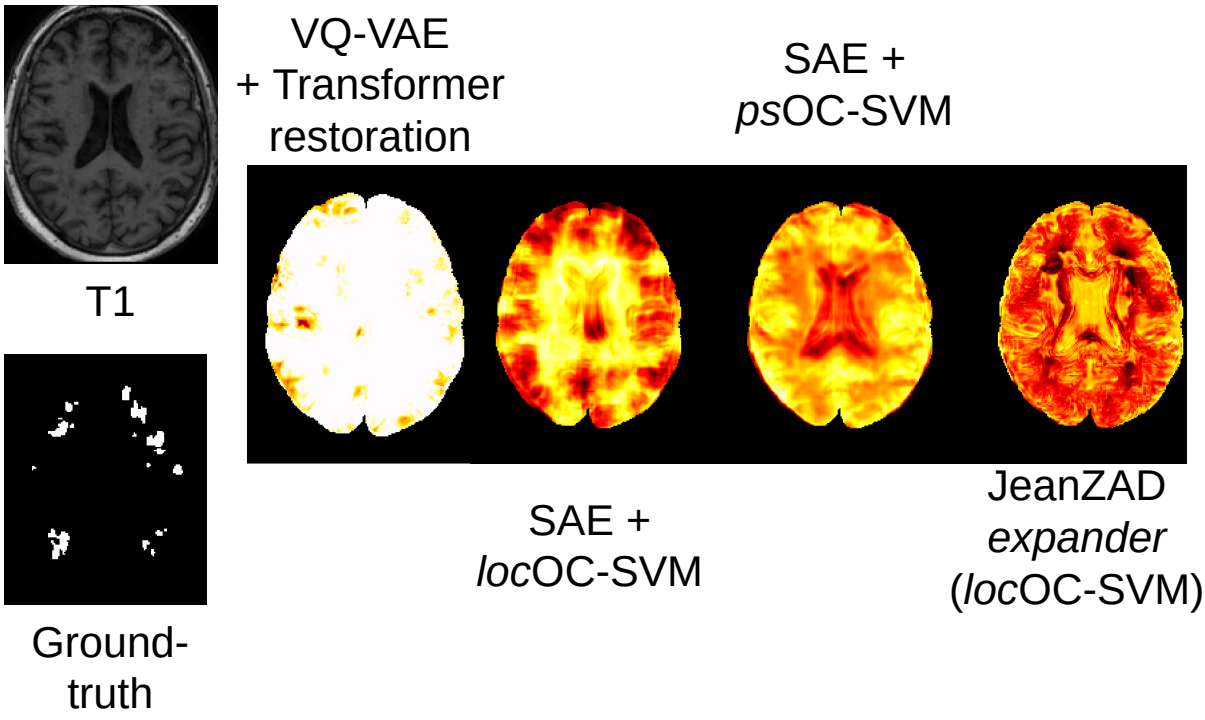

Figure V.12: Showcase of the JeanZAD framework, compared to SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and Pinaya et al. (2022b) for the patient AM114 (same patient as top row of figure IV.4)

By looking at the score maps presented in figure V.12, we see that JeanZAD, despite a saturated score map and a large number of false positives, has the most anomalous detections in the lesional zones. We think that, with the right threshold and with some additional post-processing, it has the potential to be relevant for clinicians (we develop this idea section V.4.1).

In figure V.13, the score map is less saturated for JeanZAD, but the detections seem greatly influenced by the enlarged ventricles of this patient. Despite this, we still see true positive detections. This flaw is also present in the other methods. On V.14, all the methods are able to detect the lacuna, except SAE+*ps*OC-SVM. The subtle lesions are missed by every method in this example.

In figure V.15, the score map of JeanZAD is again very saturated[1], still, the most anomalous scores are systematically true positives. In this example, SAE+*loc*OC-SVM is able to get a correct detection in the lower right area. Pinaya et al. (2022b), except for the same lower right lesion, produces a score map with many false negatives, as does SAE+*ps*OC-SVM, with the addition of false positives inside the ventricles.

V.3.3 Conclusion and perspectives

In this section, we presented a novel end-to-end model, that allows coupling of a patch-based auto-encoder and a one-class SVM framework. It was observed that the results were promising, as it was shown to outperform the method by Pinaya et al. (2022b) quantitatively, and qualitatively on the observed examples. Also, it outperforms the simple baseline presented in section V.1.2.

[1] As seem to be every anomaly score map on Amsterdam and Utrecht, indicating furthermore the need of domain adaptation between the different centers.

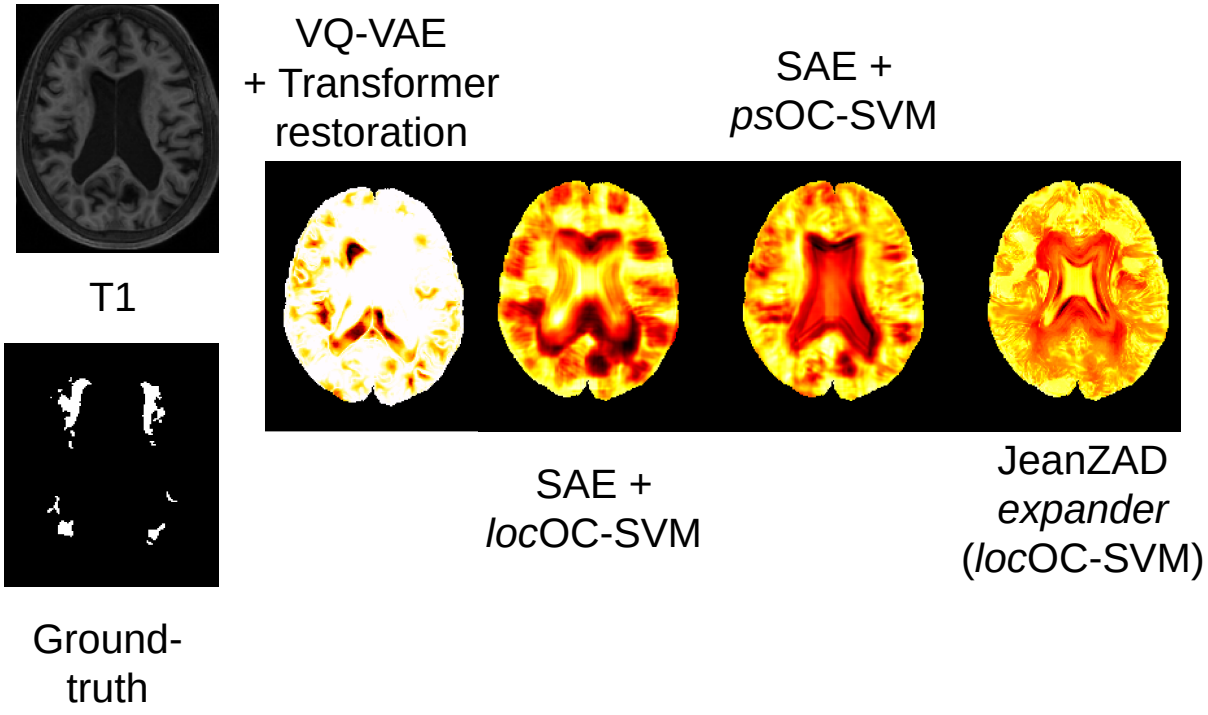

Figure V.13: Showcase of the JeanZAD framework, compared to SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and Pinaya et al. (2022b) for the patient SIN63 (same patient as middle row of figure IV.4)

The perspectives of this work are multiple. First, we introduced a novel method but evaluated it as a proof of concept. Additional experiments, including a large campaign of hyperparameter tuning, would have to be done to exploit the full potential of this method. Second, we need to include post-processing of the anomaly score maps, combination with VAE, as well as positional encoding or siamese constraint. The CSF segmentation step (described section B.2), which allows masking the ventricles in the final score maps, could also remove some false positives, despite the fact that the segmentation would be more difficult to obtain than in section IV.1.2, as it would have to be done on the T1 image only.

As we proposed a novel end-to-end representation learning combined to a support estimation method, it would seem natural to compare it to a representation learning model associated to a density estimation method[1]. The architecture of Zong et al. (2018), fusing an auto-encoder with a gaussian mixture model, would be a good candidate for this comparison.

Another perspective would be to compare the proposed method with other end-to-end representation learning + support estimation methods. To the best of our knowledge, the only method reported in the literature is the work of Zhou et al. (2021), where they combine a VAE with the deep SVDD framework (Ruff et al. 2018). However, we believe they do not compare, as there are several approximations done with Deep SVDD (presented paragraph I.3.1.iii:Deep SVDD, no radius learned, no center estimated, feature mapping learned explicitly) which are not used in our approach.

[1]Reconstruction methods are by design end-to-end.

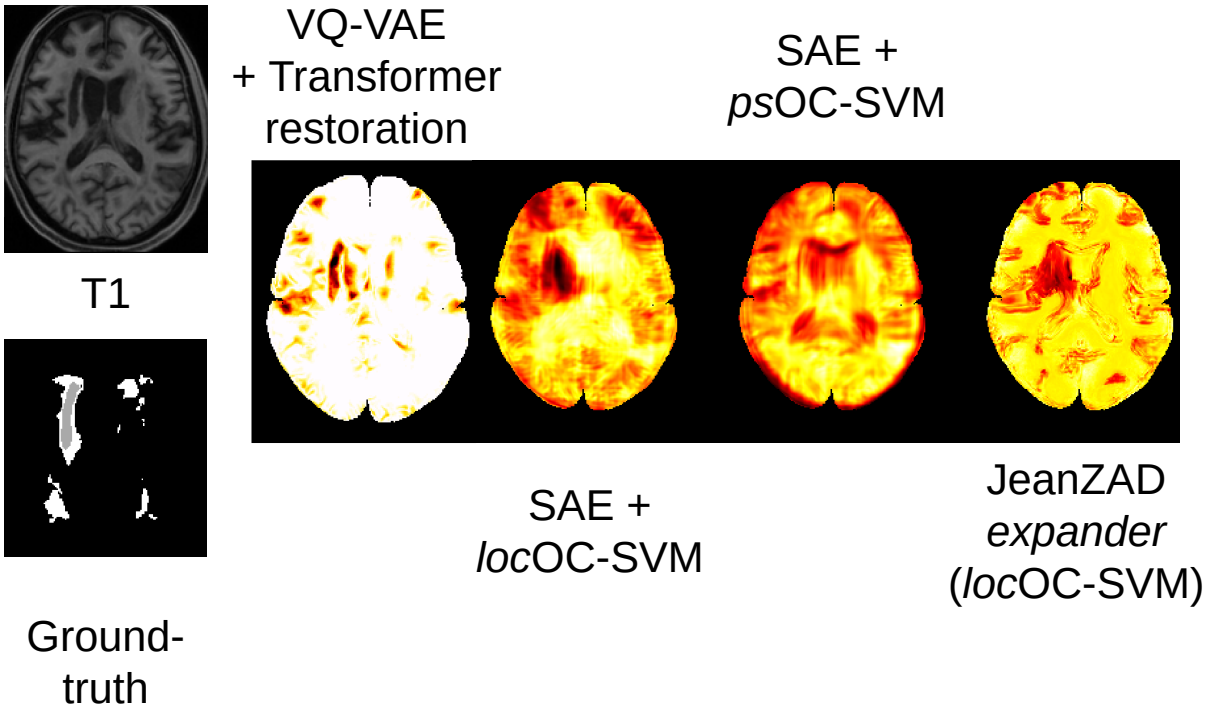

Figure V.14: Showcase of the JeanZAD framework, compared to SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and Pinaya et al. (2022b) for the patient SIN67 (same patient as presented in figure V.8). The light grey area on the ground truth corresponds to a lacuna (an example of 'other pathology')

