



HAL
open science

Individualised, interpretable and reproducible computer-aided diagnosis of dementia: towards application in clinical practice

Ninon Burgos

► **To cite this version:**

Ninon Burgos. Individualised, interpretable and reproducible computer-aided diagnosis of dementia: towards application in clinical practice. Medical Imaging. Sorbonne Université, 2022. tel-03941953

HAL Id: tel-03941953

<https://inria.hal.science/tel-03941953v1>

Submitted on 16 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HABILITATION À DIRIGER DES RECHERCHES

DE

SORBONNE UNIVERSITÉ

Individualised, interpretable and reproducible computer-aided diagnosis of dementia: towards application in clinical practice

by

Ninon Burgos

Soutenue le 12 décembre 2022 devant le jury composé de :

Nicholas Ayache	Directeur de recherche, Inria	Examineur
Isabelle Bloch	Professeure, Sorbonne Université	Examinatrice (Présidente du jury)
Olivier Colliot	Directeur de recherche, CNRS	Examineur
Michel Dojat	Directeur de recherche, Inserm	Rapporteur
Carole Lartizien	Directrice de recherche CNRS	Rapportrice
Sebastien Ourselin	Professeur, King's College London	Examineur
Julia Schnabel	Professeure, Helmholtz Munich	Rapportrice

Abstract

Neuroimaging offers an unmatched description of the brain's structure and physiology, but the information it provides is not easy to extract and interpret. A popular way to extract meaningful information from brain images is to use computational methods based on machine learning and deep learning to predict the current or future diagnosis of a patient. A large number of these approaches have been dedicated to the computer-aided diagnosis of dementia, and more specifically of Alzheimer's disease. However, only a few are translated to the clinic. This can be explained by different factors such as the lack of rigorous validation of these approaches leading to over-optimistic performance and their lack of reproducibility, but also the limited interpretability of these methods and their limited generalisability when moving from highly controlled research data to routine clinical data. This manuscript describes how we tried to address these limitations.

We have proposed reproducible frameworks for the evaluation of Alzheimer's disease classification methods and developed two open-source software platforms for clinical neuroimaging studies (Clinica) and neuroimaging processing with deep learning (ClinicaDL). We have implemented and assessed the robustness of a visualisation method aiming to interpret convolutional neural networks and used it to study the stability of the network training. We concluded that, currently, combining a convolutional neural networks classifier with an interpretability method may not constitute a robust tool for individual computer-aided diagnosis. As an alternative, we have proposed an approach that detects anomalies in the brain by generating what would be the healthy version of a patient's image and comparing this healthy version with the real image. Finally, we have studied the performance of machine and deep learning algorithms for the computer-aided diagnosis of dementia from images acquired in clinical routine.

Résumé

La neuro-imagerie offre une description inégalée de la structure et de la physiologie du cerveau, mais les informations qu'elle fournit ne sont pas faciles à extraire et à interpréter. Une façon populaire d'extraire des informations pertinentes d'images cérébrales consiste à utiliser des méthodes basées sur l'apprentissage statistique et l'apprentissage profond pour prédire le diagnostic actuel ou futur d'un patient. Un grand nombre de ces approches ont été dédiées au diagnostic assisté par ordinateur de la démence, et plus spécifiquement de la maladie d'Alzheimer. Cependant, seules quelques-unes sont transposées en clinique. Cela peut s'expliquer par différents facteurs tels que l'absence de validation rigoureuse de ces approches conduisant à des performances trop optimistes et à leur manque de reproductibilité, mais aussi l'interprétabilité limitée de ces méthodes et leur généralisation limitée lors du passage de données de recherche hautement contrôlées à des données cliniques de routine. Ce manuscrit décrit comment nous avons tenté de remédier à ces limites.

Nous avons proposé des cadres reproductibles pour l'évaluation des méthodes de classification de la maladie d'Alzheimer et développé deux plateformes logicielles open-source pour les études de neuroimagerie clinique (Clinica) et le traitement de la neuroimagerie par apprentissage profond (ClinicaDL). Nous avons implémenté et évalué la robustesse d'une méthode de visualisation visant à interpréter les réseaux neuronaux convolutifs et l'avons utilisée pour étudier la stabilité de l'entraînement du réseau. Nous avons conclu qu'actuellement, la combinaison de réseaux neuronaux convolutifs avec une méthode d'interprétabilité peut ne pas constituer un outil robuste pour le diagnostic individuel assisté par ordinateur. De façon alternative, nous avons proposé une approche qui détecte les anomalies dans le cerveau en générant ce qui serait la version saine de l'image d'un patient et en comparant cette version saine avec l'image réelle. Enfin, nous avons étudié les performances des algorithmes d'apprentissage statistique et profond pour le diagnostic assisté par ordinateur de la démence à partir d'images acquises en routine clinique.

Contents

Abstract	iii
Résumé	v
Contents	vii
List of Figures	xi
List of Tables	xiii
Introduction	1
Neuroimaging for brain disorders	1
Computer-aided diagnosis of dementia	4
Classification and prediction of dementia with machine learning	4
Detection of brain anomalies	6
1 Anomaly detection for the computer-aided diagnosis of dementia	7
1.1 Introduction	7
1.2 Methods	9
1.2.1 Database of control atlases	9
1.2.2 Atlas pre-selection	9
1.2.3 Inter-subject registration and atlas local selection	10
1.2.4 Subject-specific models of healthy PET appearance	11
1.2.5 Subject-specific abnormality maps	11
1.3 Validation on Alzheimer’s disease	12
1.3.1 Materials: Alzheimer’s disease cohort	12
1.3.2 Parameter optimisation	13
1.3.3 Validation scheme	14
1.3.4 Results	17
1.4 Application to frontotemporal dementia	18
1.4.1 Materials: frontotemporal dementia cohort	18
1.4.2 Results	19
1.5 Discussion	23
1.6 Perspectives	24
1.6.1 Anomaly detection using autoencoders	25
1.6.2 Anomaly detection framework using conditional generative adversarial networks	26
1.6.3 Uncertainty estimation of the reconstruction	27

2	Interpretable computer-aided diagnosis of dementia	29
2.1	Introduction	29
2.2	Materials and methods	31
2.2.1	Data description and preprocessing	31
2.2.2	CNN classification	31
2.2.3	Interpretability method	33
2.2.4	Metrics of evaluation	35
2.3	Results	36
2.3.1	Grid search on interpretability hyperparameters	36
2.3.2	Robustness of the interpretability method	37
2.3.3	Robustness of the CNN training	40
2.4	Discussion	41
2.5	Perspectives	42
3	Reproducible computer-aided diagnosis of dementia	43
3.1	Machine learning and deep learning for the diagnosis of Alzheimer’s disease	44
3.2	Materials	45
3.2.1	Data sets	45
3.2.2	Conversion to the Brain Imaging Data Structure	46
3.3	AD-ML: Framework for the reproducible evaluation of machine learning clas- sification experiments	47
3.3.1	Methods	48
3.3.2	Results	50
3.3.3	Discussion	55
3.4	AD-DL: Framework for the reproducible evaluation of deep learning classifi- cation experiments	58
3.4.1	Methods	59
3.4.2	Results of the cross-validation experiments	64
3.4.3	Results on the test sets	68
3.4.4	Discussion	71
3.5	Open-source contributions	73
3.5.1	Clinica: Software platform for neuroimaging studies	74
3.5.2	ClinicaDL: Software for reproducible neuroimaging processing with deep learning	77
3.6	Conclusion and perspectives	82
4	Computer-aided diagnosis of dementia from routine clinical data	83
4.1	Introduction	84
4.2	AP-HP clinical data warehouse	85
4.2.1	Data set description	85
4.2.2	Image preprocessing	86
4.3	Quality control of T1-weighted brain MRI from a clinical data warehouse	88
4.3.1	Manual labelling of the data set	89
4.3.2	Automatic quality control	93

4.3.3	Conclusion	95
4.4	Contrast-enhanced to non-contrast-enhanced image translation to exploit a clinical data warehouse of T1-weighted brain MRI	97
4.4.1	Data set description	98
4.4.2	Network architecture	98
4.4.3	Experiments and validation measures	101
4.4.4	Results	103
4.4.5	Conclusion	107
4.5	Computer-aided diagnosis of dementia in a clinical data warehouse	108
4.5.1	Materials	108
4.5.2	Methods	114
4.5.3	Results	115
4.5.4	Discussion	119
4.6	Perspectives	121
	General conclusions and perspectives	123
	Anomaly detection for the computer-aided diagnosis of dementia	123
	Interpretable computer-aided diagnosis of dementia	123
	Reproducible computer-aided diagnosis of dementia	124
	Computer-aided diagnosis of dementia from routine clinical data	124
	CV and publications	127
	A Scopus and PubMed database queries	143
	B Data access	147
	C APPRIMAGE Study Group	149
	Bibliography	151

List of Figures

1	Timeline of the main developments in neuroimaging	2
2	Distribution by imaging modality and brain disorder of 1327 articles presenting a using with machine learning.	3
3	Number of articles presenting machine learning and deep learning approaches for the computer-aided diagnosis of Alzheimer’s disease published over the years according to PubMed (query available in Appendix A).	5
1.1	Subject-specific anomaly detection framework	9
1.2	Average rMAE between the real and pseudo-healthy PET images in the grey matter and whiter matter regions for varying values of σ_G and β	14
1.3	Examples of FDG and Florbetapir abnormality maps obtained for a cognitively normal subject and patients with early mild cognitive impairment, late mild cognitive impairment and Alzheimer’s disease	16
1.4	Boxplots of the rMAE between the real and pseudo-healthy PET images for the CN amyloid negative subjects	17
1.5	NIFD participants’ demographics and clinical scores, and average Z-scores obtained with the proposed method for each subject and ROI	21
1.6	Examples of abnormality maps obtained for two subjects with bvFTD and a subject with svPPA	22
1.7	Anomaly detection framework using autoencoders	26
1.8	Anomaly detection framework using conditional generative adversarial networks (cGANs)	26
2.1	Architecture of the CNN classifier determined following to a random search procedure	33
2.2	Comparison of masks obtained for different values of the interpretability hyperparameters β_1 and β_2	36
2.3	Comparison of masks obtained for different values of the interpretability hyperparameters λ_1 and λ_2	37
2.4	Coronal view of the group masks trained on ADNI and AIBL	38
2.5	Coronal view of the group masks obtained for the five folds of the cross-validation on the first run and of the group masks obtained for five runs of the first fold	40
3.1	AD-ML: Influence of atlas	51
3.2	AD-ML: Influence of smoothing	52
3.3	AD-ML: Influence of classification method	53

3.4	AD-ML: Influence of class imbalance	54
3.5	AD-ML: Influence of training data set size	56
3.6	AD-ML: Influence of training set size when combining data sets	56
3.7	AD-DL: Architecture of the 3D subject-level CNN.	60
3.8	AD-DL: Architecture of the 3D ROI-based and 3D patch-level CNNs.	61
3.9	AD-DL: Architecture of the 2D slice-level CNN.	62
3.10	Overview of Clinica’s functionalities	75
3.11	List of the pipelines currently available in Clinica with their dependencies and outputs	76
3.12	ClinicaDL main functionalities	79
4.1	Examples of T1w brain images from the clinical data warehouse and the corresponding labels	88
4.2	General workflow of the proposed QC framework	89
4.3	Distribution of the consensus labels for the whole data set of 5500 images	91
4.4	Learning curves for the SR, gadolinium injection, tier 3 vs tier 2-1 and tier 2 vs tier 1 tasks	95
4.5	Architectures of the proposed 3D U-Net like models	99
4.6	Examples of real T1w-ce, real T1w-nce and synthetic T1w-nce obtained with the <i>cGAN Att-U-Net</i> model	103
4.7	Example of the probability grey matter maps obtained from T1w-ce, T1w-nce and synthetic T1w-nce	105
4.8	Volume differences between T1w-nce and T1w-ce images and between T1w- nce and synthetic T1w-nce images	106
4.9	Workflow describing the selection of patients belonging to the dementia category	111

List of Tables

1.1	Summary of the ADNI participants' demographics, clinical scores and amyloid status.	12
1.2	Balanced accuracy obtained when using PET images, state-of-the-art Z-maps, and the subject-specific abnormality maps as features of the linear SVM classification algorithm	18
2.1	Summary of ADNI and AIBL participant demographics, MMSE and global CDR scores at baseline	31
2.2	Similarity across different β_1 and β_2 values	39
2.3	Similarity across different λ_1 and λ_2 values	40
3.1	Summary of participant demographics and cognitive scores at baseline for ADNI, AIBL and OASIS.	47
3.2	AD-ML: Influence of feature types	52
3.3	AD-ML: Influence of data set	55
3.4	AD-DL: Results of the cross-validation experiments	65
3.5	AD-DL: Results on three independent test sets	69
4.1	Model name of all the scanners with the corresponding magnetic field strength and the number of images	87
4.2	Description and determination rules of the proposed quality control tiers	90
4.3	Weighted Cohen's kappa between the two annotators	91
4.4	Distribution of the manufacturers, field strength, sex and age according to QC grading (performed by the human raters) and on the overall population	92
4.5	Results of the CNN classifier for all the QC tasks	94
4.6	Results of three 3D CNN architectures (Conv5_FC3, Inception and ResNet) for the rating of the overall image quality	96
4.7	MAE, PSNR and SSIM obtained on the two independent test sets with various image quality	104
4.8	Absolute volume difference between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images	105
4.9	Dice scores obtained when comparing the grey matter, white matter and cerebrospinal fluid segmentations between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images	106
4.10	Description of the three categories of interest with the corresponding ICD-10 codes	110
4.11	Characteristics of the three classes of interest (D, ND and NDNL)	113

4.12	Dementia classification performance (AD vs CN) in a research data set . . .	115
4.13	Dementia classification performance (D vs NDNL and D vs NDL) in the clinical data set	116
4.14	Influence of gadolinium injection and image quality on the classification performance	117
4.15	Joint influence of gadolinium injection and image quality on the classification performance	117
4.16	Classification performance obtained after gadolinium removal using image translation, training on a set of 88 patients	118
4.17	Classification performance obtained after gadolinium removal using image translation, training on a set of 181 patients	118
4.18	Classification performance when training on a research data set and testing on a clinical data set	119

Introduction

Neuroimaging for brain disorders

Medical imaging plays an important role in the detection, diagnosis and treatment monitoring of brain disorders. Neuroimaging includes different modalities such as X-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) or single-photon emission computed tomography (SPECT).

Most neuroimaging modalities have been developed in the 1970s (Figure 1). The first CT image of a brain was acquired in 1971 (Ambrose, 1973; Hounsfield, 1973). This technology results from the discovery of X-rays by Wilhelm Röntgen in 1895 (Röntgen, 1896). A few years later, PET (Ter-Pogossian et al., 1975) and then SPECT (Jaszczak et al., 1977; Keyes et al., 1977) cameras were developed. Both modalities result from the discovery of natural radioactivity in 1896 by Henri Becquerel (Becquerel, 1903). The first MR image of a brain goes back to 1978 (Young et al., 1982) following the discovery of nuclear magnetic resonance in 1946 by Felix Bloch (Bloch, 1946). Some of these imaging modalities were later combined into hybrid scanners. The first prototype combining PET and CT was introduced into the clinical arena in 1998 (Townsend et al., 2003) while the first PET and MR images of a brain simultaneously acquired were reported in 2007 (Schlemmer et al., 2007; Schmand et al., 2007). The first commercial SPECT/CT system dates back to 1999 (Patton et al., 2000) while SPECT/MR systems are still under development (Hutton et al., 2018).

CT and MRI are the modalities of choice when studying brain anatomy while SPECT and PET are used to image particular biological processes. Note that MRI is a versatile modality that allows studying both brain structure and function, through the use of different sequences. The use of these imaging modalities differs between clinical practice and research contexts. For example, CT will be the main modality used in hospitals on adults (Smith-Bindman et al., 2019) while MRI is by far the modality that is the most used for the study of brain disorders with machine learning (Figure 2, top). The two disorders the most studied in a research context with machine learning are brain tumours and dementia, mainly Alzheimer's disease (Figure 2, bottom).

Alzheimer Europe, 2013 estimates the percentage of people with dementia in France in 2012 as being 1.85%. This percentage goes up to 23.53% in the population aged over 80 years. The prevalence of dementia keeps increasing: in Western Europe in 2015, 7.45 million people were affected by dementia and this number is expected to double in 2050, reaching 14.32 million (Alzheimer's Disease International, 2015). A recent study in the US focused on a population of 3.1 million individuals who had a claim for a service and/or treatment for any dementia subtype (Goodman et al., 2017). The most common dementia subtype was Alzheimer's (43.5%), followed by vascular (14.5%), Lewy body (5.4%), and frontotemporal

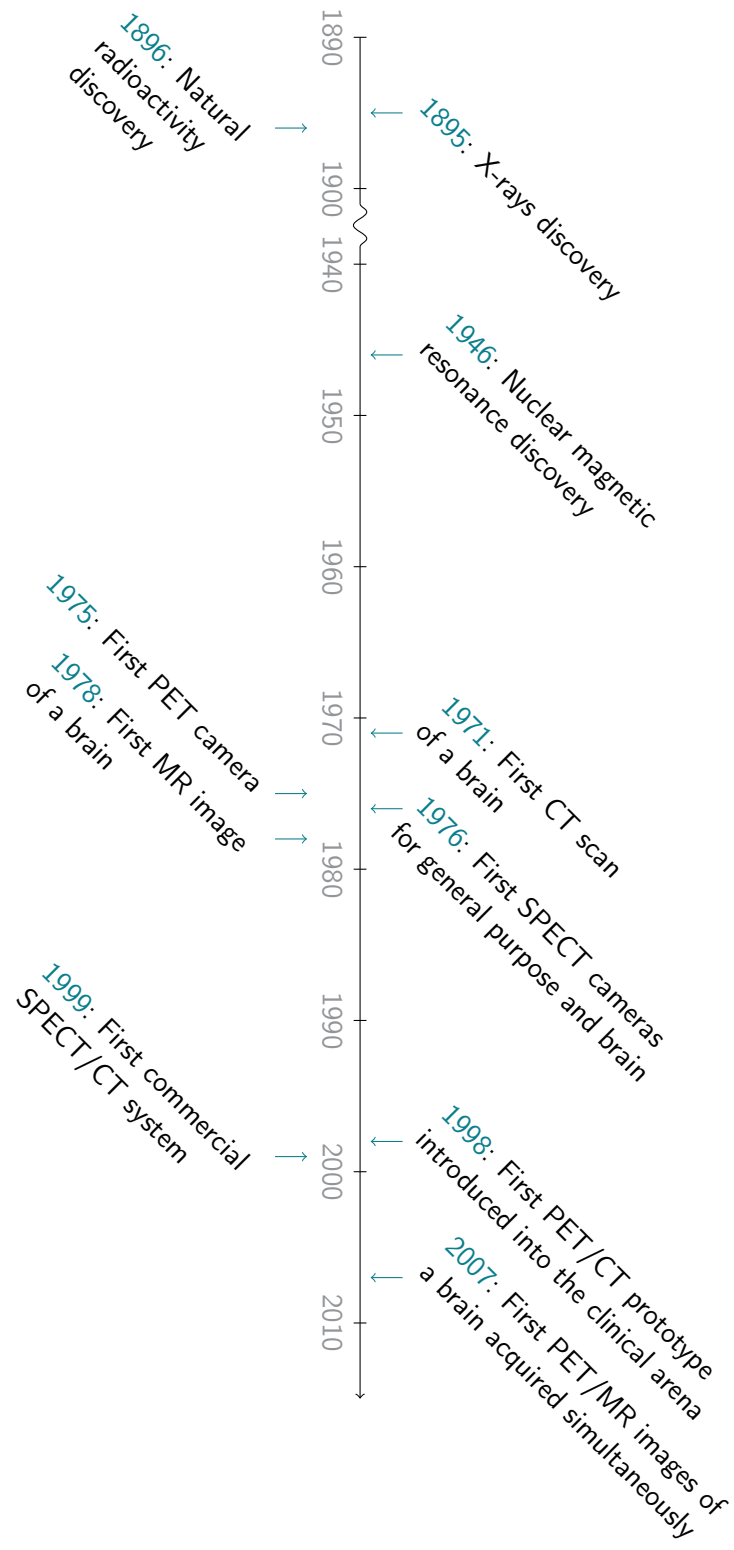


FIGURE 1: Timeline of the main developments in neuroimaging

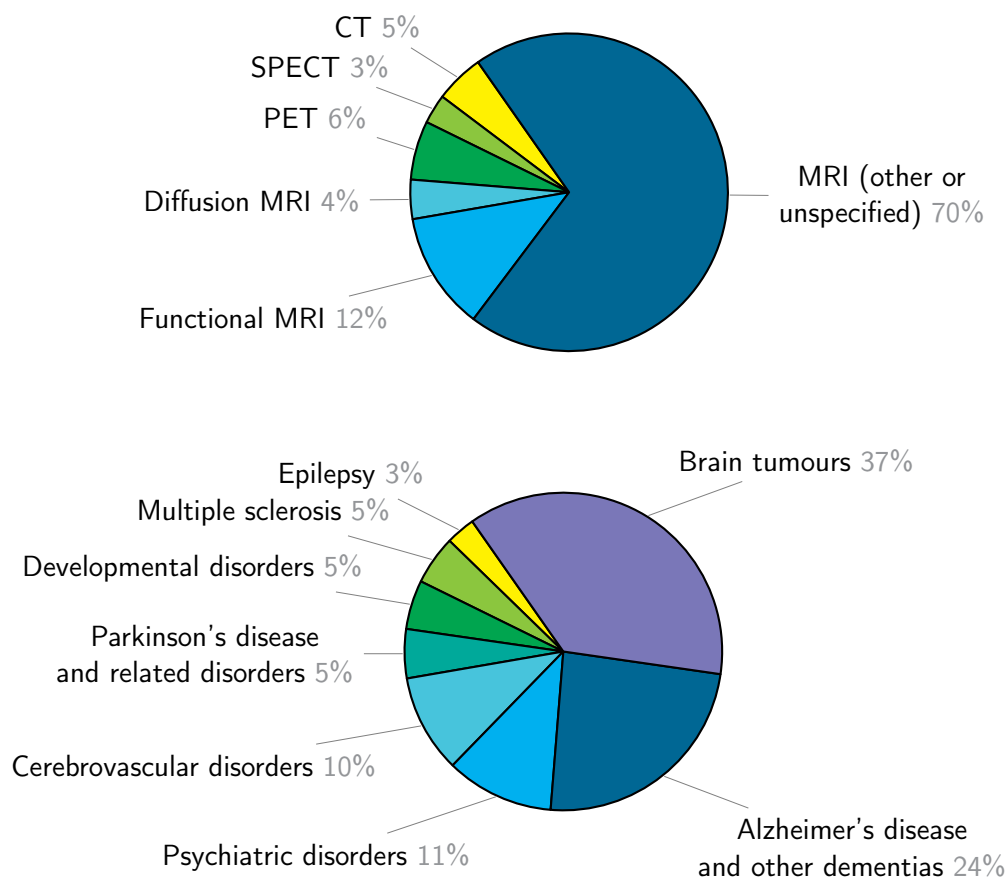


FIGURE 2: Distribution by imaging modality (top) and brain disorder (bottom) of 1327 articles presenting a study using machine learning. Note that these numbers should only be taken as rough indicators as they result from a non-exhaustive literature search. The resulting articles (after some manual filtering) are available as a public Zotero library (https://www.zotero.org/groups/4623150/neuroimaging_with_ml_for_brain_disorders/library) and the Scopus query appears in Appendix A.

(1.0%). However, the most common diagnosis (multiple diagnoses allowed to take into account mixed aetiology) was ‘dementia not otherwise specified’ (present in 92.9% of the cases). The high prevalence of this diagnostic category in the sample highlights the challenge of making a clear diagnosis of dementia subtypes in routine clinical scenario (Goodman et al., 2017). This issue will ultimately cause inappropriate cures to be delivered to patients, as well as difficulties in targeting well-phenotyped patient cohorts in clinical trials.

Diseases causing dementia such as Alzheimer’s disease are chronic non-linear progressive neurodegenerative disorders (Ewers et al., 2011a). Long before the clinical symptoms of the disease appear, neuroimaging, mainly MRI and PET, plays an important role in the diagnosis of dementias (Jagust, 2006). While structural MR images are routinely used to estimate cerebral atrophy, diffusion tensor imaging can be used to assess white matter integrity, and functional MR imaging is studied to detect changes in the functional brain networks. Information derived from PET images is of crucial value for the differential diagnosis of dementia (Herholz, 2014). ^{18}F -fluorodeoxyglucose (FDG) PET reflects the glucose consumption, which correlates with the activity of the synapses, while other PET tracers such as the Pittsburgh compound B (PiB) or Flortbetapir are used to image the deposition of beta-amyloid plaques in the brain—a core pathologic feature of certain dementia subtypes

(Bensaïdane et al., 2016). While a single imaging modality is not enough to differentiate between dementia subtypes, combining structural and functional modalities is expected to improve diagnostic accuracy (Jack et al., 2008), which explains the increased use of simultaneous PET/MR scanners to image the different aspects of dementia (Drzezga et al., 2014).

Computer-aided diagnosis of dementia

As mentioned in the previous section, neuroimaging offers an unmatched description of the brain’s structure and physiology, but the information it provides is not easy to extract and interpret. A popular way to extract meaningful information from brain images is to use computational methods based on machine learning to provide estimates of the category of pathology in a patient.

Classification and prediction of dementia with machine learning

Most classification methods rely on a number of training subjects, in the context of dementia studies usually patients with Alzheimer’s disease and cognitively normal subjects, together with a set of labels that indicate to which group each subject belongs. The first step of these methods, known as feature extraction, involves transforming the original data into a set of features, e.g. transforming a 3D image into a column vector of features that encodes the pattern of glucose consumption. An optional step, the feature selection, involves the selection of a subset of features expected to facilitate learning, e.g. selecting regions of interest expected to differ between groups. The training feature vectors and labels are then fed into a learning algorithm—kernel methods, and more particularly support vector machines, or decision trees/random forests being the most popular algorithms. Finally, the learned model can be applied to previously unseen testing subjects, which have been put through the same feature extraction and selection procedure as the training subjects. An alternative to traditional approaches using hand-crafted features is to let algorithms learn the features that optimally represent the data. This concept lies at the basis of many deep learning algorithms, which have become very popular in the medical image analysis community (Litjens et al., 2017).

A large number of machine learning and deep learning approaches have been proposed to classify and predict Alzheimer’s disease stages (Burgos et al., 2020; Burgos et al., 2021a; Haller et al., 2011; Falahati et al., 2014; Rathore et al., 2017; Jo et al., 2019; Ebrahimighahnavieh et al., 2020; Frizzell et al., 2022; Fathi et al., 2022), see Figure 3. However, only a few are translated to the clinic. This can be explained by different factors.

Lack of reproducibility Reproducibility is defined as the ability to reproduce results based on the same data and methodology. Key elements of reproducible research include data sharing, fully automatic data manipulation and sharing of code. Without these elements, results cannot be reproduced, a step essential to guarantee the robustness of a technique. Initiatives have emerged to improve the reproducibility of ML and DL approaches applied to neuroimaging.

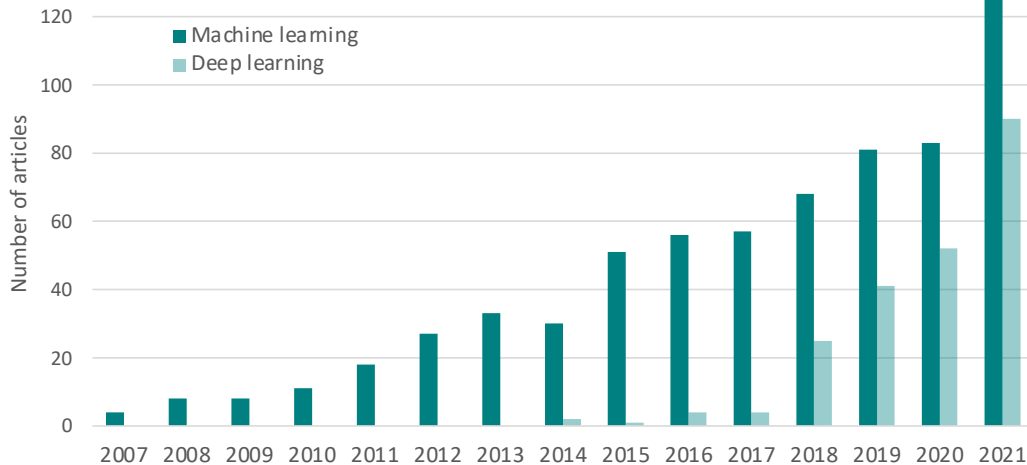


FIGURE 3: Number of articles presenting machine learning and deep learning approaches for the computer-aided diagnosis of Alzheimer’s disease published over the years according to PubMed (query available in Appendix A).

In particular, we have proposed reproducible frameworks for the evaluation of Alzheimer’s disease classification methods that comprise data management tools that rely on a community standard (Gorgolewski et al., 2016); image preprocessing and feature extraction pipelines; standard classification algorithms and CNN models; and rigorous validation procedures (Samper-González et al., 2018; Wen et al., 2020). These tools are available in the open-source software platforms Clinica (Routier et al., 2021, www.clinica.run) and ClinicaDL (Thibeau-Sutre et al., 2022b, <https://clinicadl.readthedocs.io>). These contributions will be described in Chapter 3.

Data leakage Unbiased evaluation of ML and DL algorithms is critical to assess their potential clinical value. A major source of bias is data leakage, which refers to the use of test data in any part of the training process (Kriegeskorte et al., 2009; Rathore et al., 2017). Several causes of data leakage exist and have been found in published works, as revealed in (Wen et al., 2020; Yagis et al., 2021). Not splitting the dataset at the subject-level when defining the training, validation and test sets can result in data from the same subject to appear in several sets. This problem can occur when patches or slices are extracted from a 3D image, or when images of the same subject have been acquired at multiple time points. Performing procedures such as feature selection or data augmentation before the training/validation/test split means, in the case of feature selection, that the test set is used to select the most relevant features. The absence of an independent test set implies that the same data have been used to select the optimal hyper parameters of the method and evaluate the performance.

We have tackled this issue in our software for reproducible neuroimaging processing with deep learning (Thibeau-Sutre et al., 2022b). ClinicaDL prevents data leakage as train and validation data characteristics are saved when the output structure is created. Then, when evaluating the performance of a trained model on a new data group, ClinicaDL checks that this data group does not share participants with the training and validation groups. Note

however that this only works under the assumption that participants are always named in the same way across data groups.

Generalisability Assessing the ability of an approach to generalise from highly-controlled research data of a certain cohort to another cohort, and more generally to routine data, is an essential step to ensure translation to the clinic. However, this step is rarely reached. For example, Bouts et al., 2019 assessed whether an MRI-based classification method trained to detect mild cognitive impairment on a clinical cohort could be used on a general population. Even though the model could detect mild cognitive impairment better than chance, the classification performance was moderate and probably insufficient to efficiently assist diagnosis.

We have studied the performance of machine and deep learning algorithms for the computer-aided diagnosis of dementia from anatomical MRI on clinical routine data originating from a hospital data warehouse. This work will be described in Chapter 4.

Interpretation The ability to understand why the ML and DL models take a given decision is a key issue to facilitate their acceptance, know how far they can be trusted and achieve better performance. The main idea for image-based classification methods is to highlight the parts of the image that contribute the most to the decision (Thibeau-Sutre et al., 2022a). This will be the topic of Chapter 2.

Detection of brain anomalies

In machine learning classification methods developed for dementia studies, neuroimaging features, e.g. glucose consumption extracted from PET images, are often used to draw the border that differentiates normality from abnormality. However, these features are affected by the anatomical and metabolic variabilities present in the population, which acts as confounding factors making the task of finding the frontier (i.e. the decision function) between normality and abnormality very challenging. Additionally, classification algorithms have usually been developed for two-class problems, such as differentiating patients with Alzheimer's disease from cognitively normal subjects, or differentiating two dementia subtypes (Burgos et al., 2021a). However, one of the key issues in the clinic is the differential diagnosis of all dementia subtypes.

An alternative to computer-aided diagnosis systems that predict the current or future diagnosis is to detect anomalies in an unsupervised way. This approach consists of generating a subject-specific model of healthy appearance for the targeted imaging modality, and comparing the subject's real image to the model. This results in the generation of a subject-specific map of anomalies. Different strategies exist to generate pseudo-healthy models. I proposed an approach based on a registration and fusion algorithm (Burgos et al., 2021b) but deep generative models have recently demonstrated their ability to detect anomalies in medical images (Chen et al., 2022). These approaches will be described in Chapter 1.

Chapter 1

Anomaly detection for the computer-aided diagnosis of dementia

This chapter results from the work that I started at the end of my PhD at UCL in the Centre for Medical Imaging Computing and that I pursued as a postdoc in the ARAMIS Lab at the Paris Brain Institute. Corresponding publications:

- **Burgos, N.**, Cardoso, M.J., Samper-González, J., Habert, M.-O., Durrleman, S., Ourselin, S., Colliot, O.: ‘Anomaly Detection for the Individual Analysis of Brain PET Images’. *Journal of Medical Imaging*, 8(2): 024003, 2021. [doi:10. 1117/1.JMI.8.2.024003](https://doi.org/10.1117/1.JMI.8.2.024003) • [hal-03193306](https://hal.archives-ouvertes.fr/hal-03193306)
- **Burgos, N.**, Samper-González, J., Bertrand, A., Habert, M.-O., Ourselin, S., Durrleman, S., Cardoso, M.J., Colliot, O.: ‘Individual Analysis of Molecular Brain Imaging Data through Automatic Identification of Abnormality Patterns’. In *Molecular Imaging, Reconstruction Analysis of Moving Body Organs, Stroke Imaging Treatment*, LNCS, 10555: 13–22, Springer, 2017. [doi:10.1007/978-3-319-67564-0_2](https://doi.org/10.1007/978-3-319-67564-0_2) • [hal-01567343](https://hal.archives-ouvertes.fr/hal-01567343)
- **Burgos, N.**, Cardoso, M.J., Mendelson, A.F., Schott, J.M., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Subject-Specific Models for the Analysis of Pathological FDG PET Data’. In *Medical Image Computing Computer-Assisted Intervention* • *MICCAI 2015*, LNCS, 9350: 651–658, Springer, 2015. [doi:10.1007/978-3-319-24571-3_78](https://doi.org/10.1007/978-3-319-24571-3_78)

1.1 Introduction

Neurological diseases such as epilepsy (De Tiege et al., 2004) or dementia (Mosconi et al., 2008; Cerami et al., 2016b) show heterogeneous patterns of anomalies on neuroimages, for example positron emission tomography (PET) images. These specific patterns of anomaly are important to distinguish between different syndromes and establish an accurate diagnosis (Herholz, 1995; Panegyres et al., 2009; Heiss et al., 2012; Ossenkoppele et al., 2013; Rice

et al., 2017; Bouwman et al., 2018). In clinical practice, PET images are mostly analysed visually. The sensitivity and specificity of this approach greatly depends on the observer's experience and is not in favour of centres where advanced expertise in image reading is unavailable (Perani et al., 2014). Quantitative analysis of PET images would alleviate this problem by helping define an objective limit between normal and pathological findings.

PET uptake can be quantitatively evaluated either regionally or on a voxel-by-voxel basis. In regional analysis, the regional uptake is compared with the regional uptake expected in a normal control population. This analysis usually requires prior knowledge to select the appropriate atlas and relevant discriminant regions, which should be adapted to a specific pathology, limiting its use (Signorini et al., 1999).

In voxel-wise analysis, a subject's PET image is usually aligned to a standardised group space to compare the uptake of the spatially normalised scan to a distribution obtained from normal control scans, on a voxel-by-voxel basis. The approach implemented in Neurostat (Minoshima et al., 1995b; Drzezga et al., 2005) consists of registering the PET image of the subject under investigation to a standard space and comparing it to a population of controls by means of a Z-score. The Z-score map is then projected onto different surfaces resulting in three-dimensional stereotactic surface projections that are used for image interpretation. Other software tools implementing a similar technique have been used for the analysis of PET data, such as NeuroGam (GE Healthcare, Waukesha, WI, USA) (Renard et al., 2013). Signorini et al., 1999 used the general linear model implemented in the Statistical Parametric Mapping (SPM) software package (Friston et al., 1994) to compare a subject's PET image to a population of controls. The t statistic corresponds to the difference between the mean uptake of the control group and the uptake of the subject being analysed, divided by the error estimated with the control group data (after correction for age and global metabolism). A similar approach was also used by Patterson et al., 2011, Perani et al., 2014 and Cerami et al., 2016a. Exploratory in nature, voxel-wise techniques require less prior information than regional analysis, but their sensitivity is limited by inter-subject variability in non-pathological tracer uptake, making pathological effects harder to detect (Signorini et al., 1999). The fact that the images have to be registered to a standard space can also decrease the sensitivity as the non-linear registration step may conceal subtle anomalies.

We proposed a framework for the individual analysis of PET data that consists of creating a subject-specific model of healthy PET appearance and comparing the patient's PET image to the model via a regularised Z-score (Burgos et al., 2015; Burgos et al., 2017b; Burgos et al., 2021b). The resulting voxel-wise Z-score map can be interpreted as an abnormality map, as it statistically evaluates the localised deviation of the subject-specific uptake with respect to the healthy uptake distribution. The abnormality maps are meant to help clinicians identify more easily pathological areas and also improve the interpretability of subsequent computer-aided analyses. We validated the proposed framework first by generating abnormality maps for healthy control subjects to ensure that no erroneous abnormalities are detected. We then generated abnormality maps for subjects at different stages of Alzheimer's disease (AD) and used them as features to feed a classifier. This was to ensure that the proposed method was able to extract for each individual the signal

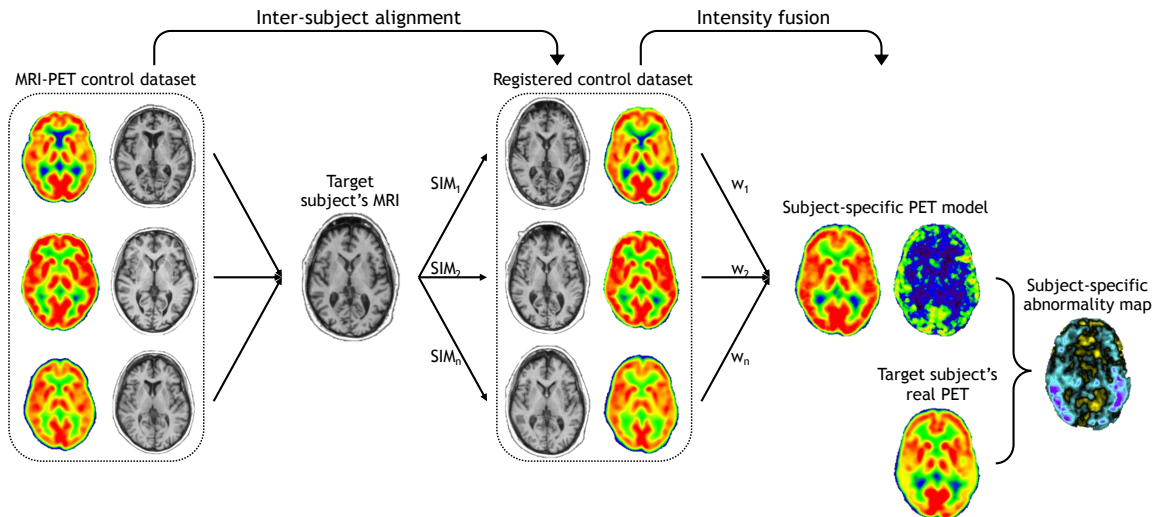


FIGURE 1.1: Subject-specific anomaly detection framework. The control dataset is first transported into the subject space. To generate the subject-specific model of healthy PET appearance, the set of registered PETs is locally selected and fused. Finally, the subject-specific abnormality map is computed by means of a regularised Z-score.

characteristic of abnormality. Finally, we applied the proposed framework to a dataset of subjects with different types of frontotemporal dementia (FTD) syndromes.