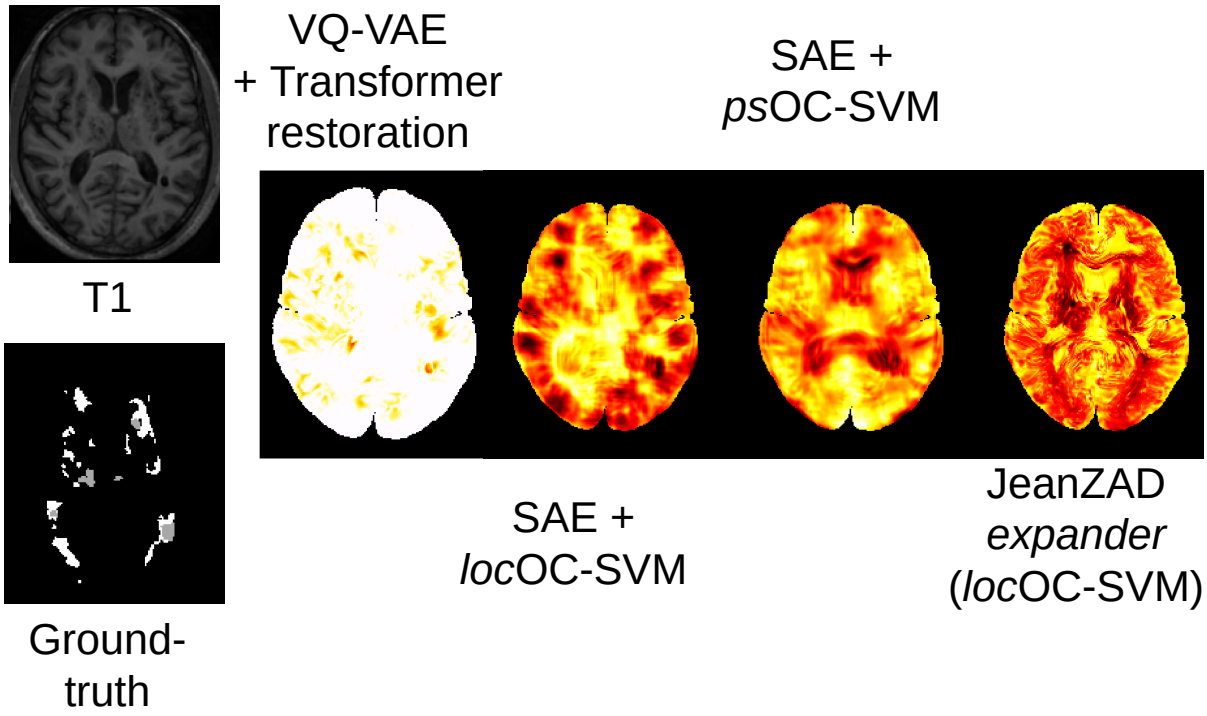

Figure V.15: Showcase of the JeanZAD framework, compared to SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and Pinaya et al. (2022b) for the patient UT4 (same patient as presented figure V.10)

V.4 Additional analyses of score maps and latent spaces for subtle lesions detection

We saw in the previous section many quantitative and qualitative results. However, these remain hard to fully understand and we wish to present a series of experiments designed to analyze the failures and successes of the proposed models. In section V.4.1, we first present additional qualitative results with cluster post-processing. In section V.4.2, we present an analysis of anomaly scores with respect to the intensity or size of the ground truth. Finally in section V.4.3, we present a visualization of the latent space obtained with the different auto-encoders.

V.4.1 Cluster analysis

As the produced score maps still generate a large number of false positives (poor specificity), we investigate in this section the possibility of proposing additional post-processing, with the main objective of generating outputs that are more suited for clinical practice.

We saw in section II.3.3.iii that Alaverdyan et al. (2020) implemented some post-processing steps to obtain cluster maps from the anomaly score maps. We implement this post-processing, consisting of thresholding the score map, extracting connected components, and removing small connected components[1], adding some morphological post-processing (3 dilatations and 1 closing). We display such example cluster maps in figure V.16 and V.17.

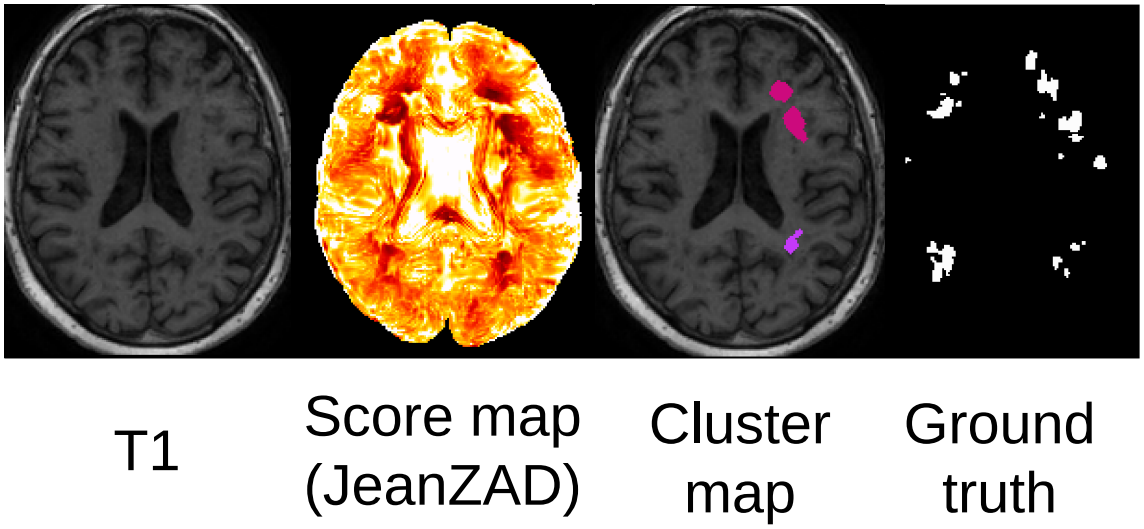

Figure V.16: Cluster map obtained from the anomaly score map, for the patient AM114 (same patient as the top row of figure IV.4)

In the first example (AM114, figure V.16), 4 clusters are detected (only 2 visible on the selected slice), and the 4 clusters intersect significantly with a ground truth lesion (true positives), thus we could say the precision is 1.0. However, the number of ground truths in the volume was 23 (23-4=19 false negatives), thus we could say the sensitivity was 0.17. In the second example (SIN63, figure V.17), 2 clusters are detected, each intersecting significantly with a ground truth lesion (precision 1.0), and the number of ground truth was 6 (sensitivity 0.33).

We see that, with this kind of analysis, we can 'clean' a lot of false positives, as we obtain very good precision. However, this improvement comes at the cost of poor sensitivity, as many true positives are removed in this process.

[1]For this analysis, we also remove the small ground truth, as they could not be detected.

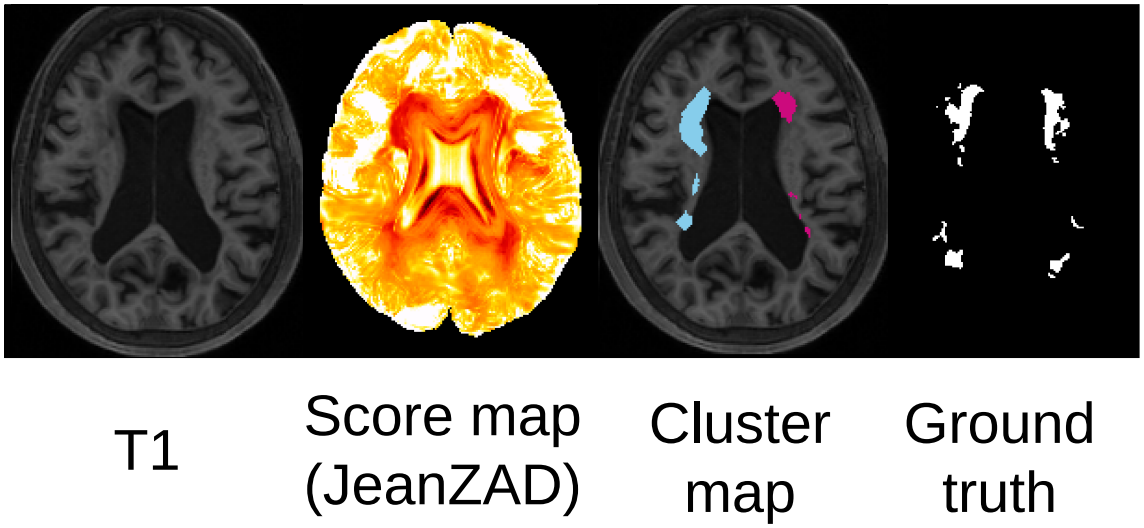

Figure V.17: Cluster map obtained from the anomaly score map, for the patient SIN63 (same patient as middle row of figure IV.4)

Nonetheless, this process allows outputting maps that are more practical to use for a radiologist, as the raw anomaly score maps have a lot of information to process.

V.4.2 True positives/False negatives analysis

In this section, we focus on lesions referred to as 'hyperintense' according to the WMH definition. We investigate how anomaly scores vary in the ground truth lesions depending on the T1 intensity values and lesion size for the three main methods studied in this thesis: SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and JeanZAD+*loc*OC-SVM. This aims to provide a different analysis of the behavior of the various models from that which can be achieved within the framework of the macroscopic performance metrics reported so far.

V.4.2.i Intensity plots

First, we want to study the relationship between intensity and anomaly scores in the lesions. We showed in section V.1.1 that we need to be careful about methods that would just be hyperintensity detectors. We saw in figure V.3 that the T1 intensity values of the so-called 'hyperintense' lesions are in fact well distributed over the range of intensities (they are neither hyperintense nor hypointense, they cover the whole intensity spectrum). Still, we want to study if the detected intensities could be the ones that are more hyperintense.

To be able to compare the anomaly scores of each method, that do not have a comparable range, we compute the z-score of each anomaly score: each score is standardized by the mean and standard deviation over every score (from every voxel in the brain). We perform this analysis separately for all lesions of all MRI exams of each of the 3 hospitals, as we already showed an important domain shift between the centers. Figure V.18 shows these standardized mean anomaly scores per true lesion (from the ground truth mask) as a function of the mean T1 intensity value in this lesion (thus one point corresponds to one ground truth lesion).

First, without taking into account the relationship between anomaly scores and intensity, we see that the different models do not perform the same across hospitals. JeanZAD seems to outperform the two other models on Amsterdam lesions because standardized anomaly scores are lower (more negatives) than those reported for the two other models, thus meaning that true

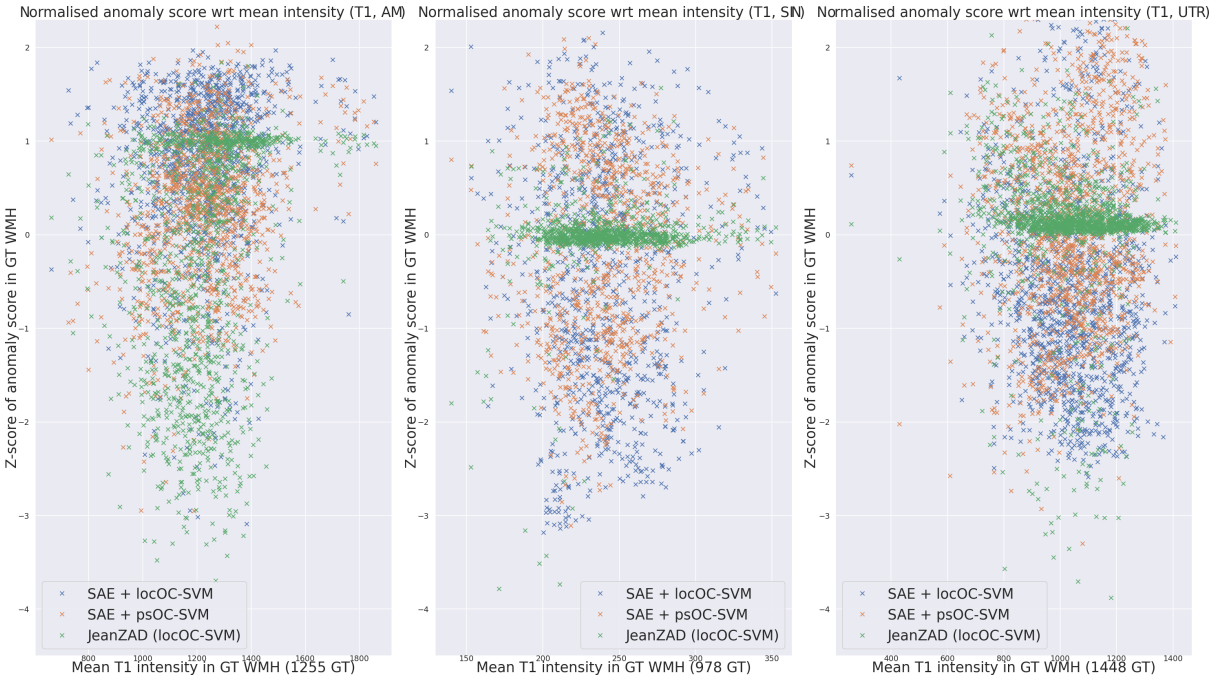

Figure V.18: Distribution of the standardized anomaly scores (the lower, the more anomalous) for SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and JeanZAD, with respect to the ground mean T1 intensity. Only 'hyperintense' lesions are displayed. From left to right: Amsterdam, Singapore, and Utrecht.

lesions are considered more anomalous than with the other models. This is the opposite on the Singapore lesions. On Utrecht, the scores output by the different models seem pretty mixed.

Overall, except on Amsterdam and Singapore where a very small fraction of hyperintense lesions are not detected (scored as normal, ie with positive values), there seems to be no correlation between the lesion intensity and the anomaly score, for every model. We believe it is a good indicator that the models do not solely rely on the T1 intensity value for classifying a voxel as suspicious.

V.4.2.ii Size plots

In this section, we investigate the link between ground truth size and anomaly scores provided by each method. The results are presented in figure V.18.

Apart from the general comment on the distribution of the scores along the hospitals, which is similar to what was observed regarding the influence of the T1 intensity value in the previous section, we notice several things. First, mean z-score values for very large regions are predominantly positive, thus meaning that none of the three considered models, on any hospital, seems to achieve the detection of very large lesions, with some exceptions with the JeanZAD model on Amsterdam lesions. Also on Amsterdam, the JeanZAD model is shown to produce very negative values on small lesions, thus enabling their detection, which is very encouraging. This is also the case, to a lesser extent, on Utrecht lesions.

V.4.3 Latent space analysis

Since from the quantitative and qualitative results only (section V.3.2.ii), it is still unclear whether we have achieved the goal of giving more structure to the latent space, in this section,

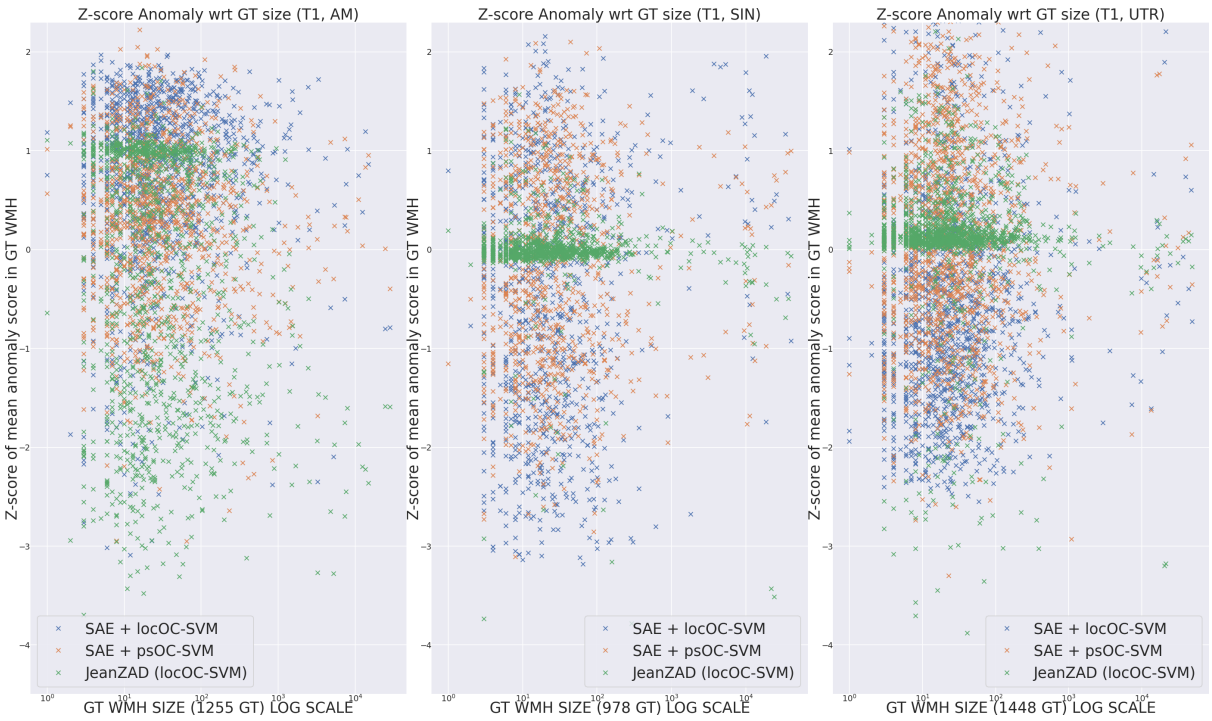

Figure V.19: Distribution of the standardized anomaly scores (the lower, the more anomalous) for SAE+*loc*OC-SVM, SAE+*ps*OC-SVM and JeanZAD, with respect to the ground truth size (in log scale). Only 'hyperintense' lesions are displayed. From left to right: Amsterdam, Singapore, and Utrecht.

we want to get more insight into the structure of latent space of the different models considered in this study. We specifically focus on comparing the JeanZAd and SAE models.

To visualize the latent spaces produced by these two auto-encoders, we use the McInnes et al. (2018) method. It is a non-linear reduction technique, similar to t-SNE (Van der Maaten and Hinton 2008). Without entering into too many details, UMAP ensures that points that are close in the original space (which is the latent space in our case) will be close in the projected space (2-dimensional for visualization).

We design three experiments with UMAP that we detail below, based on the three patch's characteristics: 1) their localization, 2) whether they are extracted from a control or a patient (from the Amsterdam, Singapore or Utrecht datasets), and 3) whether they are centered at the localization of a healthy tissue, 'hyperintense'/WMH, or other pathological lesions.

As the UMAP optimization is stochastic, the representations can vary from run to run, even though the conclusions should be the same. To account for this stochasticity, we run each experiment twice (with the same latent representations), to obtain a second UMAP projection.

V.4.3.i Control plot

For this experiment, only healthy localizations are sampled. The subjects and localizations differ from sample to samples (subjects are indicated on the plots). This "control plot" allows us to see the latent space structure in the absence of any pathology. Figure V.20 and V.21 present the results of the duplicated experiment.

Figure V.20: UMAP projection of latent vectors obtained with SAE (left) and JeanZAD (right), for healthy patches only, from different localizations and different subjects (subject number reported at each point). Best viewed zoomed in on digital format.

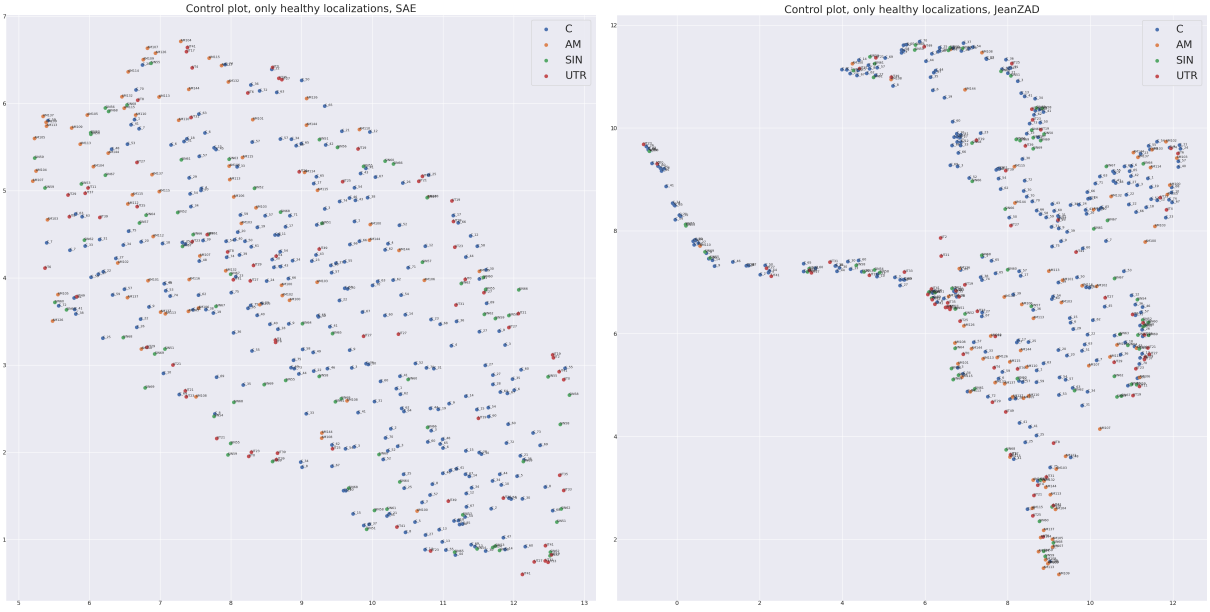

Figure V.21: UMAP projection of latent vectors obtained with SAE (left) and JeanZAD (right), for healthy patches only, from different localizations and different subjects (subject number reported at each point). Replica of figure V.20.

The first thing we notice is that the latent space of JeanZAD is more structured, in the sense that the points are not almost uniformly distributed as in SAE. It is difficult to determine whether this structure helps the outlier detection step, but we may argue that the goal of structuring the latent space is achieved to some extent.

For the SAE, we see occasionally on the border of the cloud some groups composed exclusively of non-controls. This finding is exacerbated for JeanZAD, where one of the "branches" (on the bottom for the two replicas), such a group of non-controls is found. This could be a sign of the domain shift that appears between controls and patients, even for the healthy tissues.