1.2 Methods

The proposed anomaly detection framework, illustrated in Figure 1.1, consists of selecting, in a control dataset, the subjects that are the most similar to the subject being analysed in terms of demographic characteristics and morphology, creating a subject-specific model of healthy PET uptake from the selected controls and the target subject's anatomical magnetic resonance (MR) image, and using the resulting model to create a subject-specific abnormality map.

1.2.1 Database of control atlases

The approach relied on a database of atlases, each composed of a pair of co-registered anatomical MR and PET images. The atlases were globally aligned in a common space. This was performed by mapping to a common coordinate frame the MR images from all the atlases via an affine groupwise registration (Modat et al., 2014). The transformations were then applied to the MR and PET images by updating their image coordinate system (without resampling).

1.2.2 Atlas pre-selection

An atlas pre-selection step was performed to discard the control atlases too dissimilar to the target and thus limit the computational time while maintaining a high synthesis accuracy. Two strategies were explored: one relying on the demographic characteristics, the other on the anatomical images themselves.

1.2.2.1 Demographic-based pre-selection

Both age and sex have been shown to influence brain metabolism, even though it is not clear whether this can be explained by the underlying morphology (Murphy et al., 1996; Cosgrove et al., 2007; Kalpouzos et al., 2009; Curiati et al., 2011; Knopman et al., 2014). To limit this influence, the demographic-based pre-selection first consisted of selecting the control atlases of the same gender as the target. The atlases closest in terms of age to the target were then picked.

1.2.2.2 Image-based pre-selection

The image-based strategy preselects the atlases according to their global morphological similarity to the target, as assessed by a global similarity measure, the normalised cross-correlation (NCC). The anatomical MR image of a randomly-chosen reference atlas was affinely registered to the MR image of the target subject. Because all the control atlases were pre-aligned with each other, the resulting affine transformation was applied to the anatomical MR image of each control atlas and the NCC was computed between each resampled control atlas and the target subject. The control atlases with the highest NCC were selected.

1.2.3 Inter-subject registration and atlas local selection

The anatomical MR images of the pre-selected atlases were non-rigidly registered to the target subject's MR image (Modat et al., 2010), and the PET images of the atlases, pre-aligned with the MR images, were mapped using the same transformation to the target subject. This inter-subject coordinate mapping was obtained using a symmetric global registration followed by a cubic B-spline parametrised non-rigid registration as implemented in NiftyReg (Modat et al., 2010). The normalised mutual information was used as similarity measure. The non-rigid registrations were performed with a pyramidal approach with three levels. The finer lattice of control points had a spacing of 5 mm along each axis.

Once non-rigidly aligned with the target subject, the atlases closest in terms of morphology to the target subject were identified. This morphological similarity was assessed at the voxel level using a local image similarity measure, the structural image similarity (SSIM) (Wang et al., 2004). The SSIM between two anatomical MR images I^{MRI} and J^{MRI} at voxel x is given by

$$\text{SSIM}(I^{\text{MRI}}(x), J^{\text{MRI}}(x)) = \frac{2\mu_{I^{\text{MRI}}}(x)\mu_{J^{\text{MRI}}}(x) + C_1}{\mu_{I^{\text{MRI}}}^2(x) + \mu_{J^{\text{MRI}}}^2(x) + C_1} \times \frac{2\sigma_{I^{\text{MRI}}, J^{\text{MRI}}}(x) + C_2}{\sigma_{I^{\text{MRI}}}^2(x) + \sigma_{J^{\text{MRI}}}^2(x) + C_2}. \quad (1.1)$$

C_1 and C_2 are two constants used to improve the stability of the structural similarity that depend on the range of the voxel values (Wang et al., 2004). The means and standard deviations were calculated using a Gaussian kernel G_{σ_G} with standard deviation σ_G through density normalised convolution (Cachier et al., 2003; Burgos et al., 2017a)

$$\begin{aligned} \mu_{I^{\text{MRI}}}(x) &= \frac{[G_{\sigma_G} * I^{\text{MRI}}](x)}{[G_{\sigma_G} * \Omega](x)}, & \sigma_{I^{\text{MRI}}}^2(x) &= \mu_{I^{\text{MRI}^2}(x)} - \mu_{I^{\text{MRI}}}(x)^2, \\ \sigma_{I^{\text{MRI}}, J^{\text{MRI}}}(x) &= \mu_{I^{\text{MRI}}, J^{\text{MRI}}}(x) - \mu_{I^{\text{MRI}}}(x) \cdot \mu_{J^{\text{MRI}}}(x), \end{aligned}$$

where $*$ denotes the convolution operator. $G_{\sigma_G} * \Omega$ represents a density normalisation term that compensates for areas with missing information where Ω is a density function equal to 1 where the fields of view overlap and 0 otherwise.

The local selection was performed via a weighted scheme. The weights reflect the contribution of each control atlas to the model. They were obtained at each voxel x by ranking the SSIM across the N control atlases globally pre-selected and by applying an exponential decay function:

$$w_n(x) = e^{-\beta r_n(x)} , \quad (1.2)$$

where $r_n(x)$ denotes the rank of the n^{th} control atlas. Using the rank instead of the SSIM value means that the sum and separation of the weights for different voxels are the same at each voxel location, leading to more stable results (Burgos et al., 2014). For each voxel, the atlases contributing the most to the model are the ones with the highest morphological similarity to the target subject.

1.2.4 Subject-specific models of healthy PET appearance

To generate the subject-specific model, which is composed of two elements: a spatially-varying weighted average and a spatially-varying weighted standard deviation, the control atlases locally selected were fused based on their morphological similarity to the target subject. For each of the N pre-selected atlases in the control dataset, let the n^{th} mapped PET image be denoted by J_n^{PET} . The two subject-specific model elements (I_μ^{PET} , I_σ^{PET}) were computed as follows:

$$\begin{aligned} I_\mu^{\text{PET}}(x) &= \frac{\sum_{n=1}^N w_n(x) \cdot J_n^{\text{PET}}(x)}{\sum_{n=1}^N w_n(x)} , \\ I_\sigma^{\text{PET}}(x) &= \sqrt{\frac{N_w}{N_w - 1} \frac{\sum_{n=1}^N w_n(x) \cdot (J_n^{\text{PET}}(x) - I_\mu^{\text{PET}}(x))^2}{\sum_{n=1}^N w_n(x)}} \end{aligned} \quad (1.3)$$

where N_w is the number of non-zero weights.

1.2.5 Subject-specific abnormality maps

To compare the target subject's PET image (I^{PET}) to the subject-specific model, in our preliminary work (Burgos et al., 2015) a Z-score was computed for each voxel of the image. However, we observed that this leads to the generation of high frequency signals in certain areas due to the standard deviation approaching zero. To avoid this problem, we define a regularised Z-score

$$\tilde{Z}(x) = \frac{I^{\text{PET}}(x) - I_\mu^{\text{PET}}(x)}{\hat{I}_\sigma^{\text{PET}}(x)} \quad (1.4)$$

with

$$\hat{I}_\sigma^{\text{PET}}(x) = \begin{cases} P_k(I_\sigma^{\text{PET}}) & \text{if } I_\sigma^{\text{PET}}(x) < P_k(I_\sigma^{\text{PET}}) \\ I_\sigma^{\text{PET}}(x) & \text{otherwise} \end{cases} , \quad (1.5)$$

TABLE 1.1: Summary of the ADNI participants’ demographics, clinical scores and amyloid status.

	N	Age	Gender	MMSE	CDR	Amyloid status
CN _{opti}	23	74.6 ± 5.8 [65.4; 89.0]	9 F / 14 M	28.4 ± 1.6 [24; 30]	0: 23	23 - / 0 +
CN	131	73.6 ± 6.4 [56.2; 85.6]	67 F / 64 M	29.1 ± 1.1 [25; 30]	0: 131	89 - / 42 +
EMCI	142	70.9 ± 7.0 [55.5; 88.6]	65 F / 77 M	28.4 ± 1.6 [24; 30]	0: 1; 0.5: 141	74 - / 68 +
LMCI	120	72.0 ± 7.9 [55.0; 91.4]	62 F / 58 M	27.7 ± 1.8 [24; 30]	0.5: 119; 1:1	42 - / 78 +
AD	99	74.4 ± 8.0 [55.9; 88.5]	44 F / 55 M	22.9 ± 2.2 [19; 26]	0.5: 46; 1: 52; 2: 1	12 - / 87 +

Abbreviations: CN: cognitively normal; EMCI: early mild cognitive impairment; LMCI: late mild cognitive impairment; AD: Alzheimer’s disease; MMSE: Mini-Mental State Examination; CDR: global Clinical Dementia Rating.

where $P_k(I_\sigma^{\text{PET}})$ is the k^{th} percentile of I_σ^{PET} , computed only from brain voxels using a sorting-based algorithm.

The voxel-wise regularised Z-score map can be interpreted as an abnormality map, as it statistically evaluates the localised deviation of the subject-specific uptake with respect to the healthy uptake distribution.

1.3 Validation on Alzheimer’s disease

1.3.1 Materials: Alzheimer’s disease cohort

Part of the data used in this work were obtained from the Alzheimer’s Disease Neuroimaging Initiative database (ADNI)¹. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

We selected 515 participants from the ADNI2 database who had T1-weighted (T1w) MRI, ¹⁸F-FDG PET, and Florbetapir (¹⁸F-AV45) PET images at baseline and were diagnosed as cognitively normal (CN), early MCI (EMCI), late MCI (LMCI) or AD. The diagnosis relies on three main criteria: the absence or presence of memory complaints, the Mini-Mental State Examination (MMSE) score and the Clinical Dementia Rating (CDR) score (Petersen et al., 2010). 23 CN subjects with an amyloid negative status were extracted from the main dataset for parameter optimisation purposes. This set is referred to as the CN_{opti} dataset in the following. Table 1.1 summarises the demographic characteristics, clinical scores, and amyloid status of the participants. Subjects were categorised as amyloid positive ($A\beta^+$) or negative ($A\beta^-$) based on a cortical mean cutoff of 1.11 on Florbetapir PET (Landau et al., 2012). The database of control atlases used in this paper is composed of the CN participants amyloid negative ($n = 89$).

¹adni.loni.usc.edu, see also Appendix B

1.3.1.1 Image acquisition

The acquisition protocols of the 3D T1w images can be found in (Jack et al., 2010). The images were downloaded after correction of image geometry distortion due to gradient non-linearity (gradwarp) and correction of the image intensity non-uniformity (Jack et al., 2010). The ADNI2 FDG PET protocol consisted of a dynamic acquisition of six five-minute frames, 30 to 60 minutes post-injection, and the Florbetapir PET protocol consisted of a dynamic acquisition of four five-minute frames from 50 to 70 minutes post-injection (Jagust et al., 2015). For both tracers, images were downloaded after several stages of preprocessing: frame averaging, spatial alignment, interpolation to a standard voxel size, and smoothing to a common resolution of 8 mm full width at half maximum.

1.3.1.2 Image preprocessing

For each subject, the T1w MR image was corrected for intensity non-uniformity following a non-parametric intensity non-uniformity normalisation method (Tustison et al., 2010) and was mapped to the PET images using a rigid transformation. The T1w MR images, resampled to the PET voxel grid of $1.5 \times 1.5 \times 1.5 \text{ mm}^3$, were then parcellated into 143 different regions using a multi-atlas propagation and fusion algorithm implemented in NiftySeg (Cardoso et al., 2015). The PET images were intensity-normalised using the average uptake in reference regions that were extracted from the parcellated T1w MR images. The pons was used for the FDG PET images (Minoshima et al., 1995a) and the whole cerebellum for the Florbetapir PET images (Hutton et al., 2015). The MR and PET images of all the subjects were globally aligned in the common space of the control atlases (Section 1.2.1), via an affine groupwise registration (Modat et al., 2014). This step is not necessary to generate abnormality maps but facilitates the subsequent group-space analyses.

1.3.2 Parameter optimisation

Two parameters were optimised using a leave-one-out cross validation on the 23 subjects from the CN_{opti} subset: the standard deviation of the Gaussian kernels used in Eq. (1.1) (σ_G , expressed in voxels), which controls the size of the neighbourhood where the local similarity measure is computed:

$$\sigma_G = \left[1 \quad 3 \quad 5 \quad 7 \quad 9 \right] ,$$

and β from Eq. (1.2) whose value influences the repartition of the weights:

$$\beta = \left[1 \quad 0.5 \quad 0.25 \quad 0.1 \quad 0.01 \right] .$$

The weighted sum tends to the mean when β is small.

For each tracer, a pseudo-healthy PET image (I_μ in Eq. (1.3)) was generated using the proposed method from the subject’s T1w MR image. This pseudo-healthy PET image was

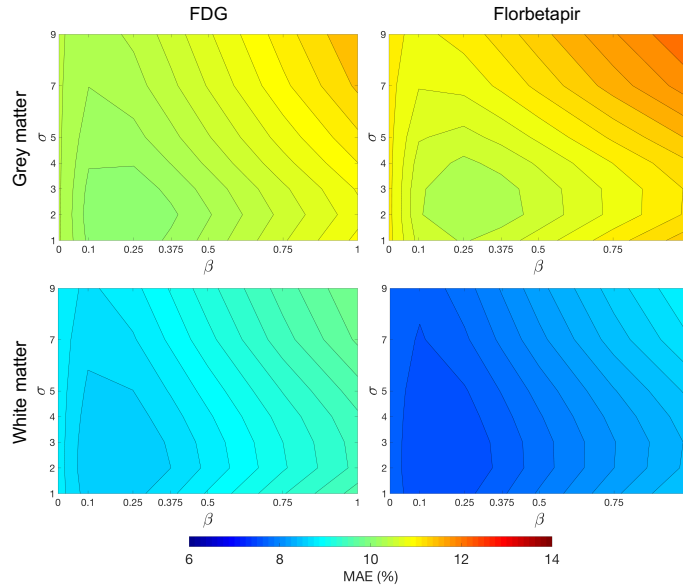


FIGURE 1.2: Average rMAE between the real and pseudo-healthy PET images in the grey matter (GM) and white matter (WM) regions for varying values of σ_G and β . The selected parameters are $\sigma_G = 3$ and $\beta = 0.25$, giving an average rMAE of 10.1% (GM) and 8.6% (WM) for FDG PET and 10.4% (GM) and 7.6% (WM) for Florbetapir PET.

then compared to the real PET image (I) using the relative mean absolute error, defined as

$$\text{rMAE} = 100 * \frac{\sum_x |I_\mu(x) - I(x)|}{\sum_x I(x)}. \quad (1.6)$$

The rMAE was computed in the grey matter (GM) and white matter (WM) regions for both tracers. Values averaged over all subjects are shown in Figure 1.2. For both tracers, the lowest MAE were obtained with $\sigma_G = 2$ or 3 and $\beta = 0.25$. We chose $\sigma_G = 3$ to favour smoother and less patchy images. The selected parameters are thus $\sigma_G = 3$ and $\beta = 0.25$, giving an average rMAE of 10.1% (GM) and 8.6% (WM) for FDG PET and 10.4% (GM) and 7.6% (WM) for Florbetapir PET.

The percentile used in Eq. (1.5) to regularise the Z-score was determined experimentally as a compromise between reducing the number of outliers and preserving enough standard deviation information. It was set to $k = 10$.

1.3.3 Validation scheme

The validation was performed in two steps. First, the pseudo-healthy PET images of the CN subjects amyloid negative were compared with their real PET images. The abnormality maps were then used as features to perform individual classification.

1.3.3.1 Synthesis accuracy

The first step of the validation consisted of comparing the pseudo-healthy PET images of the CN subjects amyloid negative to their real PET images. As these subjects should not present abnormalities, the pseudo-healthy and real PET images should be as similar as possible. This similarity was assessed using the rMAE.

1.3.3.2 Individual classification

To assess the ability of the abnormality maps to extract relevant information from PET data, the abnormality maps were used as features to feed a linear SVM classifier.

Non-linear alignment to group space A way to compare the abnormality maps, each generated in the subject’s native space, across all the subjects, is to align them with each other. As the T1 images from all the subjects were already mapped to a common coordinate frame via an affine groupwise registration, the T1 images were subsequently non-rigidly registered to the group-space. The same transformations were then applied to the abnormality maps.

Linear SVM classifier We chose a linear SVM to classify the abnormality maps. A linear kernel was calculated using the inner product for each pair of abnormality maps available in the dataset (using all the brain voxels). The cross-validation (CV) procedure included two nested loops: an outer loop evaluating the classification performance and an inner loop used to optimise the hyperparameter C that regularises the SVM. For the outer loop, we used 250 stratified shuffle splits with a test size of 30%. Note that, for a same task, the splits were kept the same between the different types of features tested. We used an inner k -fold with $k = 10$. This individual classification was performed with tools implemented in Clinica (Routier et al., 2021; Samper-González et al., 2018) that rely on scikit-learn (Pedregosa et al., 2011).

Classification tasks The experiments consisted of different tasks with varying degrees of difficulty:

- differentiating cognitively normal subjects from subjects with a disease, i.e. CN $A\beta^-$ vs AD $A\beta^+$, CN $A\beta^-$ vs LMCI $A\beta^+$, CN $A\beta^-$ vs EMCI $A\beta^+$ (using FDG PET);
- differentiating between subjects at the beginning and at the end of the early stage of the disease, i.e. EMCI $A\beta^+$ vs LMCI $A\beta^+$ (using FDG PET);
- differentiating between amyloid negative and amyloid positive subjects, i.e. $A\beta^-$ vs $A\beta^+$ (using Florbetapir PET).

For the first two experiments, 322 subjects (89 CN $A\beta^-$, 68 EMCI $A\beta^+$, 78 LMCI $A\beta^+$ and 87 AD $A\beta^+$) were considered, while for the last experiment 492 subjects (217 $A\beta^-$ and 275 $A\beta^+$) were analysed.