V.4.3.ii Patient plot

For this experiment, we sample patches extracted from different localizations within a patient. These patches may be centered on voxels located either on normal tissue (healthy) or on hyperintense (WMH) or other pathological lesions. This "patient plot" allows us to see how different patches from the same patient are distributed. Figure V.22 presents this experiment and figure V.23 its replica.

In this plot, both encoders demonstrate some structure for a fixed patient. The three categories of labels are clearly separated. However, it is observed that the SAE better separates WMH from the rest, while JeanZAD better separates other pathologies from the rest.

V.4.3.iii Localization plot

For this experiment, we sample patches extracted from a fixed localization in the brain for different subjects. The center of these patches may be located on normal tissue (healthy) or on hyperintense (WMH) or other pathological lesions depending on the subject. This "localization plot" allows us to see how patches at the same localization from different subjects are distributed. Figure V.24 presents this experiment and figure V.25 its replica.

The conclusions are hard to draw from these plots, as both seem to separate the different points in clusters, but each of these clusters often contains a WMH patch.

V.4.4 Conclusion and perspectives

In this section, we proposed additional analyses to compare the three main models studied in this thesis: SAE+*loc*OC-SVM, SAE+*ps*OC-SVM, and JeanZAD. The cluster analysis, although promising because of its ability to clean a large number of false positives, was also shown to remove a considerable number of true positives. However, we believe that more work could be done to improve these cluster maps, which ultimately provide valuable and low-burden information for clinicians.

We then showed that the proposed models are not simple 'hyperintensities' detectors for T1 and that they frequently failed to detect large lesions. A potential direction for future work would be to combine the strengths of the various models to enhance overall sensitivity.

We then showed various plots of the latent spaces obtained from SAE and JeanZAD. While these plots have to be considered very carefully, we showed that JeanZAD seemed to have a more structured space than SAE, but additional work remains to investigate how this latent space is organized (in terms of intensity for example). It would also be of great interest to analyze how these latent structures evolve as the encoder learns, i.e. through training epochs.

Figure V.22: UMAP projection of latent vectors obtained with SAE (left) and JeanZAD (right), for a fixed patient (UT0) only, from different localizations (healthy/WMH/other pathologies), coordinates reported at each point. Best viewed zoomed in on digital format.

Figure V.23: UMAP projection of latent vectors obtained with SAE (left) and JeanZAD (right), for a fixed patient (UT0) only, from different localizations (healthy/WMH/other pathologies), coordinates reported at each point. Replica of figure V.23.

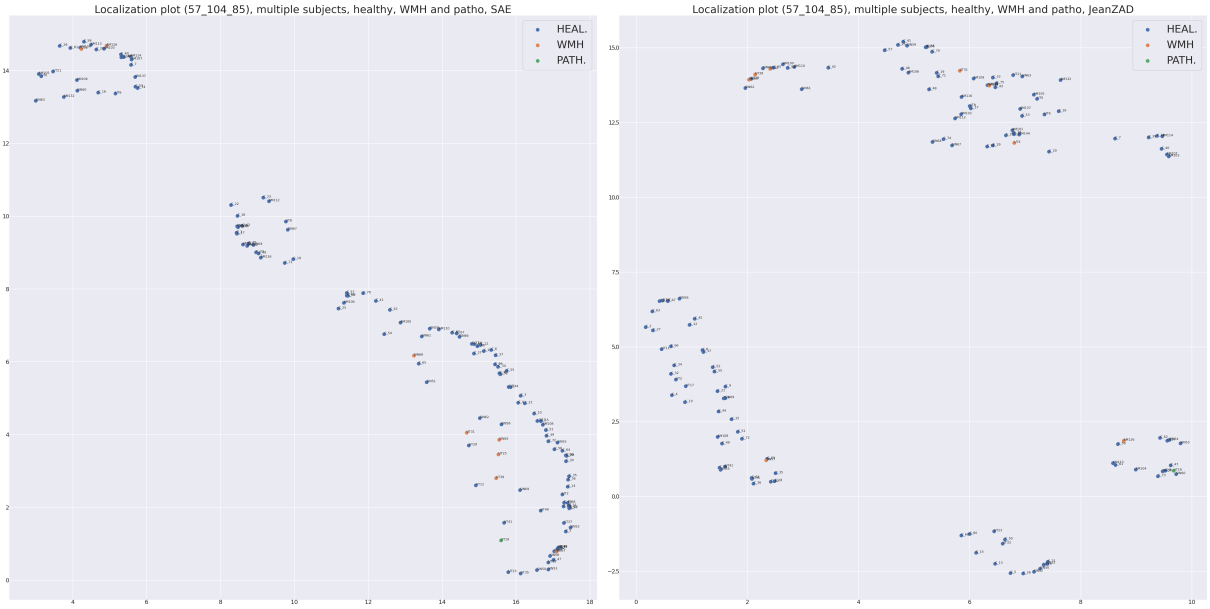

Figure V.24: UMAP projection of latent vectors obtained with SAE (left) and JeanZAD (right), for a fixed localization, for different subjects (number reported at each point) having different conditions at this localization (healthy/WMH/other pathologies). Best viewed zoomed in on digital format.

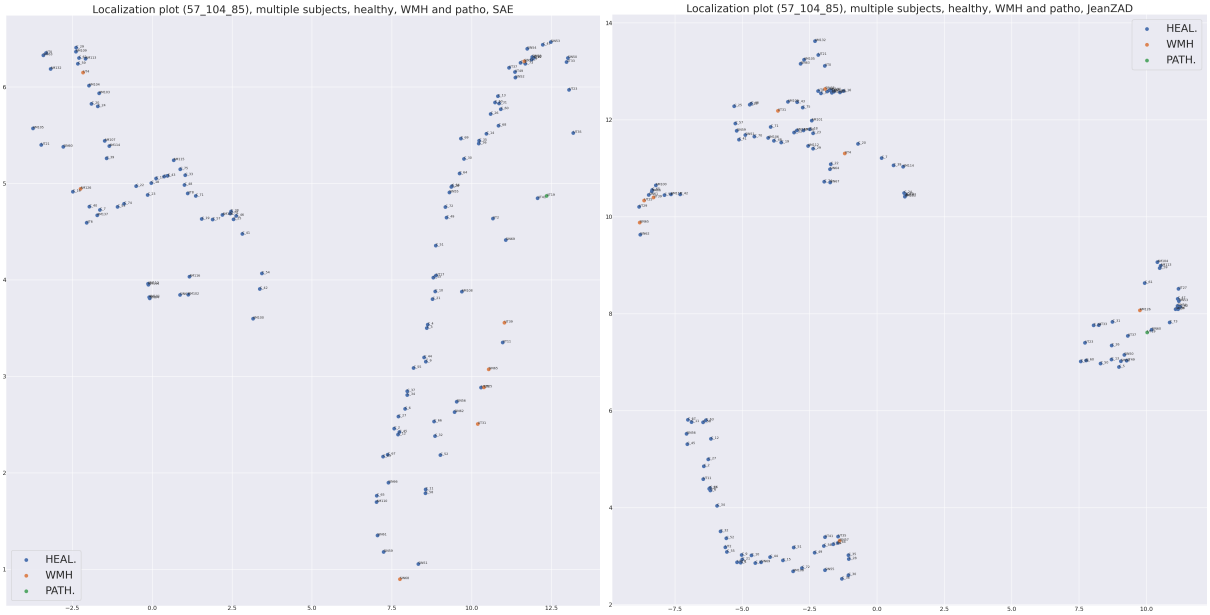

Figure V.25: UMAP projection of latent vectors obtained with SAE (left) and JeanZAD (right), for a fixed localization, for different subjects (number reported at each point) having different conditions at this localization (healthy/WMH/other pathologies).

V.5 Conclusion

In this chapter, we first demonstrated that it was necessary to tackle more challenging tasks, such as subtle lesion detection on T1 MRI. However, assessing the performance of various unsupervised detection models for this task was shown challenging due to its complexity.

We then proposed to add more structure to the latent space of the patch-based auto-encoder to improve detection performance. Our initial attempts to achieve this by adding variational constraint or positional encoding did not yield positive results. However, additional experiments remain to be done to see if the siamese constraint interfered with the proposed improvements.

Then, we introduced a novel method, that fuses the representation learning step and the outlier detection step. We derived the loss function for this model and proposed an intuitive way to think about the two loss terms (*expander*/*compactor*). We demonstrated that this model achieved superior results in detecting subtle lesions, surpassing both state-of-the-art methods and previously proposed methods. However, these findings need to be nuanced, as additional evaluation, on other tasks, would be needed to confirm the promising performance of such a model. Moreover, the influence of the *expander*/*compactor* terms on the performances remains to be studied, as well as their ability to structure the latent space.

In the last section, we presented additional plots and visualizations to aid in understanding the impact of the end-to-end framework on the latent space. We also evaluated performance in terms of intensity and size and conducted a preliminary cluster analysis. These plots, although requiring careful interpretation, showed that we achieved some level of latent space structuration and some clinical relevance, by detecting lesions that were neither hyperintense, nor too large, and we provided preliminary lesion maps of clinical interest for computer-aided diagnosis.

Conclusion

General conclusion

This work aimed at providing methodological contributions to the field of unsupervised anomaly detection (UAD) in neuroimaging. Specifically, we investigated methods that estimate the probability density support of the normative distribution in a latent representation space.

After introducing the fundamental concepts of UAD and the classical metrics and databases used, we conducted a bibliographic review of the current state-of-the-art methods for UAD in medical imaging. We also identified some of the gaps in the current literature and acknowledged the limitations of the study by Alaverdyan et al. (2020).

Our first contribution was to extend the evaluation of the method proposed by Alaverdyan et al. (2020) on multiple public databases, and compare its performances to state-of-the-art methods. We conducted this comparison on three datasets: a popular industrial anomaly detection dataset (MVTecAD), a public MRI challenge dataset for brain white matter lesion segmentation (WMH), both providing reference segmentation mask and finally, a multiparametric MRI dataset (PPMI) where anomaly detection was used as a proxy to perform classification task of Parkinson versus control subjects.

Our second contribution consisted of two parts aimed at improving the robustness of our proposed unsupervised anomaly detection model. Firstly, we proposed a new framework for the one-class SVM, that allows training a unique model specific to each patient. This model learns the support frontier based on patches extracted from the patient only, thus removing the dependence on the limited size of the control training set and less sensitive to spatial registration errors. This new strategy was successfully applied to segmentation and classification tasks, on WMH and PPMI datasets, respectively, and demonstrated superior performance compared to established state-of-the-art methods on WMH. Secondly, we addressed the problem of converting the unbounded anomaly scores to probabilistic outputs. This notably allowed for ensemble model learning or score map uniformization.

Our final contribution aimed to provide more structure to the latent space of auto-encoders for UAD. We proposed achieving this through variational regularization or positional encoding. Then, we introduced a novel UAD model that allows coupling training of the auto-encoder and the one-class SVM in an end-to-end fashion. The performance of this novel architecture was evaluated on the WMH dataset, but considering the much more challenging detection task of brain lesion detection in T1 MRI, whose signal is very subtle and not hyperintense like in FLAIR images. We also provided further analysis of the success and failure of these models.

Limits and perspectives

We found that the performance achieved on the different subdatasets of the WMH database was influenced by the characteristics of this database, including the demographic statistics of the population and the acquisition conditions of the MR images. For example, the population of the WMH database was, on average, much older than the control population used to learn the normative brain representation. Considering that brain shrinkage is a recognized aging effect, our model was sensitive to such an effect, which could lead to shrinkage of gray matter gyri and thus potentially induce a large number of false detections. The other main characteristic that affected the detection performance was the difference in scanner and/or acquisition parameters between the different datasets, i.e. between the control database and the three patient datasets acquired in different hospitals (Utrecht, Amsterdam, and Singapore). The difference in signal pattern induced by these different acquisition conditions affected the detection performance. To mitigate this problem, domain adaptation techniques could be implemented.

The developed pipeline is partially dependent on the quality of registration to a common atlas. Further work could be done to develop a registration-free algorithm, as small lesions can be erased during registration, affecting the sensitivity of the algorithm, and false positives seem to be generated at the site of registration errors, such as gray matter gyri and ventricles.

The data that we studied in this thesis were all volumetric, i.e. 3D, while the presented approaches only considered slices (2D) of the volumes. Approaching the problem as 2D makes sense because of the acquisition process of the data (usually slice by slice in MRI), the clinicians' reference view for some images, and the lower computational load. However, we experienced losses of spatial context, particularly with the anomalous detection of brain shrinkage in the elderly. A natural extension of this work would be to use 3D cubes instead of 2D patches. We also motivated the use of patches instead of whole images, but another meaningful comparison would be to study the implementation of the proposed method on whole slices, even though it would require additional adaptations to obtain a per-voxel latent representation. Lastly, the amount of available control data has increased since the start of this thesis. It would be beneficial to include these additional controls since the control database used in this work was limited in size (75 subjects).

Since we have introduced methods to convert anomaly scores into probabilistic outputs, this opens up a natural clinical application: late fusion of multi-modal image data, such as PET/MRI. Early fusion can be done by considering multi-modal data as channels, but late fusion (of score maps) can only be achieved with calibrated outputs. This fusion could also improve the sensitivity and specificity of the proposed methods. A combination of the different proposed methods could also be performed. Another approach would be to study the fusion of heterogeneous multi-modal data, such as images and text. This could be achieved by incorporating the non-image data in the latent space. Weakly-supervised approaches, which were not studied in this thesis, could also be used for this fusion. While collecting a large number of anomalies may seem unrealistic, collecting a small portion is a viable option in a clinical setting.

In this thesis, we motivated the study of support estimation methods and proposed a novel architecture combining end-to-end learning of the representation and support estimation model. A natural methodological extension of this work would be the comparison with end-to-end density estimation methods, as well as a more fundamental proof of concept on simpler datasets. Other methods, besides one class SVM could be studied and coupled with the auto-encoder latent feature extractor. Also, there have been a lot of scientific breakthroughs regarding deep learning in recent years: a more advanced analysis of the training of the models, choice of

optimization algorithms, regularizations, and hyperparameters tuning, could be undertaken to further optimize the model's capabilities.

Another perspective of this thesis is to examine how clinical characteristics such as age, sex, and risk factors affect the latent representations. One promising idea would be to explicitly structure the latent space according to these criteria, with the use of additional constraints on the auto-encoders. Also, the interactions between the siamese constraint and the other added constraints (variational, localization, end-to-end) were not studied during this thesis and could be of great interest.

False positives were sometimes generated due to a lack of clear definition of the concept of 'anomaly', such as brain shrinkage in elders producing anomalies on the cortex borders. Mitigating this pitfall could be achieved by characterizing the anomaly found with weakly-supervised or fully-supervised approaches. Also, many post-processing steps could be added to gear the score maps towards the desired outputs, by removing expected/common false positives. Finally, obtaining the most clinically relevant outputs still requires some work, as one of the goals of this research is the use of UAD algorithms in clinical practice.

Publication list

Journal articles

Zotova Daria, **Pinon Nicolas**, Trombetta Robin, Bouet Romain, Jung Julien, Lartizien Carole. (2023). GAN-based synthetic FDG PET images from T1 brain MRI can serve to improve performance of deep unsupervised anomaly detection models. *Artificial intelligence In medicine Submitted*

International conference with proceedings

Muñoz-Ramírez Verónica, **Pinon Nicolas**, Forbes Florence, Lartizen Carole, Dojat Michel. (2021). Patch vs. global image-based unsupervised anomaly detection in MR brain scans of early Parkinsonian patients. *Machine Learning in Clinical Neuroimaging: 4th International Workshop, MLCN 2021*, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4. *Poster*

Pinon Nicolas, Oudoumanessah Geoffroy, Trombetta Robin, Dojat Michel, Forbes Florence, Lartizen Carole. (2023). Brain subtle anomaly detection based on auto-encoders latent space analysis: application to de novo parkinson patients. *IEEE 20th International Symposium on Biomedical Imaging*, 2023. *Oral*

Pinon Nicolas, Trombetta Robin Lartizien Carole. (2023). One-Class SVM on siamese neural network latent space for Unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities. *2023 Medical Imaging with Deep Learning*, Proceedings of Machine Learning Research – 217:1–15, 2023. *Poster*

National conference with proceedings

Pinon Nicolas, Trombetta Robin, Lartizien Carole. (2023). Anomaly detection in image or latent space of patch-based auto-encoders for industrial image analysis. *GRETSI 2023 : XXIXème Colloque Francophone de Traitement du Signal et des Images*, Aug 2023. *Poster*

National conference

Pinon Nicolas, Oudoumanessah Geoffroy, Trombetta Robin, Dojat Michel, Forbes Florence, Lartizen Carole. (2023). Brain subtle anomaly detection based on auto-encoders latent space analysis: application to de novo parkinson patients. *Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale (IABM 2023). Poster*

Pinon Nicolas, Trombetta Robin, Lartizen Carole. (2024). Unsupervised anomaly detection of the white matter in MRI by estimation of the support of the normative distribution in siamese auto-encoder latent space. *Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale (IABM 2024). Poster*

Bibliography

Aizerman, A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837. 49

Akcay, S., Ameln, D., Vaidya, A., Lakshmanan, B., Ahuja, N., and Genc, U. (2022). Anomalib: A deep learning library for anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1706–1710. IEEE. 66

Alaverdyan, Z., Jung, J., Bouet, R., and Lartizien, C. (2020). Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening. *Medical Image Analysis*, 60:101618. xiv, xvii, xviii, xx, xxii, xxiii, xxiv, xxvi, xxxvi, 1, 30, 37, 41, 45, 46, 47, 50, 51, 55, 60, 62, 81, 84, 85, 88, 90, 132, 141

Arnaud, A., Forbes, F., Coquery, N., Collomb, N., Lemasson, B., and Barbier, E. L. (2018a). Fully automatic lesion localization and characterization: Application to brain tumors using multiparametric quantitative mri data. *IEEE Transactions on Medical Imaging*, 37(7):1678–1689. 22, 29, 97

Arnaud, A., Forbes, F., Coquery, N., Collomb, N., Lemasson, B., and Barbier, E. L. (2018b). Fully automatic lesion localization and characterization: Application to brain tumors using multiparametric quantitative mri data. *IEEE transactions on medical imaging*, 37(7):1678–1689. 37

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26:839–851. 163

Bakker, R., Tiesinga, P., and Kötter, R. (2015). The scalable brain atlas: instant web-based access to public brain atlases and related content. *Neuroinformatics*, 13:353–366. 78

Ballard, D. H. (1987). Modular learning in neural networks. In *Proceedings of the sixth National conference on Artificial intelligence - Volume 1*, AAAI'87, pages 279–284, Seattle, Washington. AAAI Press. 17

Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S. (2021a). Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952. 9, 31, 36, 45

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. (2020). Scale-space autoencoders for unsupervised anomaly segmentation in brain mri. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 552–561. Springer. 31, 38

Baur, C., Wiestler, B., Muehlau, M., Zimmer, C., Navab, N., and Albarqouni, S. (2021b). Modeling Healthy Anatomy with Artificial Intelligence for Unsupervised Anomaly Detection in Brain MRI. *Radiology: Artificial Intelligence*, 3(3):e190169. xiv, xxiv, xxviii, 31, 36, 57, 70, 71, 72, 74, 88, 89, 90, 91, 93, 114, 118, 119, 128

Behrendt, F., Bengs, M., Rogge, F., Krüger, J., Opfer, R., and Schlaefer, A. (2022). Unsupervised anomaly detection in 3d brain mri using deep learning with impured training data. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. 15, 31, 36

Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., and Schlaefer, A. (2023). Patched diffusion models for unsupervised anomaly detection in brain mri. In *Medical Imaging with Deep Learning*. 15, 33, 36

Bercea, C. I., Neumayr, M., Rueckert, D., and Schnabel, J. A. (2023a). Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*. 33, 39

Bercea, C. I., Wiestler, B., Rueckert, D., and Schnabel, J. A. (2023b). Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. In *Medical Imaging with Deep Learning*. 15, 32, 39

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2021). The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, 129(4):1038–1059. xviii, 1, 10, 11, 14, 27, 32, 45, 53, 58, 67

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, Long Beach, CA, USA. IEEE. 14

Bieder, F., Wolleb, J., Sandkühler, R., and Cattin, P. C. (2022). Position regression for unsupervised anomaly detection. In *International Conference on Medical Imaging with Deep Learning*, pages 160–172. PMLR. 33

Blackard, J. (1998). Covertype. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50K5N. 13

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., et al. (2020). Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283. 44

Bowles, C., Qin, C., Guerrero, R., Gunn, R., Hammers, A., Dickie, D. A., Valdés Hernández, M., Wardlaw, J., and Rueckert, D. (2017). Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 16:643–658. 29, 30, 33, 38, 50, 51

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. 49

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104. 14

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1993). Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6. 28, 46, 47

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27. 100

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758. 46

Chen, X. and Konukoglu, E. (2018). Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In *Medical Imaging with Deep Learning*. 15, 32, 38

Chen, X., Mishra, N., Rohaninejad, M., and Abbeel, P. (2018). Pixelsnail: An improved autoregressive generative model. In *International Conference on Machine Learning*, pages 864–872. PMLR. 26, 28

Chen, X., Pawlowski, N., Glocker, B., and Konukoglu, E. (2021). Normative ascent with local gaussians for unsupervised lesion detection. *Medical Image Analysis*, 74:102208. 15, 34, 36

Chen, X., You, S., Tezcan, K. C., and Konukoglu, E. (2020). Unsupervised lesion detection via image restoration with a normative prior. *Medical image analysis*, 64:101713. 34, 37

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., and Weller, A. (2021). Rethinking attention with performers. In *International Conference on Learning Representations*. 72

Ciuşdel, C. F., Itu, L. M., Cimen, S., Wels, M., Schwemmer, C., Fortner, P., Seitz, S., Andre, F., Buß, S. J., Sharma, P., et al. (2022). Normalizing flows for out-of-distribution detection: Application to coronary artery segmentation. *Applied Sciences*, 12(8):3839. 30, 39

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Ameli, R., Ferré, J.-C., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1):13650. 31

Correia, M. M., Rittman, T., Barnes, C. L., Coyle-Gilchrist, I. T., Ghosh, B., Hughes, L. E., and Rowe, J. B. (2020). Towards accurate and unbiased imaging-based differentiation of parkinson's disease, progressive supranuclear palsy and corticobasal syndrome. *Brain communications*, 2(1):fcaa051. 79

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297. 48

Daniel, T. and Tamar, A. (2021). Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4391–4400. 32

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 233–240, Pittsburgh, Pennsylvania. ACM Press. 9

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer. xx, 22, 27, 65, 67

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22. 95

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee. 13, 65

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., et al. (2019). Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170. 81