Comparison with standard approaches To set the results in perspective, the subjects’ PET image itself and two standard Z-maps were also used as features and fed to the classifier. For each subject, a first Z-map was obtained by comparing the subject’s PET image in the group space to the mean and standard deviation computed from the 89 subjects of the control dataset, also in the group space. A second Z-map was obtained by first pre-selecting the control subjects using demographic characteristics, as in Section 1.2.2.1, and comparing the subject’s PET image in the group space to the mean and standard deviation computed from the pre-selected control subjects, also in the group space.

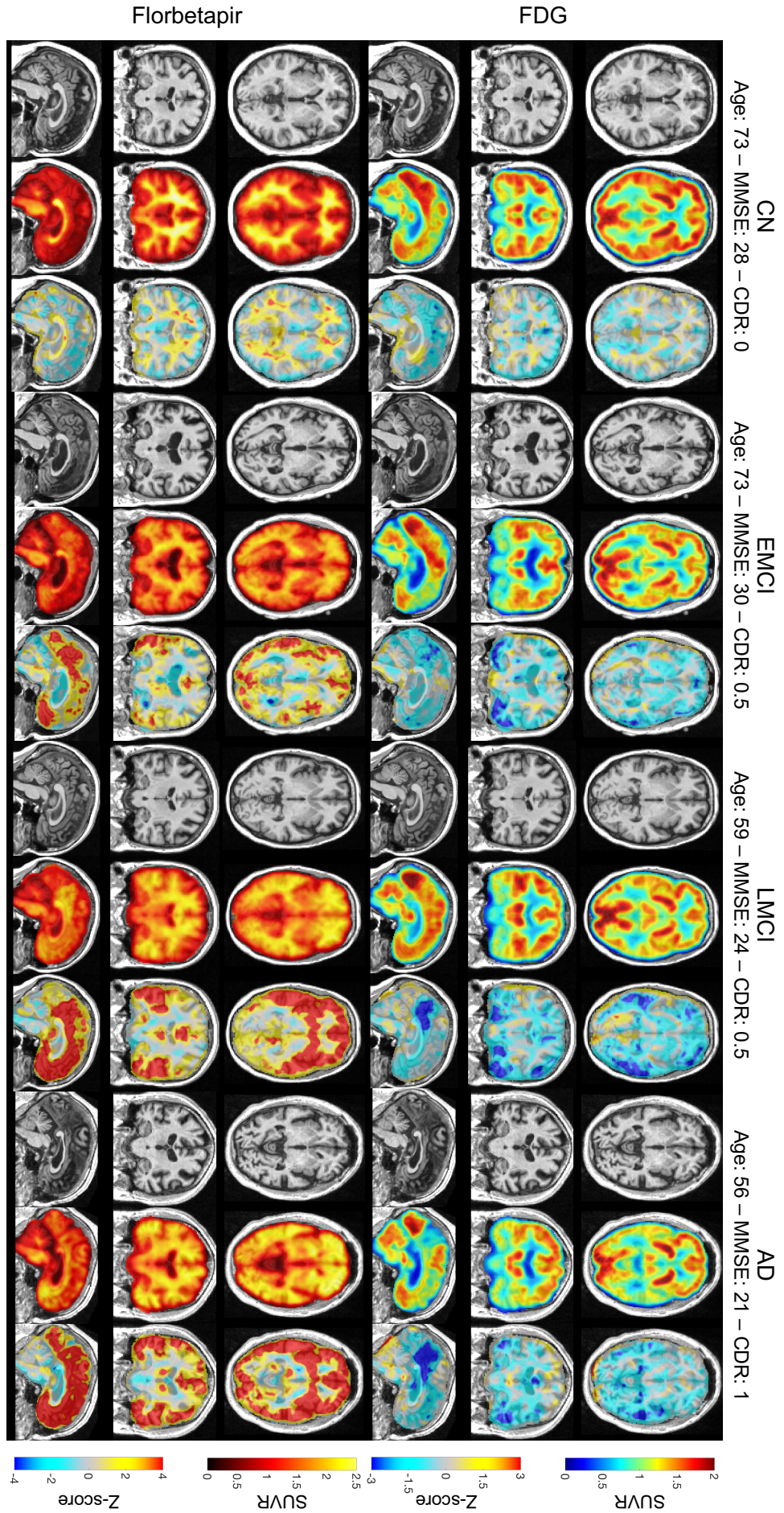


FIGURE 1.3: Examples of FDG (top) and Florbetapir (bottom) abnormality maps obtained for a cognitively normal subject (CN) and patients with early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) and Alzheimer's disease (AD). For each subject, the T1w MRI, PET and abnormality map (demographic-based pre-selection) are displayed. A highly negative Z-score (blue) indicates a reduced PET uptake in the subject relative to the controls while a highly positive Z-score (red) indicates an increased PET uptake in the subject relative to the controls.

1.3.4 Results

Abnormality maps were generated for each of the 492 ADNI2 participants selected, for both the FDG and Florbetapir PET images. Note that for the CN β^- subjects (forming the control dataset), a leave-one-out strategy was used, i.e. the images of the CN subject being processed were excluded from the control database. For the demographic-based pre-selection, the 30 control atlases of the same gender and closest in terms of age to the target subject were selected. For the image-based pre-selection, the 30 control atlases with the highest NCC were selected. The number of pre-selected atlases was chosen as a compromise between having a representative database of controls and computational time.

Examples of abnormality maps are displayed in Figure 1.3 for a CN, an early MCI, a late MCI and an AD subject. We observe that, as expected, no specific signal is being detected for the CN subject, for both the FDG and Florbetapir tracers. For the EMCI and LMCI subjects, abnormal glucose uptake is detected mainly in the precuneus and in the medial-lateral temporal lobe, and abnormal amyloid deposition is detected in the frontal, parietal, temporal and cingulate cortices, which is consistent with previous observations (Forsberg et al., 2008). Finally, for the AD subject, abnormal glucose uptake is also detected mainly in the precuneus and in the medial-lateral temporal lobe, and abnormal amyloid deposition is detected in all the cortex, which is typical of AD (Jagust, 2006).

1.3.4.1 Synthesis accuracy

The rMAE results obtained when comparing the pseudo-healthy PET images of the CN subjects amyloid negative with their real PET images are displayed in Figure 1.4.

The pre-selection strategy has no significant impact ($p > 0.05$, paired t-test) on the synthesis accuracy: for the FDG tracer, the average rMAE is of 11.6% for the image-based and 11.4% for the demographic-based pre-selection; and for the Florbetapir tracer, the average rMAE is of 11.5% for the image-based and 11.3% for the demographic-based pre-selection. These results are consistent with the ones obtained for the optimisation dataset (rMAE in the brain of 11.2% for FDG and 10.9% for Florbetapir).

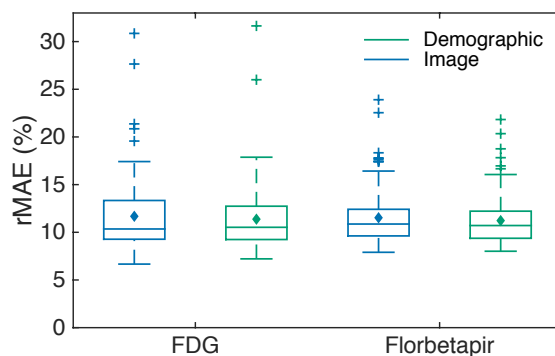


FIGURE 1.4: Boxplots displaying the mean (diamond), median, lower and upper quartiles, and minimum and maximum of the rMAE between the real and pseudo-healthy PET images for the CN amyloid negative subjects.

TABLE 1.2: Balanced accuracy obtained when using PET images, state-of-the-art Z-maps, and the subject-specific abnormality maps (both with image-based and demographic-based pre-selection) as features of the linear SVM classification algorithm. The average \pm SD balanced accuracy, obtained over 250 repeats, is expressed in percentages. For each task, the highest balanced accuracy is highlighted with bold font.

Tracer	Task	PET	Z-map		Abnormality map	
			No pre-select.	Demographic pre-select.	Image pre-select.	Demographic pre-select.
FDG	CN $A\beta^-$ vs AD $A\beta^+$	89.3 \pm 4.3	90.8 \pm 3.9	90.9 \pm 4.1	91.4 \pm 3.9	92.4 \pm 3.7
	CN $A\beta^-$ vs LMCI $A\beta^+$	73.4 \pm 6.1	75.3 \pm 5.2	77.2 \pm 5.3	74.9 \pm 5.2	76.7 \pm 5.3
	CN $A\beta^-$ vs EMCI $A\beta^+$	58.3 \pm 6.7	62.6 \pm 6.2	62.2 \pm 5.8	60.0 \pm 6.2	64.7 \pm 5.8
	EMCI $A\beta^+$ vs LMCI $A\beta^+$	58.3 \pm 6.0	59.1 \pm 6.2	60.6 \pm 6.1	56.3 \pm 6.4	61.1 \pm 6.0
Florbetapir	$A\beta^-$ vs $A\beta^+$	95.0 \pm 1.5	96.3 \pm 1.3	96.2 \pm 1.3	96.1 \pm 1.3	96.1 \pm 1.4

1.3.4.2 Individual classification

The balanced accuracies obtained for the different features and tasks are displayed in Table 1.2. No matter the task, the abnormality maps obtained with the demographic-based pre-selection lead to similar (Florbetapir PET tasks) or higher (FDG PET tasks) balanced accuracies than the abnormality maps obtained with the image-based pre-selection. The effect of the demographic-based pre-selection is also visible when comparing the Z-maps computed from the whole control dataset or from pre-selected control subjects: the accuracy is usually equivalent or higher in the later case. The classification performance is generally higher when the Z-maps and abnormality maps are used as features rather than the PET images themselves. For the majority of the FDG tasks, the balanced accuracy obtained for the proposed method with demographic-based pre-selection is slightly higher than the balanced accuracy obtained using the Z-maps with demographic-based pre-selection as features. The benefits of the abnormality maps seem to slightly increase with the difficulty of the task: 62.2% vs 64.7% for the CN vs EMCI task and 60.6% vs 61.1% for the EMCI vs LMCI task.

When analysing Florbetapir data, using the PET images themselves, the state-of-the-art Z-maps or the proposed abnormality maps leads to similar, highly accurate, classification results. These highly accurate results were expected, but are here confirmed, as differentiating amyloid negative from amyloid positive subjects based on features extracted from Florbetapir data is a quite trivial task.

Overall, the classification results obtained for the abnormality maps confirms their ability to detect meaningful signal from both FDG and Florbetapir PET images.

1.4 Application to frontotemporal dementia

1.4.1 Materials: frontotemporal dementia cohort

Data for the FTD cohort were obtained from NIFD, which uses the ADNI platform to make available data from the frontotemporal lobar degeneration neuroimaging initiative (FTLDNI). FTLDNI² is founded through the National Institute of Aging, and started in

²<http://memory.ucsf.edu/research/studies/nifd>, see also Appendix B

2010. The primary goals of FTLDNI are to identify neuroimaging modalities and methods of analysis for tracking frontotemporal lobar degeneration and to assess the value of imaging versus other biomarkers in diagnostic roles. The Principal Investigator of FTLDNI is Dr. Howard Rosen, MD at the University of California, San Francisco. The data is the result of collaborative efforts at three sites in North America.

We focused on the 12 participants who had T1w MRI and ^{18}F -FDG PET images with an iterative reconstruction at baseline. Four subjects were cognitively normal (CN), seven diagnosed as behavioural variant FTD (bvFTD) and one as semantic variant primary progressive aphasia (svPPA). The neuropsychological battery included functional measures, such as the FTD-specific Clinical Dementia Rating (CDR), and cognitive measures, such as the Mini-Mental State Examination (MMSE) that includes tests of orientation, attention, memory, language and visuospatial ability, the Boston Naming Test (BNT) that assesses word-finding ability, and a verbal fluency test assessing the ability to retrieve specific information within the animal category (Knopman et al., 2008; Staffaroni et al., 2019). The subjects' demographic characteristics and clinical scores are displayed in Figure 1.5.

1.4.1.1 Image acquisition

Both the MRI and PET images were acquired at the Mayo Clinic, Rochester. 3D T1w MR images were acquired on a GE Discovery MR750 or GE Signa HDxt 3T scanner using the following sequence parameters: $\text{TR} \approx 7$ ms, $\text{TE} \approx 3$ ms, $\text{TI} = 900$ ms, flip angle = 8° , slice thickness = 1.2 mm, in-plane resolution = 1.0×1.0 mm², matrix = $256 \times 256 \times 166$. The images were downloaded after correction of image geometry distortion due to gradient non-linearity (gradwarp) and correction of the image intensity non-uniformity. The FDG PET images were acquired on a GE Discovery RX PET/CT scanner following a protocol that consisted of a dynamic acquisition of six five-minute frames, 30 to 60 minutes post-injection. PET images were reconstructed using a 3D iterative method and displayed in a $256 \times 256 \times 47$ matrix (voxel size $1.17 \times 1.17 \times 3.27$ mm³).

1.4.1.2 Image preprocessing

The T1w MR images were corrected for intensity non-uniformity (Tustison et al., 2010) and parcellated into 143 different regions (Cardoso et al., 2015). For each PET acquisition, each frame was rigidly registered to the first frame and the co-registered frames were then averaged. The averaged PET image was mapped to the T1w MR image using a rigid transformation. The PET image, resampled to the T1w MRI, was smoothed using a Gaussian kernel with a standard deviation of two voxels (to obtain images with a resolution similar to that of the ADNI database) and intensity-normalised using the average uptake in the pons, which was extracted from the parcellated T1w MR image.

1.4.2 Results

Abnormality maps were generated for the 12 subjects of the FTD cohort using the CN subjects amyloid negative from the ADNI cohort as atlases. Demographic-based pre-selection

was performed by selecting the 30 control atlases of the same gender and closest in terms of age to the target subject.

Abnormality maps obtained for two bvFTD subjects and the svPPA subject are displayed in Figure 1.6. This figure highlights the fact that bvFTD is a very heterogeneous syndrome: neurodegeneration can affect the frontal lobe (Figure 1.6, centre), or also the temporal and parietal lobes (Figure 1.6, left) (Whitwell et al., 2009; Rohrer et al., 2013). The svPPA subject shows typical asymmetric hypometabolism affecting mainly the temporal pole, entorhinal area and hippocampus (Rabinovici et al., 2008).

Comparisons were restricted to ten clinically relevant regions for the sake of brevity. These regions were either selected to represent the areas where abnormal uptake, compared with controls, is expected for bvFTD:

- the orbitofrontal region, comprising the anterior, posterior, medial and lateral orbital gyri (Grimmer et al., 2004; Jeong et al., 2005; Diehl-Schmid et al., 2007; Tosun et al., 2016);
- the dorsolateral prefrontal region, comprising the inferior, middle and superior frontal gyri (Grimmer et al., 2004; Jeong et al., 2005; Diehl-Schmid et al., 2007; Kanda et al., 2008; Tosun et al., 2016);
- the ventromedial prefrontal region, comprising the gyrus rectus, medial frontal cortex, subcallosal area, superior frontal gyrus medial segment (Jeong et al., 2005; Diehl-Schmid et al., 2007; Kanda et al., 2008; Tosun et al., 2016);
- the lateral temporal region, comprising the inferior, middle and superior temporal gyri (Diehl-Schmid et al., 2007; Kanda et al., 2008; Tosun et al., 2016);
- the parietal region, comprising the precuneus and the supramarginal and angular gyri (Grimmer et al., 2004; Jeong et al., 2005; Diehl-Schmid et al., 2007; Tosun et al., 2016);
- the cingulate (Jeong et al., 2005; Diehl-Schmid et al., 2007; Kanda et al., 2008; Tosun et al., 2016);
- the insula (Jeong et al., 2005; Diehl-Schmid et al., 2007; Tosun et al., 2016);
- the midbrain region, comprising the caudate, pallidum, putamen and thalamus (Grimmer et al., 2004; Jeong et al., 2005; Diehl-Schmid et al., 2007; Kanda et al., 2008);

or svPPA:

- the anterior temporal region, comprising the hippocampus, amygdala, temporal pole and entorhinal area (Rabinovici et al., 2008);

or to act as a neutral region where no hypometabolism is expected:

- the occipital region, comprising the inferior, middle and superior occipital gyri, and occipital fusiform gyrus.

Each hemisphere was analysed separately to account for left/right asymmetry.

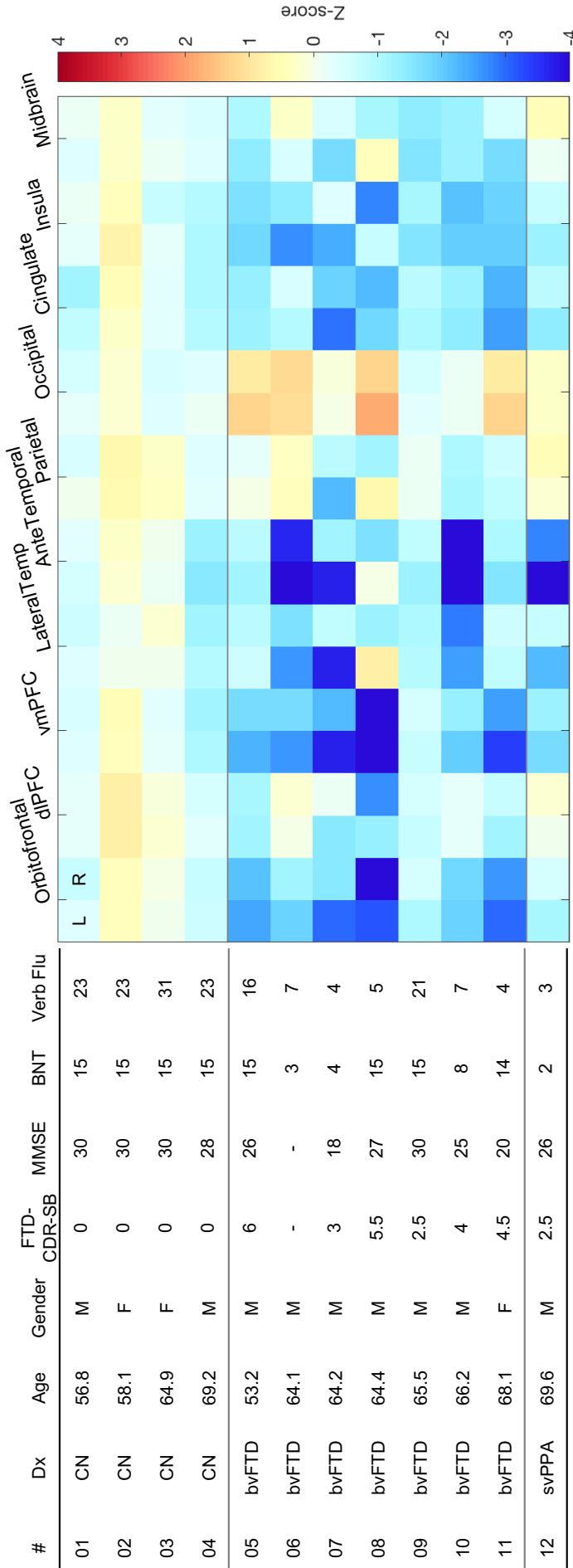


FIGURE 1.5: Left: NIFD participants' demographics and clinical scores. Right: Average Z-scores obtained with the proposed method for each subject and ROI. The subjects were sorted by diagnosis and age. A highly negative Z-score (blue) indicates a reduced FDG uptake in the subject relative to the controls while a highly positive Z-score (red) indicates an increased FDG uptake in the subject relative to the controls. Abbreviations: CN: cognitively normal; bvFTD: behavioural variant frontotemporal dementia (FTD); svPPA: semantic variant primary progressive aphasia; MMSE: Mini-Mental State Examination; FTD-CDR-SB: FTD-specific Clinical Dementia Rating sum of boxes; BNT: Boston Naming Test; Verb Flu: verbal fluency; dlPFC: dorsolateral prefrontal region; vmPLC: ventromedial prefrontal region; LateralTemp: lateral temporal region; AnteTemporal: anterior temporal region; L: left; R: right.