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302. 12

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. 27, 72

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, pages 241–252. 89

El Azami, M., Hammers, A., Jung, J., Costes, N., Bouet, R., and Lartizien, C. (2016). Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem. *PloS one*, 11(9):e0161498. 30, 37, 50, 51

El Azami, M., Lartizien, C., and Canu, S. (2017). Converting SVDD scores into probability estimates: Application to outlier detection. *Neurocomputing*, 268:64–75. 101, 102, 103, 104

European Union (2022). `https://ec.europa.eu/eurostat/fr/web/products-eurostat-news/w/ddn-20230222-1`. 71

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., and Fookes, C. (2021). Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37. 28

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102. 118

Forbes, F. and Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984. 29, 95

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press. 17, 18, 22, 61

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27. 21, 31

Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., and Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, 19(4):224–247. 163

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 60, 65

Hertz, J. A. (1991). *Introduction to the theory of neural computation*. Crc Press. p. 45-46. 31

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*. 31

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851. 22

Hoehn, M. M. and Yahr, M. D. (1967). Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427. 79

Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., and Weinberger, K. (2019). Convolutional networks with dense connectivity. *IEEE PAMI*. 96

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95. 164

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr. 60

JGraph Ltd (2021). Diagrams.net (previously draw.io). `https://github.com/jgraph/drawio`. 164

Kascenas, A., Pugeault, N., and O'Neil, A. Q. (2022a). Denoising autoencoders for unsupervised anomaly detection in brain mri. In *International Conference on Medical Imaging with Deep Learning*, pages 653–664. PMLR. 15, 33, 38

Kascenas, A., Young, R., Jensen, B. S., Pugeault, N., and O'Neil, A. Q. (2022b). Anomaly detection via context and local feature matching. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE. 15, 34, 37

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR), Poster*. 17, 89

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. *preprint, arXiv:1312.6114 [cs, stat]*. 19, 20

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, pages 583–621. 89

Kuijf, H. J., Casamitjana, A., Collins, D. L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Llado, X., Biesbroek, J. M., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.-Y., Park, H., Park, S. H., Pezold, S., Puybareau, E., De Bresser, J., Rittner, L., Sudre, C. H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Heinen, R., Chen, C., van der Flier, W., Barkhof, F., Viergever, M. A., Biessels, G. J., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., and Cardoso, M. J. (2019). Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568. xviii, xxiv, 1, 31, 33, 44, 54, 70

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86. 19

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 13, 14, 45

Lee, G. and Scott, C. D. (2007). The one class support vector machine solution path. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–521. IEEE. 103

Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., and Špiclin, Ž. (2018). A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16:51–63. 31

Li, W., Mo, W., Zhang, X., Squiers, J. J., Lu, Y., Sellke, E. W., Fan, W., DiMaio, J. M., and Thatcher, J. E. (2015). Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging. *Journal of Biomedical Optics*, 20(12):121305. 29, 39

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE. 13, 14

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., and Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31. 120, 121

Loshchilov, I. and Hutter, F. (2018). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. 17

Luo, G., Xie, W., Gao, R., Zheng, T., Chen, L., and Sun, H. (2023). Unsupervised anomaly detection in brain MRI: Learning abstract distribution from massive healthy brains. *Computers in Biology and Medicine*, 154:106610. 15, 31, 38

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644, ICLR 2016 Submission*. 31, 32

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C. M., Trojanowski, J. Q., et al. (2018). The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477. xviii, xxiv, 1, 15, 31, 43, 44

Marimont, S. and Tarroni, G. (2021a). Implicit field learning for unsupervised anomaly detection in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 189–198. Springer. 15, 34, 37

Marimont, S. N. and Tarroni, G. (2021b). Anomaly Detection Through Latent Space Restoration Using Vector Quantized Variational Autoencoders. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767, Nice, France. IEEE. 33, 37

McCarthy, P. (2023). Fsleyes. 164

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861. xxxiv, 135

Meissen, F., Kaissis, G., and Rueckert, D. (2021a). Challenging current semi-supervised anomaly segmentation methods for brain mri. In *International MICCAI brainlesion workshop*, pages 63–74. Springer. 15, 35, 36, 43, 45, 70, 114, 115

Meissen, F., Lagogiannis, I., Kaissis, G., and Rueckert, D. (2022). Domain shift as a confounding variable in unsupervised pathology detection. In *Medical Imaging with Deep Learning*. 44

Meissen, F., Wiestler, B., Kaissis, G., and Rueckert, D. (2021b). On the pitfalls of using the residual as anomaly score. In *Medical Imaging with Deep Learning*. 43, 45, 114, 115

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., and Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024. 15, 31, 33, 42, 43, 44

Mu, N. and Gilmer, J. (2019). Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*. 14

Muñoz-Ramírez, V. M., Kmetzsch, V., Forbes, F., and Dojat, M. (2020). Deep Learning Models to Study the Early Stages of Parkinson's Disease. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1534–1537, Iowa City, IA, USA. IEEE. 15, 31, 37, 44, 76, 77

Muñoz-Ramírez, Pinon, V. N., Forbes, F., Lartizen, C., and Dojat, M. (2021). Patch vs. global image-based unsupervised anomaly detection in mr brain scans of early parkinsonian patients. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 34–43. Springer. 76

Mérida, I., Jung, J., Bouvard, S., Le Bars, D., Lancelot, S., Lavenne, F., Bouillot, C., Redouté, J., Hammers, A., and Costes, N. (2021). CERMEP-IDB-MRXFDG: a database of 37 normal adult human brain [18F]FDG PET, T1 and FLAIR MRI, and CT images available for research. *EJNMMI Research*, 11(1):91. xxiv, 73, 88

Oudoumanessah, G., Dojat, M., and Forbes, F. (2022). Unsupervised scalable anomaly detection: application to medical imaging. Research report. `https://hal.archives-ouvertes.fr/hal-03824951`. 95

Oudoumanessah, G., Lartizien, C., Dojat, M., and Forbes, F. (2023). Towards frugal unsupervised detection of subtle abnormalities in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 411–421. Springer. 95

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076. 23

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011a). https://scikit-learn.org/stable/modules/outlier_detection.html. 14

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 5, 87, 100

Pernet, C., Gorgolewski, K., and Ian, W. (2016). A neuroimaging dataset of brain tumour patients. https://reshare.ukdataservice.ac.uk/851861/. 33

Pinaya, W. H., Graham, M. S., Gray, R., Da Costa, P. F., Tudosiu, P.-D., Wright, P., Mah, Y. H., MacKinnon, A. D., Teo, J. T., Jager, R., David, W., Geraint, R., Parashkev, N., Sebastien, O., and Jorge, C. (2022a). Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 705–714. Springer. 15, 33, 39, 70

Pinaya, W. H., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J. (2022b). Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475. xiv, xxiv, xxviii, xxxiv, 15, 33, 36, 57, 70, 71, 72, 74, 88, 89, 90, 91, 93, 114, 119, 123, 128, 129, 130, 131

Pinon, N., Oudoumanessah, G., Trombetta, R., Dojat, M., Forbes, F., and Lartizien, C. (2023a). Brain subtle anomaly detection based on auto-encoders latent space analysis : application to de novo parkinson patients. In *ISBI 2023 - IEEE 20th International Symposium on Biomedical Imaging*, Cartegena de Indias, Colombia. IEEE. 94

Pinon, N., Trombetta, R., and Lartizien, C. (2023b). Détection d'anomalies dans l'espace image ou l'espace latent d'auto-encodeurs par patch pour l'analyse d'images industrielles. In *GRETSI 2023, XXIXème Colloque Francophone de Traitement du Signal et des Images*. 65

Pinon, N., Trombetta, R., and Lartizien, C. (2023c). One-Class SVM on siamese neural network latent space for Unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities. In *MIDL 2023, International Conference on Medical Imaging with Deep Learning*, page 27:1783–1797. PMLR, Proceedings of Machine Learning Research. 70, 88

Poldrack, R. A., Huckins, G., and Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, 77(5):534–540. 77

Prastawa, M. (2004). A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, 8(3):275–283. 22, 29, 39

Prasuhn, J., Heldmann, M., Münte, T. F., and Brüggemann, N. (2020). A machine learning-based classification approach on parkinson's disease diffusion tensor imaging datasets. *Neurological research and practice*, 2(1):1–5. 97

Que, Z. and Lin, C.-J. (2023). One-class SVM Probabilistic Outputs. Technical report. preprint, under submission, https://www.csie.ntu.edu.tw/~cjlin/papers/oneclass_prob/oneclass_prob.pdf. 100

Quinlan, R. (1987). Thyroid Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C5D010. 14

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer. 18, 19, 32

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837. 23

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Muller, K.-R. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5):756–795. xix, 4, 5, 6, 8, 13, 14, 17, 28

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR. 25, 130

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44. 32, 36

Schuff, N., Wu, I.-W., Buckley, S., Foster, E. D., Coffey, C. S., Gitelman, D. R., Mendick, S., Seibyl, J., Simuni, T., Zhang, Y., et al. (2015). Diffusion imaging of nigral alterations in early parkinson's disease with dopaminergic deficits. *Movement Disorders*, 30(14):1885–1892. 79

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2). 29

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471. 7, 24, 47, 49, 87

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass. 49, 101, 103

Scott, C. and Nowak, R. (2005). Learning minimum volume sets. *Advances in neural information processing systems*, 18. 7

Seeböck, P., Orlando, J. I., Schlegl, T., Waldstein, S. M., Bogunović, H., Klimscha, S., Langs, G., and Schmidt-Erfurth, U. (2019). Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE transactions on medical imaging*, 39(1):87–98. 31, 38

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR. 22

Sone, D. (2021). Making the invisible visible: Advanced neuroimaging techniques in focal epilepsy. front neurosci. 2021 jul 27; 15: 699176. 51

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34. 12

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020). Osqp: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672. 162

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779. 72, 91

Tabak, E. G. and Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164. 23, 66

Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., et al. (2022). Detecting outliers with foreign patch interpolation. *Machine Learning for Biomedical Imaging*, 1(April 2022 issue):1–27. 34, 37

Tarassenko, L. (1995). Novelty detection for the identification of masses in mammograms. In *4th International Conference on Artificial Neural Networks*, volume 1995, pages 442–447, Cambridge, UK. IEE. 28, 38

Tax, D. M. and Duin, R. P. (2004). Support Vector Data Description. *Machine Learning*, 54(1):45–66. 25, 47, 101

Tian, Y., Pang, G., Liu, F., Chen, Y., Shin, S. H., Verjans, J. W., Singh, R., and Carneiro, G. (2021). Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 128–140. Springer. 32, 38

Tong, A., Wolf, G., and Krishnaswamy, S. (2022). Fixing bias in reconstruction-based anomaly detection with lipschitz discriminators. *Journal of Signal Processing Systems*, 94(2):229–243. 45

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE CVPR*, pages 6450–6459. 96

Tschuchnig, M. E. and Gadermayr, M. (2022). Anomaly detection in medical imaging-a mini review. In *Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021*, pages 33–38. Springer. 28

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30. 20, 65

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11). 135

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., and Suetens, P. (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688. 29, 39

Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media. 7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017a). Attention is all you need. *Advances in neural information processing systems*, 30. 26, 33

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017b). Attention Is All You Need. *arXiv:1706.03762 [cs]*. arXiv: 1706.03762. 65

Vert, R., Vert, J.-P., and Schölkopf, B. (2006). Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7(5). 101, 103

Wang, L., Zhang, D., Guo, J., and Han, Y. (2020). Image Anomaly Detection Using Normal Data Only by Latent Space Resampling. *Applied Sciences*, 10(23):8660. 27, 28, 33