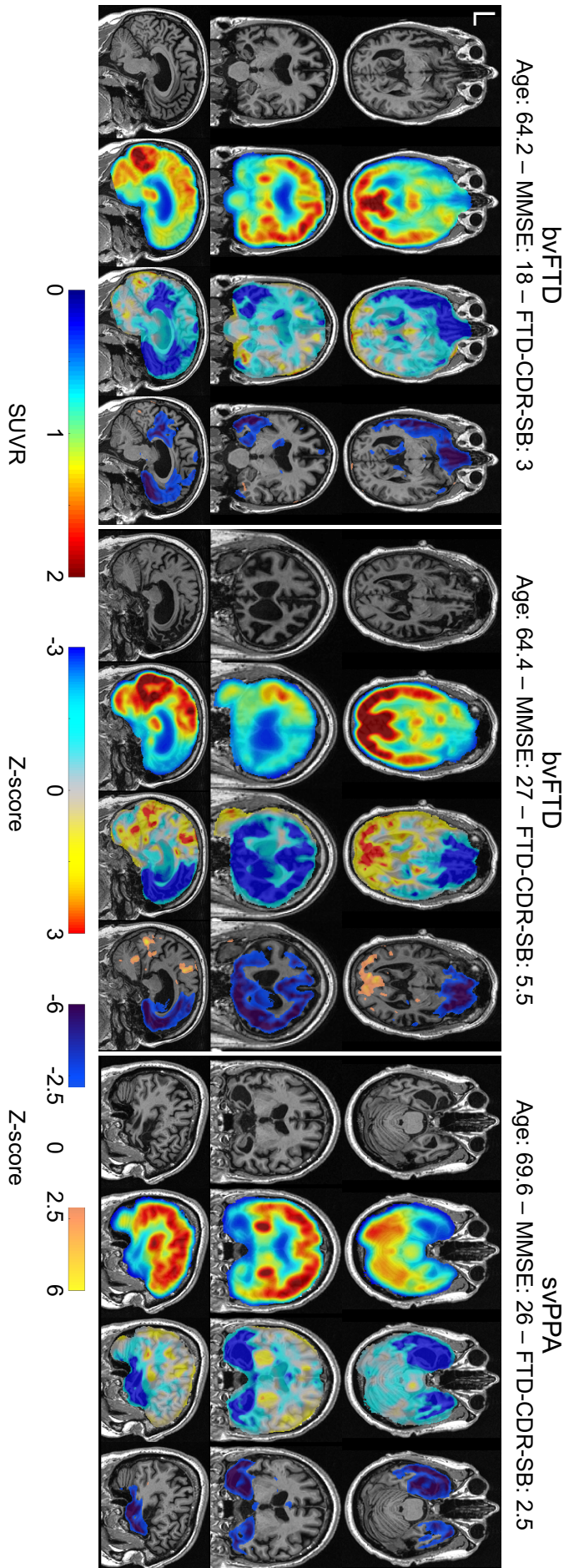


FIGURE 1.6: Examples of abnormality maps obtained for two subjects with bvFTD and a subject with svPPA. For each subject, the T1w MRI, PET and abnormality map, both with and without thresholding, are displayed. The threshold is set at 2.5. A highly negative Z-score (blue) indicates a reduced FDG uptake in the subject relative to the controls while a highly positive Z-score (red) indicates an increased FDG uptake in the subject relative to the controls. The first bvFTD subject corresponds to subject #07 in Figure 1.5, and the second to subject #08.

Figure 1.5 displays average Z-scores obtained with the proposed method for each ROI. We note that the controls have approximately zero Z-scores in all regions while highly negative Z-scores are observed for bvFTD and svPPA subjects in regions where a reduced uptake is expected when compared with healthy controls. The svPPA subject shows typical asymmetric hypometabolism affecting the anterior temporal region (Rabinovici et al., 2008). This figure again highlights the heterogeneity of bvFTD: the areas affected and the degree of hypometabolism vary across subjects. Several subjects present hypometabolism only in the frontal lobe (e.g. subject #08), while for others hypometabolism is present in both the frontal and temporal lobes (e.g. subject #07). The parietal region can also be affected, even though not as much as the frontal and temporal regions (subject #07). On the contrary, we observe for some patients a hypermetabolism in the occipital region (e.g. subject #08), which has already been described (Kanda et al., 2008). The subjects may or may not present left/right observed asymmetries. The degree of hypometabolism appears consistent with the neuropsychological scores. For example, the low degree of anomaly observed for subject #09 can be explained by his high MMSE, BNT and verbal fluency scores, and low FTD-CDR score, compared with the other bvFTD subjects. Subjects with the highest degree of anomaly in the left anterior temporal region are the ones with the lowest verbal fluency score, which is in accordance with the fact that this region has been shown to be critical for semantic abilities (Rogers et al., 2006).

1.5 Discussion

We presented a method for the individual analysis of PET data providing voxel-wise statistics of normality/abnormality. The method consists of creating a subject-specific model of healthy PET appearance and comparing the patient's PET image to the model via a regularised Z-score. We validated the proposed method by generating abnormality maps for healthy controls and subjects at different stages of Alzheimer's disease, and we applied the framework to the analysis of frontotemporal dementia.

We first ensured that the method was able to generate accurate subject-specific models of healthy PET appearance by applying it to normal control subjects and showed that no anomalies were detected for this population (Figures 1.3, 1.4 and 1.5). We then applied it to subjects at different stages of Alzheimer's disease and to subjects with two subtypes of frontotemporal dementia. In both cases, we observed that the abnormality maps obtained with the proposed method coincided with the regions where uptake abnormalities were expected (Figures 1.3, 1.5 and 1.6).

The proposed framework was also validated using the abnormality maps as inputs of a classifier. For comparison, we also used the PET images themselves and standard Z-maps as features. The different approaches produced comparable accuracies for all the tasks tested: differentiating CN from early MCI, late MCI and AD, differentiating early from late MCI, and differentiating amyloid negative subjects from amyloid positive subjects. More accurate results were systematically obtained with the proposed method than with the PET images themselves. These results demonstrated that the proposed approach was able to automatically locate and characterise the areas characteristic of dementia. The

classification results also highlighted the importance of selecting control subjects matching the demographic characteristics of the subject under investigation. This may be due to the fact that age and sex influence brain metabolism, and that these differences are not captured when comparing structural MR images. Note that the objective of the proposed framework was not to generate features leading to a higher classification accuracy, but rather to provide feature maps easily interpretable.

A limitation of the proposed strategy is the computational cost. The bottleneck is the inter-subject registration, which takes on average 18 min per atlas when run on two standard CPUs for a reference image of size $256 \times 256 \times 166$. When the inter-subject registrations are run in parallel for all the atlases, an abnormality map can be generated in less than 25 min. The inter-subject registration step could be accelerated using deep learning-based approaches that can perform pair-wise registration in less than a minute (Vos et al., 2019; Fan et al., 2019; Ding et al., 2019).

The abnormality maps could have two complementary uses. They could i) help clinicians in their diagnosis by highlighting, in a data-driven fashion, the pathological areas obtained from the individual PET data, and ii) improve the interpretability of subsequent analyses, such as computer-aided diagnosis or spatio-temporal modelling.

1.6 Perspectives

The subject-specific model of healthy appearance is composed of two elements, a spatially-varying weighted average and a spatially-varying weighted standard deviation, that are currently obtained by registering the images of a population of healthy controls to the target subject, locally selecting the registered controls that are morphologically the most similar to the target subject, and locally fusing the PET images of the selected controls.

The spatially-varying weighted average could be obtained using other image synthesis approaches, such as deep learning methods (Li et al., 2014; Bi et al., 2017; Sikka et al., 2018; Wei et al., 2018; Wei et al., 2019; Xia et al., 2020), which have been shown to be more accurate than atlas-based methods (Nie et al., 2016; Kläser et al., 2018; Xiang et al., 2018). However, they would not generate the latter in a straightforward manner. The standard deviation is an important component of the model as it decreases the degree of the anomalies detected in areas where the uncertainty is high, reducing the amount of false positives.

Others have used image synthesis for anomaly detection. For example, Ye et al., 2013 proposed a patch-based modality propagation technique to synthesise pseudo-healthy T2-weighted from T1w MR images for tumour detection. A patch-based approach has also been developed by Tsunoda et al., 2014 to synthesise pseudo-normal chest radiographs. Bowles et al., 2017 proposed a regression-based method proposed to synthesise pseudo-healthy FLAIR from T1w images for white matter lesion segmentation. Yang et al., 2016 synthesised pseudo-normal images from images with lesions to improve registration. They used a variational auto-encoder to learn the brain appearance from the normal areas of the pathological images only and estimate the reconstruction uncertainty of the predicted pseudo-normal image. Various auto-encoder models have been explored in (Baur et al., 2019; Choi et al., 2019; Uzunova et al., 2019; You et al., 2019; Zimmerer et al., 2019; Chen et al.,

2020; Baur et al., 2021) for brain lesion detection. Adversarial learning has been used as well to learn mappings between abnormal and normal tissues, and detect tumours on brain MR images (Andermatt et al., 2019; Xia et al., 2020; Sun et al., 2020b) or fluid on optical coherence tomography images of the retina (Schlegl et al., 2019). Generative adversarial networks have also been used to generate pseudo-healthy PET images from T1w MR images to detect hypometabolic lesions in the context of epilepsy (Yaakub et al., 2019). Except for (Yang et al., 2016), these approaches do not estimate the synthesis uncertainty, which is important to assess the significance of the anomalies detected. Note that these approaches have mostly been used to detect well-characterised lesions such as brain tumours (Uzunova et al., 2019; You et al., 2019; Zimmerer et al., 2019; Andermatt et al., 2019; Chen et al., 2020; Xia et al., 2020; Sun et al., 2020b; Baur et al., 2021; Pinaya et al., 2022) or white matter lesions (Baur et al., 2019; Baur et al., 2021; Pinaya et al., 2022) from brain MRI and that very few have tried to detect diffuse lesions such as that characteristic of dementia (Choi et al., 2019; Baydargil et al., 2021).

Future work will consist in implementing such deep learning-based approaches with the hope of detecting subtler anomalies, and in a more computationally efficient manner, compared with the current approach. Two strategies will be explored, based on the use of auto-encoders and generative adversarial networks, respectively. These two strategies mainly differ in the images they use as input and it is not clear from the current literature which would perform best.

1.6.1 Anomaly detection using autoencoders

Autoencoders are a type of generative models that try to learn a compressed representation of the input data, here images. In the encoder network, the input data pass through a series of neural network layers that output a low dimensional set of latent variables. To learn this latent variable set, a decoder network, also composed of a series of neural network layers, is used to reconstruct a resemblance of the original data from the latent variables. The result is an imperfect reconstruction of the input. Autoencoders are trained with the objective to minimise the difference between the input and reconstructed images.

The idea is to use autoencoders to learn the distribution of the healthy brain, which should enable the model to fully reconstruct healthy brain areas while failing to reconstruct areas with anomalies in images of a diseased brain (Baur et al., 2019; Uzunova et al., 2019). During the training phase, only images of healthy subjects would be used. The training data set should be as large and diverse as possible to capture as best as possible the diversity that exists in the healthy population. During the application phase, the image of a patient would be fed to the encoder-decoder network, which would only know how to reconstruct healthy images. As a result, the reconstructed image would be a pseudo-healthy representation of the input image. Comparing the input image and its pseudo-healthy reconstruction would highlight the areas of the brain presenting anomalies, which would be presented in the form of an abnormality map of the same dimensions as the input image. This framework is displayed in Figure 1.7.

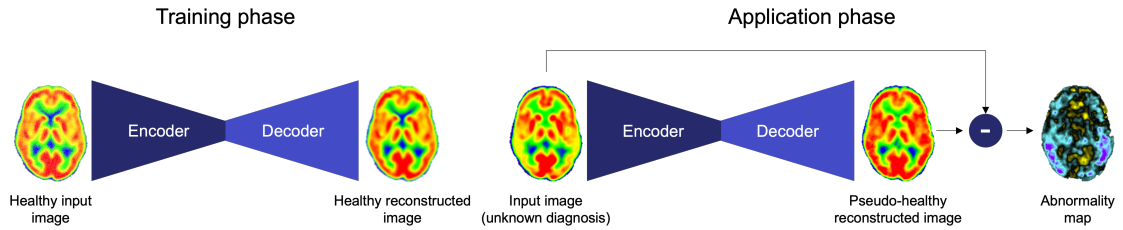


FIGURE 1.7: Anomaly detection framework using autoencoders

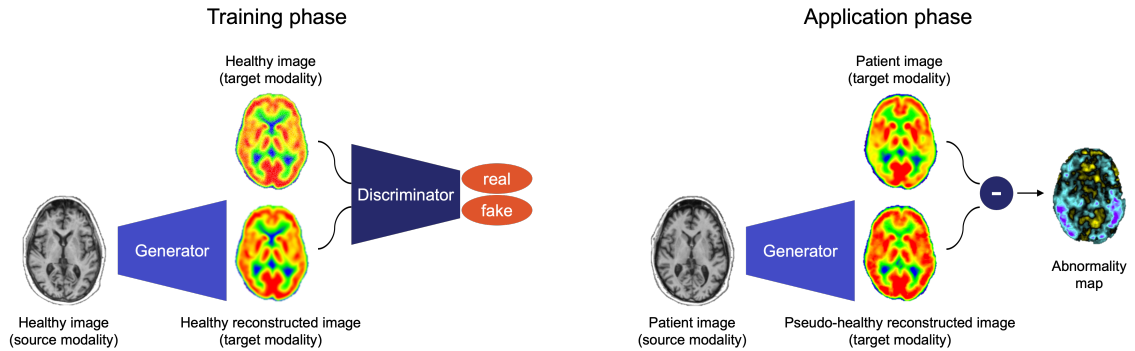


FIGURE 1.8: Anomaly detection framework using conditional generative adversarial networks (cGANs)

1.6.2 Anomaly detection framework using conditional generative adversarial networks

Anomaly detection could also be performed with the use of conditional generative adversarial networks (cGANs) (Yaakub et al., 2019). cGANs are composed of two neural networks, the generator and the discriminator, that are competing against each other. The generator network is trained to learn a transformation going from an image of a source modality (e.g. T1w MRI) to an image of a target modality (e.g. PET). This generated fake image must be as realistic as possible. Then the discriminator network learns to distinguish between fake and real images. By having these two networks competing against each other, the discriminator is forced to become as good as possible at distinguishing fake and real, which in turns forces the generator to become as good as possible at generating images that are as realistic as possible to try to fool the discriminator. After training, the generator network can be used to reconstruct an image of the target modality from an image of the source modality.

As for the framework using autoencoders, only healthy subjects would be used during the training phase. The cGAN would learn to generate healthy images (e.g. PET images) for a specific anatomy obtained from a structural imaging modality (e.g. T1w MRI). During the application phase, the structural image of a patient would be given as input. As the generator would only know how to reconstruct healthy images, the generated image would be a pseudo-healthy representation, specific to the patient's anatomy. As previously, comparing the real image of the patient and the pseudo-healthy reconstruction would highlight the areas of the brain presenting anomalies. This framework is displayed in Figure 1.8.

1.6.3 Uncertainty estimation of the reconstruction

The current deep learning based anomaly detection approaches (Schlegl et al., 2019; Baur et al., 2019; Uzunova et al., 2019; Yaakub et al., 2019) do not model the uncertainty of the reconstruction, which is crucial for an application to the clinic (Barbano et al., 2022). Uncertainty measures can provide information as to how confident the model was on reconstructing pseudo-healthy images. Voxel-wise uncertainty maps would be useful to detect potential defects in a specific area of the image while subject-level uncertainty, for example obtained by aggregating voxel-wise uncertainties, would provide a global measure of success. Several general-purpose methods developed to assess the uncertainty in deep learning models (Gal et al., 2016; Kendall et al., 2017; Lakshminarayanan et al., 2017) have been applied to a medical imaging context (Jungo et al., 2019; Barbano et al., 2022). The general idea would be to generate several pseudo-healthy reconstructions for a given input to obtain a variance estimate. This could be done by training a single model and repeatedly ignoring part of the neurons in the application phase (Gal et al., 2016), or by training several models and reconstructing a pseudo-healthy image for each of these models in the application phase (Lakshminarayanan et al., 2017). The voxel-wise uncertainty maps could be presented to clinicians along with the abnormality maps or they could be used to modulate the abnormality maps (i.e. the abnormality maps could be computed by subtracting the real patient image by the pseudo-healthy reconstruction and subsequently dividing by the uncertainty map, which would mimic a Z-score).

Chapter 2

Interpretable computer-aided diagnosis of dementia

This chapter results from the PhD work of Elina Thibeau-Sutre, who I supervised with Didier Dormont and Olivier Colliot. Corresponding publications:

- Thibeau-Sutre, E., Colliot, O., Dormont, D., **Burgos, N.**: ‘Visualisation Approach to Assess the Robustness of Neural Networks for Medical Image Classification’. In *SPIE Medical Imaging 2020*, 11313: 113131J, 2020. [doi:10.1117/12.2548952](https://doi.org/10.1117/12.2548952) • [hal-02370532](https://hal.archives-ouvertes.fr/hal-02370532)
- Thibeau-Sutre, E., Collin, S., **Burgos, N.**, and Colliot, O.: ‘Interpretability of Machine Learning Methods Applied to Neuroimaging’. In *Machine Learning for Brain Disorders*, edited by Colliot O., Springer. To be published in 2022. [hal-03615163](https://hal.archives-ouvertes.fr/hal-03615163)

2.1 Introduction

The anomaly detection approach described in the previous chapter is by construction interpretable. This is not the case of all computer-aided diagnosis tools, especially those relying on deep learning classifiers. The users of such tools (patients and clinicians) may want more information than simply a prediction performance before relying on such systems. They may want to know on which features the model is relying to reach a particular results, but also whether these features are close to the way a clinician thinks.

According to Lipton, [2018](#), model interpretability can be broken down into two categories: transparency and post-hoc explanations. A model can be considered as transparent when it (or all parts of it) can be fully understood as such, or when the learning process is understandable. A natural and common candidate that fits, at first sight, these criteria is the linear regression algorithm, where coefficients are usually seen as the individual contributions of the input features, but also decision trees, where model predictions can be broken down into a series of understandable operations. However, one may need to be cautious about the real interpretability allowed by these models. Indeed, in some cases a feature may have not been kept by the model, but this does not mean that it is not associated with the

target. This is the case for example for sparse models like LASSO, but also multiple non-regularized linear regressions. Moreover, features given as input to transparent models are often highly-engineered, and choices made before the training step (preprocessing, feature selection) may also hurt the transparency of the whole framework. The second category of interpretability methods, post-hoc interpretations, allows dealing with non-transparent models. Xie et al., 2020 proposed a taxonomy in three categories: *visualisation* methods consist in extracting an attribution map of the same size as the input whose intensities allow knowing where the algorithm focused its attention, *distillation* approaches consist in reproducing the behaviour of a black-box model with a transparent one, and *intrinsic* strategies include interpretability components within the framework, which are trained along with the main task (for example, a classification).

Our work focuses on post-hoc interpretability, which is the category the most used nowadays as it allows interpreting deep learning methods that were recently adapted to neuroimaging studies. We proposed a new taxonomy inspired from that of Xie et al., 2020 but including other methods of interpretation (Thibeau-Sutre et al., 2022a):

1. **weight visualisation** consists in directly visualising weights learned by the model, which is natural for linear models but quite less informative for deep learning networks,
2. **feature map visualisation** consists in displaying intermediate results produced by a deep learning network to better understand its operation principle,
3. **back-propagation methods** are back-propagating a signal through the machine learning system from the output node of interest o_c to the level of the input to produce an attribution map,
4. **perturbation methods** evaluate the difference in performance between an original input and its locally perturbed versions to infer which parts of the input is relevant for the machine learning system,
5. **distillation** approximates the behaviour of a black-box model with a more transparent one, and then draw conclusions from this new model,
6. **intrinsic** corresponds to non post-hoc explanations: in this case, interpretability is obtained thanks to components of the framework that are trained at the same time as the model.

The following sections describe how we adapted the perturbation method of Fong et al., 2017 to 3D medical images to find on which basis a network classifies quantitative data. Quantitative data can be obtained from different medical imaging modalities, for example binding potential maps obtained with positron emission tomography (PET) or grey matter probability maps extracted from structural magnetic resonance imaging (MRI). Our application focuses on the detection of Alzheimer’s disease (AD), which induces grey matter atrophy. We used as inputs grey matter probability maps, a proxy for atrophy, extracted from T1-weighted (T1w) MRI. The process includes two distinct parts: first a convolutional neural network (CNN) is trained to classify AD from control subjects, then the weights of the network are fixed and a mask is trained to prevent the network from classifying correctly

all the subjects it has correctly classified after training. The goals of this work were to assess whether the interpretability method initially developed for natural images was suitable for 3D medical images and could be exploited to better understand the decisions taken by classification networks.

2.2 Materials and methods

2.2.1 Data description and preprocessing

We used T1w MR images from two public data sets: the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database and the Australian Imaging, Biomarkers and Lifestyle (AIBL) study. Two diagnosis groups were considered:

- CN: sessions of subjects who were cognitively normal (CN) at baseline and stayed stable during the follow-up;
- AD: sessions of subjects who were diagnosed as AD at baseline and stayed stable during the follow-up.

The populations of ADNI and AIBL are described in Table 2.1.

TABLE 2.1: Summary of ADNI and AIBL participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores at baseline. Values are presented as mean (standard deviation) [range]. M: male, F: female

Data set	Label	Subjects	Sessions	Age	Gender	MMSE	CDR
ADNI	CN	330	1 830	74.4 (5.8) [59.8, 89.6]	160 M / 170 F	29.1 (1.1) [24, 30]	0: 330
	AD	336	1 106	75.0 (7.8) [55.1, 90.9]	185 M / 151 F	23.2 (2.1) [18, 27]	0.5: 160; 1: 175; 2: 1
AIBL	CN	429	730	72.5 (6.2) [60, 92]	183 M / 246 F	28.8 (1.2) [25, 30]	0: 406, 0.5: 22, 1: 1
	AD	76	108	73.9 (8.0) [55, 93]	33 M / 43 F	20.6 (5.5) [6, 29]	0.5: 31; 1: 36; 2: 7, 3: 2

Preprocessing of T1w MR images was performed with the Clinica software platform (Routier et al., 2021). First the data sets were converted to the BIDS format, then the `t1-volume` preprocessing pipeline of Clinica was applied (Samper-González et al., 2018). This pipeline performs bias field correction, non-linear registration and tissue segmentation using the Unified Segmentation approach (Ashburner et al., 2005) available in SPM12. The grey matter maps in MNI space were retrieved for the image analysis.

2.2.2 CNN classification

The following sections describe the evaluation procedure, the hyperparameters selection and implementation details that are linked to the classification of AD vs CN subjects with CNNs. During training, the weights and biases of the network are optimised to maximise the score function f on a set of images X .

2.2.2.1 Evaluation procedure

The ADNI data set was split into training/validation and test sets. The ADNI test set consisted of 100 randomly chosen age- and sex-matched subjects for each diagnostic class (i.e. 100 CN subjects, 100 AD patients). The rest of the ADNI data set was used as training/validation set. We ensured that age and sex distributions between training/validation and test sets were not significantly different. The model selection procedure, including model architecture selection and training hyperparameter fine-tuning, was performed using only the training/validation data set. For that purpose, a 5-fold cross-validation was performed, which resulted in one fold (20%) of the data for validation and the rest for training. Note that the 5-fold data split was performed only once for all the experiments with a fixed seed number ($random_state = 2$), thus guaranteeing that all the experiments used exactly the same subjects during cross-validation. The AIBL data set was used as an independent test set to assess the CNN generalisation ability. Test and validation sets included only one session per subject.

2.2.2.2 Hyperparameter selection

We performed a random search (Bergstra et al., 2012) to select the architecture and optimisation hyperparameters of our CNN. The hyperparameters explored for the architecture were the number of convolutional blocks, of filters in the first layer and of convolutional layers in a block, the dimension reduction strategy (by using a max pooling layer or by setting the stride of the last convolutional layer of the convolutional block to 2), the number of fully-connected layers and the dropout rate. Other hyperparameters such as the learning rate, the weight decay, the batch size, the data preprocessing and the intensity normalisation were also part of the search.

Only one experiment was performed per architecture tested using the first split of the cross-validation due to the computational cost of the random search. The chosen architecture was the one that obtained the best balanced accuracy on the validation set. This architecture (displayed in Figure 2.1) is composed of seven convolutional blocks followed by a dropout layer and a fully-connected layer. Each convolutional block (C1, C2 or C3) is made of 1 to 3 sub-blocks and a max pooling layer with a kernel size and a stride of 2. Each sub-block is composed of a convolutional layer with kernel size of 3, a batch normalisation layer and a leaky ReLU activation. The predicted label of the input image is the class with the highest output probability.

2.2.2.3 CNN training

The weights of the convolutional and fully connected layers were initialised as described in (He et al., 2015), which corresponds to the default initialisation method in PyTorch. We applied the following early stopping strategy for all the classification experiments: the training procedure does not stop until the validation loss is continuously higher than the lowest validation loss for N epochs ($N=5$); otherwise, the training continues to the end of a pre-defined number of epochs (30). The training and validation loss were computed with the cross-entropy loss. For each experiment, the final model was the one that obtained the

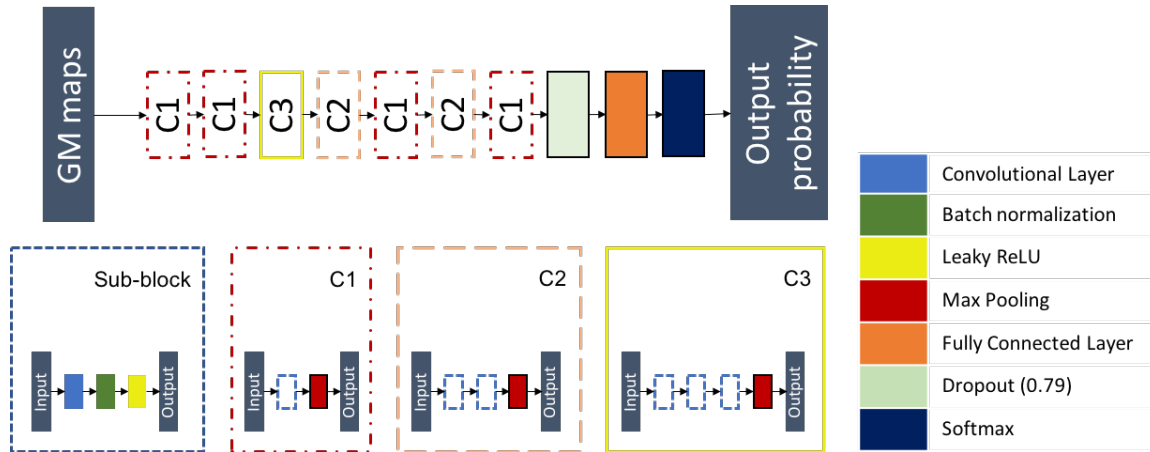


FIGURE 2.1: Architecture of the CNN classifier determined following to a random search procedure

highest validation balanced accuracy during training. The balanced accuracy of the model was evaluated at the end of each epoch.

2.2.3 Interpretability method

The proposed interpretability method extends the framework of (Fong et al., 2017). Once the classification network has been trained, its parameters are fixed to the best value found. Then the interpretability method consists in computing a mask that will overlay the most meaningful parts of an image to prevent the network from classifying it correctly. In the following, the goal is to mask AD images that were correctly classified by the CNN so that it systematically classifies them with the CN label. The mask m is a 3D volume of the same size as the input image and hide parts of the image in a voxel-wise manner. In this application, each voxel u of the input image X will be masked by a constant value μ according to the value of the mask for this voxel. The mask values are included in $[0, 1]$. The masked input image X^m at voxel u is defined as:

$$X^m(u) = m(u)X(u) + (1 - m(u))\mu \quad (2.1)$$

As AD patients suffer from grey matter atrophy, the goal of the masking method would be to artificially simulate grey matter restoration in a minimal number of brain regions to make them look like CN subjects. By setting $\mu = 1$, the mask was trained to artificially increase the probability of grey matter for the minimum set of voxels which will lead to the maximum decrease of the performance of the CNN. The optimal mask m^* is the mask for which the following loss function is minimised:

$$m^* = \underset{m}{\operatorname{argmin}} f(X^m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \sum_u \|\nabla m(u)\|_{\beta_2}^{\beta_2} \quad (2.2)$$

The first term prevents the network from finding the correct class when the mask is applied, the second term ensures that a minimum set of voxels is selected, and the third one ensures that the mask is smooth enough and is not made of scattered voxels.

Once the mask training is finished, values above 0.95 are set to 1. This ensures that the CNN is only perturbed by the zones identified by the mask, and not by the small gradients that can be found on all the surface of the mask.

2.2.3.1 Quality check procedure

As the interpretability method is very sensitive to outliers when applied to a group of images, a quality check procedure was performed before mask optimisation. This quality check includes two steps:

1. The grey matter maps were sorted in increasing order by their maximal value. Images with a maximal value lower than 0.95 were automatically rejected. Eight sessions were removed during this procedure.
2. One image was removed after training a group mask. During the training of the first group mask this session led to a significant increase of the loss. This image was removed as it suffered from defects (the eyes were segmented as grey matter).

2.2.3.2 Grid search on interpretability hyperparameters

A grid search was performed to choose the set of hyperparameters linked to the computation of the mask: the coefficients for the regularisation λ_1 , λ_2 , β_1 , β_2 in equation 2.2. The learning rate was arbitrarily fixed to 0.1. The grid search was only performed on the group level masking for AD label.

2.2.3.3 Group level masking

To find the regions most related to Alzheimer’s disease according to the CNN, we computed a mask based on the subset of images of AD subjects that were in the training set and validation set and all correctly classified by the network. For AIBL, the subset of all AD sessions correctly classified by the network was used. We exploit the fact that a voxel-wise correspondence exists between the grey matter maps, thanks to the non-linear registration, to iteratively build a group mask: the mask is initialised with all its values set to 1 and is then updated each time with a different image. The subset of well classified AD of the validation set used for the CNN training was again used as a validation set for mask optimisation. At the end of each epoch, the masking loss was evaluated on this set to save the best mask according to the validation masking loss. To assess the robustness of the CNN training, the masks obtained for different folds (i.e. different input images and initialisations) and different runs of the same folds (i.e. same input images but different initialisations) were compared by pairs. The mean value of all pairs gave the similarity between folds or runs of the same fold.

2.2.3.4 Session level masking

Masks were also produced at the session level based on a single image. To avoid the overfitting risk due to the use of only one image instead of a set of images as in the previous section, the regularisation terms λ_1 and λ_2 were multiplied by 100. No validation set was

used for these experiments as the goal is precisely to fit the individual pattern of one image instead of finding a general pattern that may correspond to a group of images. Session level experiments include a longitudinal and cross-sectional analysis. In the longitudinal analysis, all the sessions of one subject were compared by pairs and the mean value of these comparisons gave the intra-subject similarity for this subject. The mean intra-subject similarity is then the mean value of the intra-subject similarity of all AD subjects. For the cross-sectional analysis, the mean value of all pairwise comparisons of baseline sessions of all AD subjects gave the inter-subject similarity measure. These analyses are performed to assess the stability of the interpretability method and provide a baseline value by using the inter-subject similarity for the different metrics.

2.2.3.5 Interpretability method training

The mask was initialised with a matrix of the same size as the input images ($121 \times 145 \times 121$) full of ones. We applied a similar early stopping strategy than for the classification experiments: the training procedure for group level masking on ADNI does not stop until the relative difference between the validation loss and the lowest validation loss is superior to a tolerance of 0.05 for a patience of N epochs ($N=5$); otherwise, the training continues to the end of a pre-defined maximum number of epochs (150). For the group level masking on AIBL, the patience was increased to 25 and the maximum number of epochs to 300 as the number of AD subjects is smaller than for ADNI. For the session level masking, the patience was increased to 200 and the maximum number of epochs to 5,000, while the tolerance was decreased to 0.01. The loss corresponds to the argmin argument of equation 2.2. For each experiment, the final mask was the one that obtained the lowest validation loss during training. The loss of the mask was evaluated at the end of each epoch.

2.2.4 Metrics of evaluation

The similarity between masks was evaluated in two ways. The output probabilities of the CNN for the true class (prob_{CNN}) for an input masked by two masks optimised in two different contexts (e.g. different runs for the group level masking, different sessions of the same subject for the session level masking) are used to establish a comparison based on the CNN perception of the input. A mean output probability close to 1 means that the first model is not perturbed by the mask optimised for the second model, meaning that the two models are dissimilar. A ROI-based similarity was also computed to assess the similarity of two masks according to the 120 regions-of-interest (ROIs) of the AAL2 atlas (Rolls et al., 2015). For each ROI, 1 minus the sum of the values in the ROI is computed, resulting in a ROI-vector of size 120 for each mask. Each value in the ROI-vector represents the density of the mask in the associated ROI. The ROI-based similarity between two masks is then the cosine similarity of two ROI-vectors. A value close to 1 means that the densities of the masks are the same between the ROIs, a value close to 0 means that the locations of the masks have no intersection.

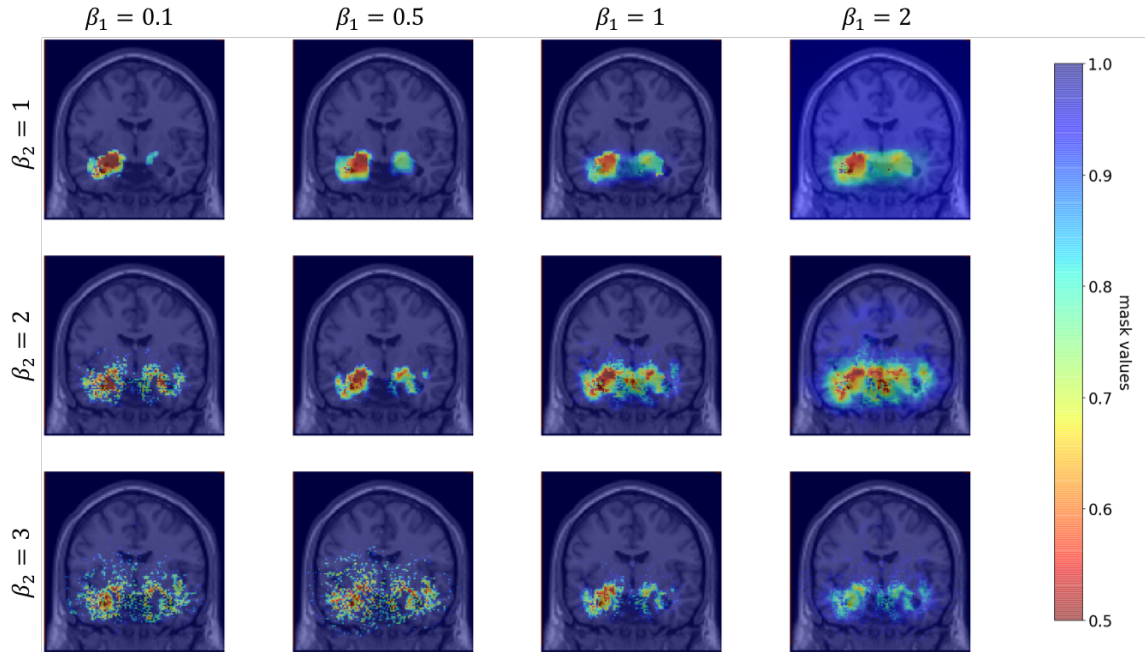


FIGURE 2.2: Comparison of masks obtained for different values of the interpretability hyperparameters β_1 and β_2

2.3 Results

Once the architecture was chosen and the CNN was trained on all folds, the classification performance was evaluated on the independent test set to ensure the absence of overfitting. The validation balanced accuracies on the five folds were 0.95, 0.82, 0.96, 0.85 and 0.87, giving an average of 0.89. The test balanced accuracies on the five folds were 0.89, 0.87, 0.90, 0.86 and 0.87, giving an average of 0.88. Moreover, the balanced accuracies obtained on the independent test set AIBL were 0.85, 0.92, 0.91, 0.92 and 0.92, giving an average value of 0.90. We could thus conclude that the network was not overfitting and we could use it for the interpretability task.

2.3.1 Grid search on interpretability hyperparameters

First the hyperparameters β_1 and β_2 were chosen with fixed $\lambda_1 = 0.0001$ and $\lambda_2 = 0.001$. The choice of the values $\beta_1 = 0.1$ and $\beta_2 = 1$ was made based on visual inspection. We observe on Figure 2.2 that when β_1 decreases, the minimal value of the mask decreases and this prevents from producing a mask with a large set of values close but different from 1. When β_2 increases, the value of the second term becomes negligible before the first term of equation 2.2. This leads to a very scattered mask as it is dominated by the first term of the regularisation.

The hyperparameters λ_1 and λ_2 were then chosen with fixed $\beta_1 = 0.1$ and $\beta_2 = 1$. The choice of the values $\lambda_1 = 0.0001$ and $\lambda_2 = 0.01$ was made based on visual inspection and the stability of the loss during mask training. We observe on Figure 2.3 that when λ_1 increases, the surface covered by the mask decreases until it only becomes scattered points. When λ_2 increases, the surface covered by the mask increases.