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612. 27

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021. 164

Wyatt, J., Leach, A., Schmon, S. M., and Willcocks, C. G. (2022). AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 649–655, New Orleans, LA, USA. IEEE. 33, 36

Xiao, Y., Fonov, V., Bériault, S., Subaie, F. A., Chakravarty, M. M., Sadikot, A. F., Pike, G. B., and Collins, D. L. (2015). Multi-contrast unbiased mri atlas of a parkinson's disease population. *International journal of computer assisted radiology and surgery*, 10:329–341. 78

Yi, J. and Yoon, S. (2021). Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Computer Vision–ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30–December 4, 2020, Revised Selected Papers, Part VI 15*, pages 375–390. Springer. 28

Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. (2021). FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows. preprint, arXiv:2111.07677 [cs]. xx, 27, 30, 65, 66, 67

Zhang, Y., Brady, M., and Smith, S. (2001a). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57. 89

Zhang, Y., Brady, M., and Smith, S. (2001b). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57. 163

Zhao, R., Yaman, B., Zhang, Y., Stewart, R., Dixon, A., Knoll, F., Huang, Z., Lui, Y. W., Hansen, M. S., and Lungren, M. P. (2022a). fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 9(1):152. 15, 44

Zhao, Y., Ding, Q., and Zhang, X. (2022b). Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection. In *The Eleventh International Conference on Learning Representations*. 15, 30, 31, 39

Zhou, Y., Liang, X., Zhang, W., Zhang, L., and Song, X. (2021). Vae-based deep svdd for anomaly detection. *Neurocomputing*, 453:131–140. 130

Zimmerer, D., Full, P. M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., Jensen, B. S., O'Neil, A. Q., Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., Shvetsova, N., Fedulova, I., Dylov, D. V., Yu, B., Zhai, J., Hu, J., Si, R., Zhou, S., Wang, S., Li, X., Chen, X., Zhao, Y., Marimont, S. N., Tarroni, G., Saase, V., Maier-Hein, L., and Maier-Hein, K. (2022). MOOD 2020: A Public Benchmark for Out-of-Distribution Detection and Localization on Medical Images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738. Conference Name: IEEE Transactions on Medical Imaging. 15, 16

Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., and Maier-Hein, K. (2019). Unsupervised anomaly localization using variational auto-encoders. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 289–297. Springer. 15, 29, 31, 38

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. (2018). Deep autoencoding gaussian mixture model fore unsupervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, page 19. 130

A | Mathematical and technical details for cSVDD

A.1 Lagrangian derivation of cSVDD

We have seen in section IV.2.1.ii the primal optimization problem of cSVDD that we recall hereafter in equation A.1.

$$\begin{aligned} \min_{\mathbf{c},\mathbf{R},\underline{\underline{\boldsymbol{\xi}}}} \quad & \sum_{j=1}^{q} R_j^2 + \sum_{j=1}^{q} \frac{1}{\nu_j n} \sum_{i=1}^{n} \xi_{ji} \\ \text{subject to} \quad & ||\Phi(\mathbf{z}_i) - \mathbf{c}||^2 \leq R_j^2 + \xi_{ji} \qquad i \in [1,n] \quad j \in [1,q] \\ & \xi_{ji} \geq 0 \qquad\qquad\qquad\qquad\quad i \in [1,n] \quad j \in [1,q] \end{aligned} \tag{A.1}$$

To derive the dual of this primal problem, we write its Lagrangian. Let $\alpha_{ji} \geq 0$ and $\lambda_{ji} \geq 0$ be the Lagrange multiplier associated with the two inequality constraints, the lagrangian then writes:

$$\begin{aligned} \mathcal{L}(\boldsymbol{c}, R_j, \xi_j, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = & \sum_{j=1}^{q} R_j^2 + \sum_{j=1}^{q} \frac{1}{\nu_j n} \sum_{i=1}^{n} \xi_{ji} \\ & - \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} \left[R_j^2 + \xi_{ji} - (\varphi(\mathbf{z}_i) - \boldsymbol{c})^T (\varphi(\mathbf{z}_i) - \boldsymbol{c}) \right] \\ & - \sum_{j=1}^{q} \sum_{i=1}^{n} \lambda_{ji} \xi_{ji} \end{aligned}$$

By computing the partial derivative with respect to the primal variables we obtain:

- $\nabla_{\boldsymbol{c}} \mathcal{L}(\boldsymbol{c}, R_j, \xi_j, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = -2 \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji} (\varphi(\mathbf{z}_i) - \boldsymbol{c})$
- $\nabla_{R_j} \mathcal{L}(\boldsymbol{c}, R_j, \xi_j, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = 2R_j \times (1 - \sum_{i=1}^{n} \alpha_{ji})$
- $\nabla_{\xi_j} \mathcal{L}(\boldsymbol{c}, R_j, \xi_j, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \frac{1}{\nu_j n} \boldsymbol{e} - \boldsymbol{\alpha}_j - \boldsymbol{\lambda}_j$

with $\boldsymbol{e}^T = (1, \ldots, 1)^T \in \mathbb{R}^n$, $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jn})^T \in \mathbb{R}^n$ et $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jn})^T \in \mathbb{R}^n$
At the optimum, the Karush-Kuhn-Tucker gives:

- $-2 \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji}^* (\varphi(\mathbf{z}_i) - \boldsymbol{c}) = 0 \Rightarrow \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji}^* \varphi(\mathbf{z}_i) = \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji}^* \boldsymbol{c} \Rightarrow \sum_{j=1}^{q} \sum_{i=1}^{n} \alpha_{ji}^* \varphi(\mathbf{z}_i) = q\boldsymbol{c}$
- $2R_j \times \left(1 - \sum_{i=1}^{n} \alpha_{ji}^*\right) = 0 \Rightarrow \sum_{i=1}^{n} \alpha_{ji}^* = 1$ with $j = 1, \ldots, q$
- $\frac{1}{\nu_j n} \boldsymbol{e} - \boldsymbol{\alpha}_j^* - \boldsymbol{\lambda}_j^* = 0$ with $j = 1, \ldots, q$

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{c}, R_j, \xi_j, \boldsymbol{\alpha}, \boldsymbol{\lambda}) &= \sum_{j=1}^{q} R_j^2 + \sum_{j=1}^{q} \frac{1}{\nu_j n} \sum_{i=1}^{n} \xi_{ji} \\
&\quad - \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} \left[R_j^2 + \xi_{ji} - \left(\varphi(\mathbf{z}_i) - \boldsymbol{c}\right)^T \left(\varphi(\mathbf{z}_i) - \boldsymbol{c}\right) \right] - \sum_{j=1}^{q}\sum_{i=1}^{n} \lambda_{ji}\xi_{ji} \\
&= \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_i) > + \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} < \mathbf{a}, \mathbf{a} > -2 \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji}\alpha_{ji} < \mathbf{a}, \varphi(\mathbf{z}_i) > \\
&= \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_i) > + \frac{1}{q} \sum_{j,k=1}^{q}\sum_{i,l=1}^{n} \alpha_{ji}\alpha_{kl} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_l) > \\
&\quad - \frac{2}{q} \sum_{j,k=1}^{q}\sum_{i,l=1}^{n} \alpha_{ji}\alpha_{kl} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_l) > \\
&= \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_i) > - \frac{1}{q} \sum_{j,k=1}^{q}\sum_{i,l=1}^{n} \alpha_{ji}\alpha_{kl} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_l) > \\
&= \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} k\left(\mathbf{z}_i, \mathbf{z}_i\right) - \frac{1}{q} \sum_{j,k=1}^{q}\sum_{i,l=1}^{n} \alpha_{ji}\alpha_{kl} k\left(\mathbf{z}_i, \mathbf{z}_l\right)
\end{aligned}
$$

Furthermore, with the equation $\boldsymbol{\alpha}_j = \frac{1}{\nu_j n}\boldsymbol{e} - \boldsymbol{\lambda}_j$ and the fact that $\alpha_{ji} \geq 0$, $\lambda_{ji} \geq 0$, Lagrange multipliers λ_{ji} can be removed if we impose that $0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n}$.

Finally, the dual problem writes:

$$
\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{q} \sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_{ji}\alpha_{kl} k(\mathbf{z}_i, \mathbf{z}_l) - \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} k(\mathbf{z}_i, \mathbf{z}_i) \\
\text{subject to} \quad & 0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n} && j \in [1, q] \quad i \in [1, n] \qquad \text{(A.2)} \\
& \sum_{i=1}^{n} \alpha_{ji} = 1 && j \in [1, q]
\end{aligned}
$$

This problem corresponds to a quadratic program and can thus be solved efficiently with any state-of-the-art solver.

Additionally, a new sample $\mathbf{z}$ belongs to the support of the j^{th} SVDD if:

$$
f_{\mathbf{c}}(\mathbf{z}) = \left(\varphi(\mathbf{z}) - \boldsymbol{c}\right)^T \left(\varphi(\mathbf{z}) - \boldsymbol{c}\right) \leq R_j^2
$$

we developp:

$$
\begin{aligned}
f_{\mathbf{c}}(\mathbf{z}) &= \\
\left(\varphi(\mathbf{z}) - \boldsymbol{c}\right)^T \left(\varphi(\mathbf{z}) - \boldsymbol{c}\right) &= \varphi(\mathbf{z})^T\varphi(\mathbf{z}) - 2\mathbf{a}\varphi(\mathbf{z}) + \mathbf{a}^T\mathbf{a} \\
&= k(\mathbf{z}, \mathbf{z}) - \frac{2}{q} \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}) > + \frac{1}{q^2} \sum_{j,k=1}^{q}\sum_{i,l=1}^{n} \alpha_{ji}\alpha_{kl} < \varphi(\mathbf{z}_i), \varphi(\mathbf{z}_l) > \\
&= k(\mathbf{z}, \mathbf{z}) - \frac{2}{q} \sum_{j=1}^{q}\sum_{i=1}^{n} \alpha_{ji} k(\mathbf{z}_i, \mathbf{z}) + \frac{1}{q^2} \sum_{j,k=1}^{q}\sum_{i,l=1}^{n} \alpha_{ji}\alpha_{kl} k(\mathbf{z}_i, \mathbf{z}_l)
\end{aligned}
$$

which means $\mathbf{z}$ belongs to the support of the j^{th} SVDD if:

$$f_{\mathbf{c}}(\mathbf{z}) = k(\mathbf{z},\mathbf{z}) - \frac{2}{q}\sum_{j=1}^{q}\sum_{i=1}^{n}\alpha_{ji}k(\mathbf{z}_i,\mathbf{z}) + \frac{1}{q^2}\sum_{j,k=1}^{q}\sum_{i,l=1}^{n}\alpha_{ji}\alpha_{kl}k(\mathbf{z}_i,\mathbf{z}_l) \leq R_j^2$$

We can formulate the problem A.2 in vector/matrix manner, using:

- $\underline{\underline{\boldsymbol{\alpha}}} = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{q1} & \dots & \alpha_{qn} \end{pmatrix} \in \mathbb{R}^{q\times n}$
- $\boldsymbol{\alpha_j}^T = (\alpha_{j1},\dots,\alpha_{jn})^T \in \mathbb{R}^n$
- $\underline{\underline{\mathbf{K}}} \in \mathbb{R}^{n\times n}$ where $K_{il} = k(\mathbf{z_i},\mathbf{z_l})$
- $\mathbf{f} = diag(\underline{\underline{\mathbf{K}}}) \in \mathbb{R}^n$

the problem thus writes:

$$\begin{aligned} \min_{\underline{\underline{\boldsymbol{\alpha}}}} \quad & \frac{1}{q}\,\underline{\underline{\boldsymbol{\alpha}}}^T\underline{\underline{\mathbf{K}}}\,\underline{\underline{\boldsymbol{\alpha}}}\boldsymbol{e} - \boldsymbol{f}\underline{\underline{\boldsymbol{\alpha}}}\boldsymbol{e} \\ \text{subject to} \quad & 0 \leq \boldsymbol{\alpha_j} \leq \frac{1}{\nu_j n}\boldsymbol{e} \qquad j \in [1,q] \\ & \underline{\underline{\boldsymbol{\alpha}}}\boldsymbol{e} = \boldsymbol{e} \end{aligned} \tag{A.3}$$

A.2 Implementation and technical details

For solver-related manner, the problem A.3 has to be written in a way that it is linear in parameters, and not quadratic as we can see in equation A.2 (i.e. K_{il} would be multiplied by some $K_{i'l'}$). We thus need to utilize the fact that $\underline{\underline{\mathbf{K}}}$ is positive semi-definite (because it is a gram matrix), to express it as: $\underline{\underline{\mathbf{K}}} = \underline{\underline{\mathbf{K}}}^{\frac{1}{2}T}\underline{\underline{\mathbf{K}}}^{\frac{1}{2}}$.