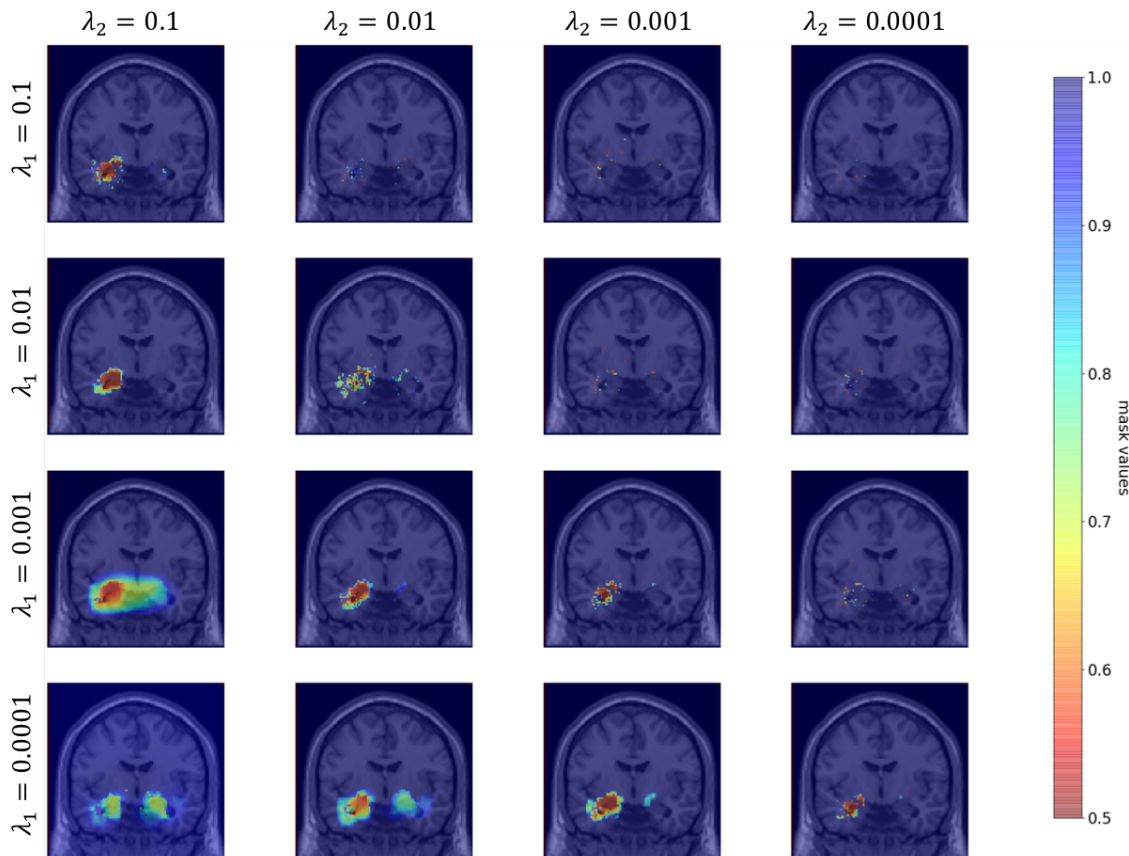


FIGURE 2.3: Comparison of masks obtained for different values of the interpretability hyperparameters λ_1 and λ_2 . Note that for $\lambda_1 = 0.0001$ and $\lambda_2 = 0.1$, the learning rate was fixed to 0.01 as the mask optimisation did not converge with a learning rate of 0.1.

2.3.2 Robustness of the interpretability method

Different experiments were conducted to assess whether the method was robust enough to help interpret the results of the CNN. Indeed Adebayo et al., 2018 highlighted that some interpretability methods developed to interpret the results of neural networks (for example guided back-propagation and guided grad-CAM) did not depend on model parameters, as they gave the same result with a pre-trained network or a randomised one. Hence we need to check if the interpretability method gives coherent results based on the CNN training.

2.3.2.1 Group level masking

This experiment aims to assess the coherence of the proposed interpretability approach with the a priori knowledge of the disease. One mask was optimised for each of the five models trained on the five folds of the cross-validation. Though these masks do not always overlap, they focus on a set of ROIs known to be particularly affected during AD progression. To confirm this visual observation, the list of the 5 ROIs in which the mask has the lowest values was extracted for each fold. All masks include in this list at least one hippocampus and parahippocampal gyrus. Moreover, the fusiform gyri (4 masks out of 5) and the amygdalae (3 masks out of 5) are frequently highlighted by the masks. Other regions such as the

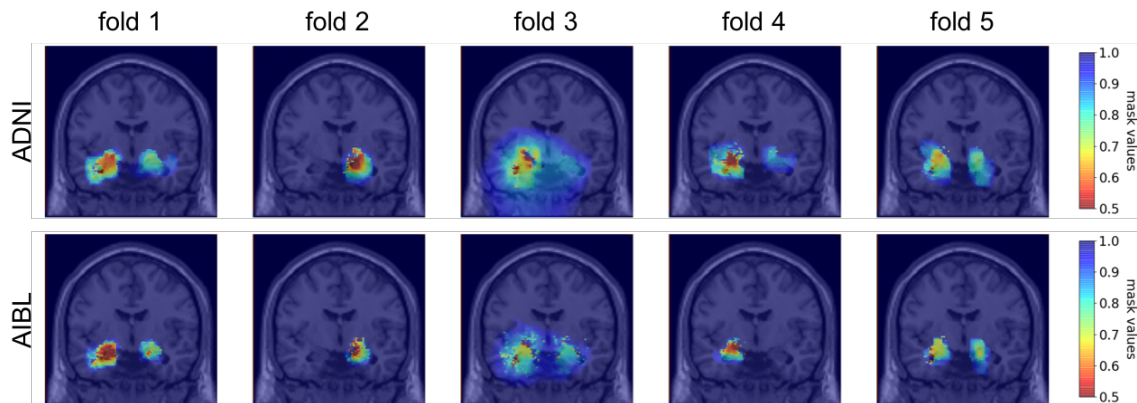


FIGURE 2.4: Coronal view of the group masks trained on ADNI (first line) and AIBL (second line). Each column corresponds to a model trained on one fold of the cross-validation on training/validation ADNI set.

putamen, the pallidum, the inferior temporal gyrus and the thalamus appear only once in these lists.

Moreover, to assess the robustness of the method towards data used for mask optimisation we compared the masks obtained by applying the interpretability method on ADNI or AIBL data using the five networks trained on the five folds of the cross-validation on ADNI training/validation set. The corresponding masks are displayed on Figure 2.4. The ROI-based similarities between the pairs of masks were 0.92, 0.99, 0.93, 0.89 and 0.97. These are comparable to the intra-subject ROI-based similarity (0.94). The prob_{CNN} dissimilarities were very small as all the dissimilarities were smaller than 10^{-3} . This indicates that for a given pre-trained network, a mask optimised for images of ADNI (resp. AIBL) correctly occludes the images of AIBL (resp. ADNI). However, the comparison of the masks in this way may not be completely fair. Even though the number of epochs and the patience of the early stopping procedure were increased for AIBL masking, the masks on ADNI and AIBL did not benefit from the same number of iterations. This factor leads to masks that comprise a different number of points, as the effect of the regularisation terms is correlated to the number of iterations. This means that though the masks highlight the same locations in the brain, the difference in regularisation makes the masks more dissimilar than they would be if we could find an equivalence for the hyperparameters that control the number of epochs (patience, tolerance and maximum number of epochs). Hence the dissimilarity here may not be due to the difference of data sets, but to the different number of iterations done during the mask optimisation.

Finally, as the method is not deterministic we computed ten times the mask on the first fold of the cross-validation to ensure that the mask optimisation is robust to rerun. We obtained high ROI-based similarities for all pairs of runs (≥ 0.97) with a mean similarity of 0.99.

2.3.2.2 Session level masking

The inter-subject similarity and dissimilarity were evaluated to 0.80 and 0.58 for the ROI-based and the prob_{CNN} metrics respectively. The intra-subject similarity and dissimilarity

were evaluated to 0.94 and 0.11 for the ROI-based and the prob_{CNN} metrics respectively. The higher intra-subject similarity compared to the inter-subject similarity ensures that the interpretability metric is robust as the same pattern is generated for different sessions of the same subject.

2.3.2.3 Similarity across hyperparameters

With the ROI-based similarity, we can assess whether the masks produced by varying one hyperparameter value are similar. To observe this similarity, we reused the same masks as those produced for the grid search (see Figure 2.2 and 2.3).

First, the similarities using different β_1 and β_2 values were computed with fixed $\lambda_1 = 0.0001$ and $\lambda_2 = 0.001$ and a learning rate of 0.1. The similarities between masks produced with β_1 values in 0.1, 0.5, 1, 2 and fixed $\beta_2 = 1$ are given in Table 2.2a. As expected when looking at the masks obtained in Figure 2.2, the masks are highly similar except for the value $\beta_1 = 2$ for which the first regulation term became negligible in front of the second regulation term in equation 2.2. It resulted in a very smooth mask which is dense in all regions of the brain as 97.5% of values are below 0.95. This explains why the ROI-based similarity is so low between this mask and the others, though the regions identified seem similar at visual inspection. Other masks have a high similarity (> 0.95 in all cases). The similarities between masks produced with β_2 values in 1, 2, 3 and fixed $\beta_1 = 0.1$ are given in Table 2.2b. There is more variability for this hyperparameter, though the similarity between two consecutive values is still high (> 0.90).

TABLE 2.2: Similarity across different β_1 and β_2 values

	$\beta_1 = 0.1$	$\beta_1 = 0.5$	$\beta_1 = 1$	$\beta_1 = 2$
$\beta_1 = 0.1$		0.98	0.97	0.32
$\beta_1 = 0.5$	0.98		1.00	0.36
$\beta_1 = 1$	0.97	1.00		0.37
$\beta_1 = 2$	0.32	0.36	0.37	

(A) Similarity across different β_1 with fixed $\beta_2 = 1$

	$\beta_2 = 1$	$\beta_2 = 2$	$\beta_2 = 3$
$\beta_2 = 1$		0.90	0.70
$\beta_2 = 2$	0.90		0.91
$\beta_2 = 3$	0.70	0.91	

(B) Similarity across different β_2 with fixed $\beta_1 = 0.1$

The similarities using different λ_1 and λ_2 values were then computed with fixed $\beta_1 = 0.1$ $\beta_2 = 1$ and a learning rate of 0.1. For both hyperparameters the similarity is high between two consecutive values (>0.90), as can be seen in Table 2.3.

These results highlight the stability of the method toward the hyperparameters choice, as two consecutive hyperparameter values led to two masks with a ROI-based similarity superior to the inter-subject similarity (0.80). Moreover, all the masks involved in this section analysis correctly occlude the CNN perception: for all masks, the mean output probability of the AD class on the validation data set is below 10^{-6} .

TABLE 2.3: Similarity across different λ_1 and λ_2 values

	$\lambda_1 = 0.1$	$\lambda_1 = 0.01$	$\lambda_1 = 0.001$	$\lambda_1 = 0.0001$
$\lambda_1 = 0.1$		0.93	0.84	0.83
$\lambda_1 = 0.01$	0.93		0.95	0.91
$\lambda_1 = 0.001$	0.84	0.95		0.91
$\lambda_1 = 0.0001$	0.83	0.91	0.91	

(A) Similarity across different λ_1 with fixed $\lambda_2 = 0.01$

	$\lambda_2 = 0.1$	$\lambda_2 = 0.01$	$\lambda_2 = 0.001$	$\lambda_2 = 0.0001$
$\lambda_2 = 0.1$		0.98	0.85	0.72
$\lambda_2 = 0.01$	0.98		0.92	0.82
$\lambda_2 = 0.001$	0.85	0.92		0.96
$\lambda_2 = 0.0001$	0.72	0.82	0.91	

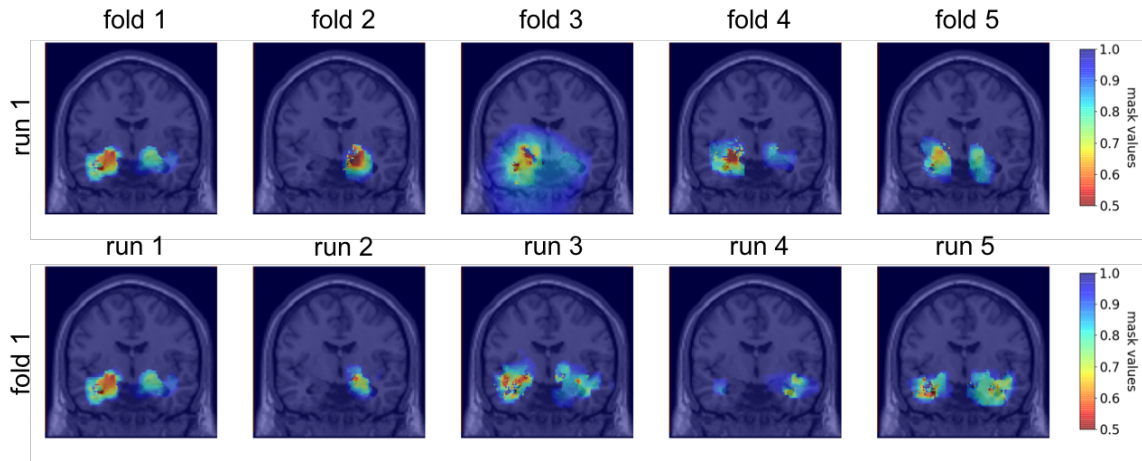
(B) Similarity across different λ_2 with fixed $\lambda_1 = 0.0001$ 

FIGURE 2.5: Coronal view of the group masks obtained for the five folds of the cross-validation on the first run (first line) and of the group masks obtained for five runs of the first fold (second line).

2.3.3 Robustness of the CNN training

After having assessed the robustness of the interpretability method, we applied it to better understand the factors influencing the training process of the CNN classifier based on several scenarios: for different folds (different initialisation, different training/validation split) and different runs (different initialisation, same training/validation split). Figure 2.5 displays the masks obtained for the five folds and five runs of the first fold.

With the prob_{CNN} metric using the validation set, the dissimilarity between folds and 5 runs of the same fold are equivalent with respectively 0.78 and 0.82. The similarities computed with the ROI-based metric are also equivalent with respectively 0.65 and 0.69 between folds and between runs of the same fold. This indicates that the impact of the distribution of the data between training and validation is minimal compared to the initialisation of the CNN and the training process. Moreover, we observe that the dissimilarity between folds / runs of the same fold is higher than the intersubject dissimilarity obtained with session level masking. This could mean that the regions on which the CNN relies on to identify the

diagnosis mainly depend on the initialisation and the training process and that the CNN training is not robust towards the regions identified.

2.4 Discussion

We extended an interpretability method to 3D medical imaging data and used it to better understand the decisions made by a classification network.

We first assessed the robustness of the proposed interpretability approach. We showed that it gave coherent results as the regions identified by the mask are representative of AD (Whitwell et al., 2007). This coherence is also confirmed by the fact that the intra-subject similarity is higher than the inter-subject similarity. Moreover, the high similarity across neighbouring values of the hyperparameters of the masking method indicates that this method is stable towards hyperparameter selection. Finally, we assessed that the method appears robust towards the data used for the construction of group masks by comparing masks computed using ADNI and AIBL data sets.

We then applied the interpretability approach to assess the robustness of the CNN training. We demonstrated that even if the classification performance on the test set is very similar between the different folds, the training of the CNN for our application is not robust as the inter-subject similarity for one training is higher than the similarity between two retrainings or two folds of the network. This problem of robustness in CNN training may exist for many medical applications in which the number of samples is not sufficient for the network to learn stable meaningful features. This means that it may not be possible to study individual variations using interpretability methods on deep learning applied to imaging data. This problem might be resolved by using more samples and with a better initialisation, given for example by an autoencoder pre-training. Moreover, we found that the regions identified by the networks were very small (restricted to the hippocampus, amygdala and part of the temporal lobe) and that most of the image is not exploited by the CNN to find the diagnosis. This focus of the CNN on the hippocampi only may be partly due to the data set: ADNI is a research cohort from which patients with multiple phenotypes are excluded, leading to a very homogeneous cohort in which the main symptom of the patients is memory loss. It does not fully represent the diversity of AD phenotype, and they may be biased towards hippocampus atrophy.

The interpretability method we used has several limitations. First, the quality check of the data is crucial otherwise the training of the group level mask is not stable. Second, our method is only meaningful for quantitative data: for T1w MRI it would not have been sensible to increase the value of the voxels as it would have deformed the image in a non-meaningful way. This is an issue as the advantage of deep learning is precisely to be able to adapt to the rawest data possible. Finally, though we explored the effect of four hyperparameters of the interpretability method (β_1 , β_2 , λ_1 , λ_2) we did not conduct an exhaustive study on the impact of the learning rate and the number of epochs performed (correlated to the patience, the tolerance and the maximum number of epochs). As we have seen when comparing masks trained on ADNI and AIBL, these parameters impact the amount of regularisation of the masks.

2.5 Perspectives

As we mentioned at the beginning of this chapter, current machine learning methods have a limited interpretability and appear as a ‘black-box’ to clinicians, which hinders their adoption in clinical routine. It is particularly the case for deep learning approaches, as the features are learned automatically by the algorithm and are often hidden from the user. We have seen that even though strategies exist to interpret neural networks, their ability to precisely and robustly interpret decision appears limited. This could of course evolve positively in the future with the development of new approaches. These would need to be thoroughly evaluated, both in controlled settings using simulated data and in realistic scenarios, as proposed in the PhD work of Elina Thibeau-Sutre (Thibeau-Sutre, 2021, Chapter 5).

Another limitation is that current classification algorithms, usually developed for two-class problems, are too rigid to be successfully applied in clinical setting where patients may have several pathologies or a pathology that has never been seen during training, and where more data become available every day. Search and retrieval of brain images is a promising strategy to overcome the ‘black-box’ effect and handle multiple pathologies, thus facilitating adoption by clinicians. A future area of research could consist in building new decision support systems based on methods that can, for a given patient, retrieve similar clinical cases from existing databases in an efficient manner. Similarity between cases could for example be obtained by comparing the abnormality maps generated for each patient as described in Chapter 1. Such patient retrieval framework could be used to diagnose a new patient by attributing the diagnosis of the most similar patient retrieved from the database or the most common diagnosis if several patients are retrieved. It would also offer the opportunity to clinicians to visualise the images retrieved and thus to evaluate themselves the ability of the method to retrieve similar cases, which should increase their trust in the proposed tool for guided diagnosis and its adoption in clinical practice. The ability to visualise the images retrieved could be exploited to create a tool for training new radiologists or assist non-specialists in their diagnosis.