Also, as $\frac{1}{\nu_j n}$ can get very small as n increase, this only leaves a tight bound for the constraint $0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n}$. Thus, for numerical stability reasons, we solve a scaled problem of variable $\tilde{\alpha}_{ji} = n\nu_j\alpha_{ji}$, which will scale the constraint to a more appropriate range.

With these two modifications, the problem A.3 will be solved as:

$$\begin{aligned} \min_{\underline{\underline{\tilde{\boldsymbol{\alpha}}}}} \quad & \frac{1}{q}\,\underline{\underline{\tilde{\boldsymbol{\alpha}}}}^T\underline{\underline{\mathbf{K}}}^{\frac{1}{2}T}\underline{\underline{\mathbf{K}}}^{\frac{1}{2}}\underline{\underline{\tilde{\boldsymbol{\alpha}}}}\boldsymbol{e} - \boldsymbol{f}\underline{\underline{\tilde{\boldsymbol{\alpha}}}}\boldsymbol{e} \\ \text{subject to} \quad & \underline{\mathbf{0}} \leq \underline{\underline{\tilde{\boldsymbol{\alpha}}}} \leq \underline{\mathbf{1}} \qquad \text{(element-wise)} \\ & \tilde{\boldsymbol{\alpha}}_j^T\boldsymbol{e} = \nu_j n \qquad j \in [1,q] \end{aligned} \tag{A.4}$$

and α_{ji} recovered as $\alpha_{ji} = \frac{1}{n\nu_j}\tilde{\alpha}_{ji}$. For numerical stability, we computed $\underline{\underline{\mathbf{K}}}$ as $\underline{\underline{\mathbf{K}}} + 1e^{-8}I$. We used the OSQP solver Stellato et al. (2020).

B | MRI volumes pre/post-processing

B.1 Pre-processing of the MRI volumes

Preprocessing of the T1w MR images was performed based on the reference methods implemented in SPM12. The spatial normalization was performed using the unified segmentation algorithm (UniSeg) Ashburner and Friston (2005) which includes 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 to the standard brain template of the Montreal Neurological Institute (MNI). In this work, we used the default parameters for normalization and a voxel size of $1 \times 1 \times 1$ mm. Next, FLAIR image of each subject was rigidly co-registered to its corresponding individual T1w MR image in the native space and then transformed to the MNI space by applying the transformation field produced by the UniSeg algorithm on the T1w image.

The cerebellum and brain stem were excluded from the spatially normalized images. The masking image in the reference MNI space was derived from the Hammersmith maximum probability atlas (Hammers et al. 2003).

On top of that, each image X was intensity-normalized into X_{norm} with:
$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$.

B.2 Segmentation of the cerebrospinal fluid (CSF)

We used the FMRIB's Automated Segmentation Tool (FAST) by Zhang et al. (2001b) to segment the grey and white matter, allowing us to exclude the CSF from the anomaly maps, as we found a high number of false positive in our method belong in these regions.
FAST is here used to provide two CSF segmentation maps, one based on the T1 image and the second based on the T1 and FLAIR images. The union of the two segmentations, after being masked by a gross brain segmentation to remove the skull, is then processed with some basic mathematical morphology operators, namely : two dilatations followed by two erosions, using a basic cross-shaped structuring element of width 1 voxel. A last erosion on the convex hull of the segmentation is performed to remove a thin outer border of the cortex. Note that this whole processing is done in 3D.

C | Visualization software acknowledgments

Throughout this manuscript, written with LaTeX, a combination of Matplotlib (Hunter 2007) and Seaborn (Waskom 2021) was used to generates plots, graphs, and generally every non-image data visualization. Draw.io (JGraph Ltd 2021) was used to generate the diagrams. FSLeyes (McCarthy 2023) was used for medical images visualization.

D | Computational complexity of *loc*OC-SVM and *ps*OC-SVM

We claimed in IV.1.4 that the *ps*OC-SVM algorithm was faster than the *loc*OC-SVM algorithm, with the intuitive fact that we train as many SVM as there is patients for *ps*OC-SVM and as many as there is brain localization for *loc*OC-SVM. We derive the exact computational complexity hereafter, with:

- M the number of brain localizations (1.5M in our experiments)
- N the number of controls (75 in our experiments)
- N' the number of patients (60 in our experiments)
- n the number of patches sampled for *ps*OC-SVM (500 in our experiments)
- d the dimension of the input vectors (16 in our experiments)

With p samples, SVM training complexity of $O(p^2 \times d)$ and SVM inference complexity of $O(p \times d)$, assuming all vectors are support vectors (worst case scenario, i.e. upper bound on the complexity), we get the following:

_loc_OC-SVM

- Training: M OC-SVM per voxel with N sample: $M \times O(N^2 \times d)$
- Inference: M OC-SVM per voxel with N' samples: $M \times O(N' \times d)$
- Total: $O((N^2 + N') \times M \times d)$

With approximation $N^2 >> N' => O(N^2 \times M \times d)$ (meaning inference negligible compared to training)

_ps_OC-SVM

- Training: N' OC-SVM with n samples: $N' \times O(n^2 \times d)$
- Inference: N' OC-SVM with M samples: $N' \times O(M \times d)$
- Total: $O((n^2 + M) \times N' \times d)$

With approximation $n^2 << M => O(M \times N' \times d)$ (meaning training negligible compared to inference)

_loc_OC-SVM VS _ps_OC-SVM

Approximated ratio *loc*OC-SVM/*ps*OC-SVM: $\frac{O(N^2 \times M \times d)}{O(M \times N' \times d)} = O(N^2/N')$
If we consider $N' \simeq N$ (number of controls same magnitude as number of patients):
_ps_OC-SVM $N \times$ **faster than _loc_OC-SVM** with N the number of control/patient.

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment INSA direction, 1er étage
37, av. J. Capelle
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH1091_PINON Nicolas

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : PINON DATE de SOUTENANCE : 11/04/2021

Prénoms : Nicolas, Eric, Olivier

TITRE (EN) : *Unsupervised anomaly detection in neuroimaging: contributions to representation learning and density support estimation in the latent space.*

TITRE (FR) : *Détection d'anomalies non supervisée en neuro-imagerie : contributions à l'apprentissage de représentation et à l'estimation de support de densité dans l'espace latent.*

NATURE : Doctorat Numéro d'ordre : 2024ISAL0031

Ecole doctorale : ED EEA (160)

Spécialité : Traitement du Signal et de l'Image

RESUME (EN) : This PhD thesis covers the topic of deep unsupervised anomaly detection (UAD) in neuroimaging. This research is partially grounded on the UAD model that was proposed in [Alaverdyan, MEDIA 2020], whose novelty was to perform the detection step in the latent representation space by adjusting density support model of the normative distribution. This model developed was applied to the detection of subtle (MRI negative) epileptogenic zones in multiparametric MRI and evaluated on a private database. As a first part of this PhD, we optimize the architecture and hyperparameter setting of this UAD model, and evaluate its performance on different open datasets, including the non medical MVTec anomaly detection [Pinon, GRETSI 2023], the WMH challenge, and the Parkinson's Progression Markers Initiative database [Ramirez, Pinon, MLCN 2021][Pinon, ISBI 2023]. This allows comparison with state of the art UAD methods, especially with the most common methods based on reconstruction error in the image space. The second main phase of this PhD work is to build on the limits of this model [Alaverdyan, MEDIA 2020] and propose original methodological contributions to 1) design patient specific models, relaxing the strong constraint to accurately coregister all control subjects and patients [Pinon, MIDL 2023], 2) provide a probabilistic detection framework to enable ensemble learning and score map uniformization, 3) fuse the representation learning step and the outlier detection step, by proposing a novel deep learning model.

RESUME (FR) : Cette thèse de doctorat porte sur la détection d'anomalies non supervisée (UAD) en neuro-imagerie. Ce travaille s'appuie en partie sur le modèle proposé dans [Alaverdyan, MEDIA 2020], dont la nouveauté est d'effectuer l'étape de détection dans l'espace de représentation latent en estimant le support de densité de probabilité de la distribution normative. Ce modèle développé a été appliqué à la détection de zones épileptogènes subtiles (IRM négative) en IRM multiparamétrique et évalué sur une base de données privée. Dans la première partie de cette thèse, nous optimisons l'architecture et les paramètres de ce modèle UAD, et évaluons ses performances sur différents jeux de données publics, y compris pour la la détection d'anomalies non médicales sur MVTecAD [Pinon, GRETSI 2023], la base de challenge WMH, et la base de données Parkinson's Progression Markers Initiative [Ramirez, Pinon, MLCN 2021][Pinon, ISBI 2023]. Ceci permet une comparaison avec l'état de l'art des méthodes d'UAD , en particulier avec les méthodes les plus courantes basées sur l'erreur de reconstruction dans l'espace image. La deuxième partie de ce travail de thèse est de s'appuyer sur les limites de ce modèle [Alaverdyan, MEDIA 2020] et de proposer des contributions méthodologiques originales pour 1) concevoir des modèles patient-spécifiques, en relâchant la contrainte forte de recaler avec précision tous les controles et patients [Pinon, MIDL 2023], 2) fournir une calibration en probabilité pour permettre l'apprentissage d'ensemble et l'uniformisation de cartes de score, 3) fusionner l'étape d'apprentissage de représentation et l'étape de détection d'anomalie, en proposant un nouveau modèle bout à bout.

MOTS-CLÉS (EN) : Anomaly detection, Neuroimaging, Unsupervised learning, MRI

MOTS-CLÉS (FR) : Détection d'anomalie, Neuro-imagerie, Apprentissage non-supervisé, IRM

Laboratoire (s) de recherche : CREATIS – UMR 5220 et U1206 – Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé

Directrice de thèse: Carole Lartizien

Président de jury : Fabrice Meriaudeau

Composition du jury : Julia Schnabel (Rapportrice), Isabelle Bloch (Rapportrice), Fabrice Meriaudeau (Examinateur), Nicolas Duchateau (Examinateur), Florrence Forbes (Invitée), Carole Lartizien (Directrice de thèse)