Chapter 3

Reproducible computer-aided diagnosis of dementia

This chapter results from the PhD works of Jorge Samper-González, who I co-supervised with Olivier Colliot and Theodoros Evgeniou, and Elina Thibeau-Sutre, who I co-supervised with Didier Dormont and Olivier Colliot. Corresponding journal publications:

- Samper-González, J., **Burgos, N.**, Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., Colliot, O.: ‘Reproducible Evaluation of Classification Methods in Alzheimer’s Disease: Framework Application to MRI PET Data’. *NeuroImage*, 183: 504–521, 2018. [doi:10.1016/j.neuroimage.2018.08.042](https://doi.org/10.1016/j.neuroimage.2018.08.042) • [hal-01858384](https://hal.archives-ouvertes.fr/hal-01858384)
 - Wen*, J., Thibeau-Sutre*, E., Samper-González, J., Routier, A., Bottani, S., Durrleman, S., **Burgos, N.**, Colliot, O.: ‘Convolutional Neural Networks for Classification of Alzheimer’s Disease: Overview Reproducible Evaluation’, *Medical Image Analysis*, 63: 101694, 2020 (*: joint first authorship). [doi:10.1016/j.media.2020.101694](https://doi.org/10.1016/j.media.2020.101694) • [hal-02562504](https://hal.archives-ouvertes.fr/hal-02562504)
 - Routier, A., **Burgos, N.**, Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., Wen, J., Wild, A., Habert, M.-O., Durrleman, S., Colliot, O.: ‘Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies’. *Frontiers in Neuroinformatics*, 15: 39, 2021. [doi:10.3389/fninf.2021.689675](https://doi.org/10.3389/fninf.2021.689675) • [hal-02308126](https://hal.archives-ouvertes.fr/hal-02308126)
 - Thibeau-Sutre*, E., Díaz*, M., Hassanaly, R., Routier, A., Didier, D., Colliot, O., **Burgos, N.**, ‘ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing’, *Computer Methods and Programs in Biomedicine*, 220: 106818, 2022 (*: joint first authorship). [doi:10.1016/j.cmpb.2022.106818](https://doi.org/10.1016/j.cmpb.2022.106818) • [hal-03351976](https://hal.archives-ouvertes.fr/hal-03351976)
-

3.1 Machine learning and deep learning for the diagnosis of Alzheimer’s disease

Alzheimer’s disease (AD) affects over 20 million people worldwide. Identification of AD at an early stage is important for adequate care of patients and for testing new treatments. Neuroimaging provides useful information to identify AD (Ewers et al., 2011b): atrophy due to grey matter loss with anatomical magnetic resonance imaging (MRI), hypometabolism with ^{18}F -fluorodeoxyglucose positron emission tomography (FDG PET), accumulation of amyloid-beta and tau proteins with amyloid and tau PET imaging. A major interest is then to analyse those markers to identify AD at an early stage. Machine learning and deep learning methods have the potential to assist in identifying patients with AD by learning discriminative patterns from neuroimaging data.

A large number of machine learning and deep learning approaches have been proposed to classify and predict AD stages (see Haller et al., 2011; Falahati et al., 2014; Rathore et al., 2017; Jo et al., 2019; Ebrahimighahnavieh et al., 2020; Frizzell et al., 2022; Fathi et al., 2022 for reviews). Validation and comparison of such approaches require a large number of patients followed over time. Many published work uses the publicly available Alzheimer’s Disease Neuroimaging Initiative (ADNI) data set. However, the objective comparison between their results is almost impossible because they differ in terms of: i) subsets of patients (with unclear specification of selection criteria); ii) image preprocessing pipelines (and thus it is not clear if the superior performance comes from the classification or the preprocessing); iii) feature extraction and selection; iv) machine learning algorithms; v) cross-validation procedures and vi) reported evaluation metrics. Because of these differences, it is arduous to conclude which methods perform the best. As a result, the practical impact of these works has remained very limited. Moreover, the vast majority of these works use the ADNI data set (ADNI 1 for earlier papers and most often a combination of ADNI 1, ADNI GO, ADNI 2 and ADNI 3 for more recent works). Therefore, assessment of generalisation to another data set is rarely done, even though other publicly available data sets exist such as the Australian Imaging Biomarker and Lifestyle study (AIBL) and the Open Access Series of Imaging Studies (OASIS).

Comparison papers (Cuingnet et al., 2011; Sabuncu et al., 2014) and challenges (Allen et al., 2016; Bron et al., 2015; Marinescu et al., 2018) have been an important step towards objective evaluation of machine learning methods by allowing the benchmark of different methods on the same data set and with the same preprocessing. Nevertheless, such studies provide a ‘static’ assessment of methods. Evaluation data sets are used in their current state at the time of the study, whereas new patients are continuously included in studies such as ADNI. Similarly, they are limited to the classification and preprocessing methods that were used at the time of the study. It is thus difficult to complement them with new approaches.

We proposed two frameworks for the reproducible evaluation of machine learning (AD-ML) and deep learning (AD-DL) algorithms in AD and demonstrated their use on classification of structural MRI data obtained from three publicly available data sets: ADNI, AIBL and OASIS. Note that FDG PET data were also used when evaluating machine learning algorithms, see Samper-González et al., 2018. Specifically, our contributions were three-fold:

i) a framework for the management of publicly available data sets and their continuous update with new subjects, and in particular tools for fully automatic conversion into the Brain Imaging Data Structure¹ (BIDS) format (Gorgolewski et al., 2016); ii) a modular set of preprocessing pipelines, feature extraction and classification methods, together with an evaluation framework, that provide a baseline for benchmarking different components; iii) a large-scale evaluation on T1w MRI from three publicly available neuroimaging data sets (ADNI, AIBL and OASIS).

All the code of the frameworks and the experiments is publicly available: general-purpose tools have been integrated into Clinica² (Routier et al., 2021) and ClinicaDL³ (Thibeau-Sutre et al., 2022b), which are open-source software platforms that we developed to process data from neuroimaging studies and perform deep learning analyses, and the experiments are available in specific repositories (AD-ML: <https://github.com/aramis-lab/AD-ML>, AD-DL: <https://github.com/aramis-lab/AD-DL>).

3.2 Materials

3.2.1 Data sets

Three publicly available data sets have mainly been used for the study of AD: the Alzheimer’s Disease Neuroimaging Initiative (ADNI), the Australian Imaging, Biomarkers and Lifestyle (AIBL) and the Open Access Series of Imaging Studies (OASIS) (see also Appendix B). These are the ones we also used in the AD-ML and AD-DL studies.

In the following, we describe the diagnostic labels extracted from each data set for the AD-DL study. Note that similar criteria were used for AD-ML, but AD-DL includes more participants as we included participants that had T1-weighted (T1w) MR images available (and not both T1w MRI and FDG PET as in AD-ML) and because new participants regularly join the ADNI study. For the detailed MRI protocols, one can see (Samper-González et al., 2018).

The ADNI data set used in our experiments comprises 1455 participants for whom a T1w image was available at at least one visit. Five diagnostic groups were considered:

- CN: sessions of subjects who were diagnosed as CN at baseline and stayed stable during the follow-up;
- AD: sessions of subjects who were diagnosed as AD at baseline and stayed stable during the follow-up;
- MCI: sessions of subjects who were diagnosed as MCI, early MCI or late MCI at baseline, who did not encounter multiple reversions and conversions and who did not convert back to CN;
- pMCI: sessions of subjects who were diagnosed as MCI, early MCI or late MCI at baseline, and progressed to AD during the 36 months following the current visit;

¹BIDS: <http://bids.neuroimaging.io>

²Clinica: <http://www.clinica.run>

³ClinicaDL: <https://clinicadl.readthedocs.io>

- sMCI: sessions of subjects who were diagnosed as MCI, early MCI or late MCI at baseline, and neither progress nor regress to AD during the 36 months following the current visit.

AD and CN participants whose label changed over time were excluded. This was also the case for MCI patients with two or more label changes (for instance progressing to AD and then reverting back to MCI). We made this choice because one can assume that the diagnosis of these subjects is less reliable. Naturally, all the sessions of the pMCI and sMCI groups are included in the MCI group. Note that the reverse is false, as some MCI subjects did not convert to AD but were not followed long enough to state whether they were sMCI. For 30 sessions, the preprocessing did not pass the quality check and these images were removed from our data set. Two pMCI subjects were entirely removed because the preprocessing failed for all their sessions.

The AIBL data set considered in this work is composed of 598 participants for whom a T1w MR image and an age value was available at at least one visit. The criteria used to create the diagnosis groups are identical to the ones used for ADNI. After the preprocessing pipeline, seven sessions were removed without changing the number of subjects.

Our OASIS data set is composed of 193 participants aged 62 years or more (minimum age of the participants diagnosed with AD). As this data set is not longitudinal, we consider as AD (resp. CN) participants who were diagnosed as AD (resp. CN) at baseline. After the preprocessing pipeline, 22 AD and 17 CN subjects were excluded.

Table 3.1 summarises the demographics and cognitive scores of the ADNI, AIBL and OASIS participants. Note that for the ADNI and AIBL data sets, three diagnostic labels (CN, MCI and AD) exist and were assigned by a physician after a series of clinical tests (Ellis et al., 2009; Ellis et al., 2010; Petersen et al., 2010) while for OASIS only two diagnostic labels exist, CN and AD (the MCI subjects are labelled as AD), and were assigned based on the CDR only (Marcus et al., 2007). As the diagnostic criteria of these studies differ, there is no strict equivalence between the labels of ADNI and AIBL, and those of OASIS.

3.2.2 Conversion to the Brain Imaging Data Structure

Even though public data sets are extremely valuable, an important difficulty with these studies lies in the organisation of the clinical and imaging data. As an example, the ADNI and AIBL imaging data, in the state they are downloaded, do not rely on community standards for data organisation and lack of a clear structure. Multiple image acquisitions exist for a given visit of a participant and the complementary image information is contained in numerous csv files, making the exploration of the database and subject selection very complicated. To organise the data, we selected the BIDS format (Gorgolewski et al., 2016), a community standard enabling the storage of multiple neuroimaging modalities. Being based on a file hierarchy rather than on a database management system, BIDS can be easily deployed in any environment. Very importantly, we provide the code that automatically performs the conversion of the data as they were downloaded to the BIDS organised version, for all the data sets used: ADNI, AIBL and OASIS. This allows direct reproducibility by other groups without having to redistribute the data set, which is not allowed in the case of

TABLE 3.1: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores at baseline for ADNI, AIBL and OASIS. Values are presented as mean (standard deviation) [range]. M: male, F: female

ADNI

	Subjects	Sessions	Age	Gender	MMSE	CDR
CN	330	1 830	74.4 (5.8) [59.8, 89.6]	160 M / 170 F	29.1 (1.1) [24, 30]	0: 330
MCI	787	3 458	73.3 (7.5) [54.4, 91.4]	464 M / 323 F	27.5 (1.8) [23, 30]	0: 2; 0.5: 785
sMCI	298	1 046	72.3 (7.4) [55.0, 88.4]	175 M / 123 F	28.0 (1.7) [23, 30]	0.5: 298
pMCI	295	865	73.8 (6.9) [55.1, 88.3]	176 M / 119 F	26.9 (1.7) [23, 30]	0.5: 293; 1: 2
AD	336	1 106	75.0 (7.8) [55.1, 90.9]	185 M / 151 F	23.2 (2.1) [18, 27]	0.5: 160; 1: 175; 2: 1

AIBL

	Subjects	Age	Gender	MMSE	CDR
CN	429	72.5 (6.2) [60, 92]	183 M / 246 F	28.8 (1.2) [25, 30]	0: 406; 0.5: 22; 1: 1
MCI	93	75.4 (6.9) [60, 96]	50 M / 43 F	27.0 (2.1) [20, 30]	0: 6; 0.5: 86; 1: 1
sMCI	13	76.7 (6.5) [64, 87]	8 M / 5 F	28.2 (1.5) [26, 30]	0.5: 13
pMCI	20	78.1 (6.6) [63, 91]	10 M / 10 F	26.7 (2.1) [22, 30]	0.5: 20
AD	76	73.9 (8.0) [55, 93]	33 M / 43 F	20.6 (5.5) [6, 29]	0.5: 31; 1: 36; 2: 7; 3: 2

OASIS

	Subjects	Age	Gender	MMSE	CDR
CN	76	76.5 (8.4) [62, 94]	14 M / 62 F	29.0 (1.2) [25, 30]	0: 76
AD	78	75.6 (7.0) [62, 96]	35 M / 43 F	24.4 (4.3) [14, 30]	0.5: 56; 1: 20; 2: 2

ADNI and AIBL. We also provide tools for subject selection according to desired imaging modalities, duration of follow up and diagnoses, which makes possible the use of the same groups with the largest possible number of subjects across studies. Finally, we propose a BIDS-inspired standardised structure for all the outputs of the experiments. These tools are available in Clinica (more in Section 3.5.1).

3.3 AD-ML: Framework for the reproducible evaluation of machine learning classification experiments

We demonstrate the use of the AD-ML framework for automatic classification of T1w MRI data obtained from three data sets (ADNI, AIBL and OASIS). We assess the influence of various components on the classification performance: image preprocessing, feature type (voxel or regional features), classification algorithm and data set. Experiments were first performed on the ADNI, AIBL and OASIS data sets independently, and the generalisation of the results was assessed by applying classifiers trained on ADNI to the AIBL and OASIS data. The complete set of experiments performed is available in (Samper-González et al., 2018).

3.3.1 Methods

3.3.1.1 Preprocessing and feature extraction

The preprocessing pipeline of the T1w MRI data is based on SPM12⁴. First, the Unified Segmentation procedure (Ashburner et al., 2005) is used to simultaneously perform tissue segmentation, bias correction and spatial normalisation of the input image. Next, a group template is created using DARTEL, an algorithm for diffeomorphic image registration (Ashburner, 2007), from the subjects' tissue probability maps on the native space, usually grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) tissues, obtained at the previous step. Here, not only the group template is obtained, but also the deformation fields from each subject's native space into the DARTEL template space. Lastly, the DARTEL to MNI method (Ashburner, 2007) is applied, providing a registration of the native space images into the MNI space: for a given subject its flow field into the DARTEL template is combined with the transformation of the DARTEL template into MNI space, and the resulting transformation is applied to the subject's different tissue maps. As a result, all the images are in a common space, providing a voxel-wise correspondence across subjects.

Two types of features were extracted from the imaging data: voxel and region features. The first type of features simply corresponds, for each image, to all the voxels in the brain. The signal obtained from the T1w MR images is the grey matter density. Regional features correspond to the average signal (grey matter density) computed in a set of regions of interest (ROIs) obtained from different atlases. Five atlases containing both cortical and subcortical regions, and covering the brain areas affected by AD, were selected: AAL2 (Tzourio-Mazoyer et al., 2002), AICHA (Joliot et al., 2015), Hammers (Gousias et al., 2008; Hammers et al., 2003), LPBA40 (Shattuck et al., 2008) and Neuromorphometrics⁵.

3.3.1.2 Classification models

We considered three different classifiers: linear support vector machine (SVM), logistic regression with L2 regularisation, and random forest, all available in Clinica. The linear SVM was used with both the voxel and the regional features because its computational complexity depends only on the number of subjects when using its dual form. On the other hand, the logistic regression with L2 regularisation and random forest models were only used for the region-based analyses given that their complexity depends on the number of features, which becomes infeasible with images containing about 1 million voxels. We used the implementations of the scikit-learn library (Pedregosa et al., 2011).

Linear SVM The first method included is linear SVM. To reduce computational load, the Gram matrix $K = (k(x_i, x_j))_{i,j}$ was precalculated using a linear kernel k for each pair of images (x_i, x_j) (using the region or voxel features) for the provided subjects. This Gram matrix is used as input for the generic SVM. We chose to optimise the penalty parameter C of the error term. An advantage of SVM is that, when using a precomputed Gram matrix (dual SVM), computing time depends on the number of subjects, and not on the number of

⁴SPM 12: <http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>

⁵Neuromorphometrics atlas: <http://www.neuromorphometrics.com>

features. Given its simplicity, linear SVM is useful as a baseline to compare the performance of the different methods.

Logistic regression with L2 regularisation The second method is logistic regression with L2 regularisation (which is classically used to reduce overfitting). We optimised, as for the linear SVM, the penalty parameter C of the error term. Logistic regression with L2 regularisation directly optimises the weights for each feature, and the number of features influences the training time. This is the reason why we only used it for regional features.

Random forest The third classifier used is the random forest. Unlike both linear SVM and logistic regression, random forest is an ensemble method that fits a number of decision trees on various sub-samples of the data set. The combined estimator prevents overfitting and improves the predictive accuracy. Based on the implementation provided by the scikit-learn library (Pedregosa et al., 2011), there is a large number of parameters that can be optimised. After preliminary experiments to assess which had a larger influence, we selected the following two hyperparameters to optimise: i) the number of trees in the forest; ii) the number of features to consider when looking for the best split. Random forest was only used for regional features and not voxel features, due to its high computational cost.

3.3.1.3 Evaluation strategy

Cross-validation Evaluation of classification performance mainly followed the recent guidelines provided by Varoquaux et al., 2017. Cross-validation (CV), the classical strategy to maintain the independence of the train set (used for fitting the model) and the test set (used to evaluate the performance), was performed. The CV procedure included two nested loops: an outer loop evaluating the classification performance and an inner loop used to optimise the hyperparameters of the model (C for SVM and L2 logistic regression, the number of trees and features for a split for the random forest). It should be noted that the use of an inner loop of CV is important to avoid biasing performance upward when optimising hyperparameters. This step has not always been appropriately performed in the literature (Querbes et al., 2009; Wolz et al., 2011) leading to over-optimistic results, as presented in (Eskildsen et al., 2013; Maggipinto et al., 2017).

We implemented in Clinica three different outer CV methods: k-fold, repeated k-fold and repeated random splits (all of them stratified), using scikit-learn based tools (Pedregosa et al., 2011). The choice of the method would depend on the computational resources at hand. However, whenever possible, it is recommended to use repeated random splits with a large number of repetitions to yield more stable estimates of performance and better estimates of empirical variance. Therefore, we used for each experiment 250 iterations of random splits. We report the full distribution of the evaluation metrics in addition to the mean and empirical standard-deviation, as done in (Raamana et al., 2017) that uses neuropredict (Raamana, 2017). It should nevertheless be noted that there is no unbiased estimate of variance for cross-validation (Bengio et al., 2004; Nadeau et al., 2003) and that the empirical variance largely underestimates the true variance. This should be kept in mind when interpreting the empirical variance values. Also, we chose not to perform statistical

testing of the performance of different classifiers. This is a complex matter for which there is no universal solution. In many publications, a standard t-test on cross-validation results is used. However, such an approach is way too liberal and should not be applied, as shown by Nadeau et al., 2003. Better behaved approaches have been proposed such as the conservative Z or the corrected resampled t-test (Nadeau et al., 2003). However, such approaches must be used with caution because their behaviour depends on the data and the cross-validation set-up. We thus chose to avoid the use of statistical tests in the present paper, in order not to mislead the reader. Instead, we reported the full distributions of the metrics.

For hyperparameter optimisation, we implemented an inner k-fold. For each split, the model with the highest balanced accuracy is selected, and then these selected models are averaged across splits to profit of model averaging, that should have a stabilizing effect. In the present paper, experiments were performed with $k = 10$ for the inner loop.

Metrics As output of the classification, we report the balanced accuracy, area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity and, in addition, the predicted class for each subject, so the user can calculate other desired metrics with this information.

3.3.1.4 Classification experiments

In general, we performed clinical diagnosis classification tasks, or ‘predictive’ tasks of the evolution of MCI subjects. Note that tasks involving progression from MCI to AD were not performed for AIBL due to the small number of participants in the sMCI and pMCI categories. However, the framework would allow performing these experiments very easily when more progressive MCI subjects become publicly available in AIBL. Depending on the type of features, the performance of several classifiers with different parameters was tested. For voxel features, the only classifier was the linear SVM. Four different levels of smoothing were applied to the images using a Gaussian kernel, from no smoothing to up to 12 mm full width at half maximum (FWHM). For region-based classification experiments, three classifiers were tested: linear SVM, logistic regression and random forest. The features were extracted using five atlases: AAL2, AICHA, Hammers, LPBA40 and Neuromorphometrics.

3.3.2 Results

We present a selection of the results that we believe are the most valuable. The complete results of all experiments (including other tasks, preprocessing parameters, features or classifiers) are available in the repository containing all the code and experiments (<https://github.com/aramis-lab/AD-ML>). In the following subsections, we present the results using the balanced accuracy as performance metric but all the other metrics are available on GitHub.

3.3.2.1 Influence of the atlas

To assess the impact of the choice of atlas on the classification accuracy and to potentially identify a preferred atlas, the linear SVM classifier using regional features was selected.

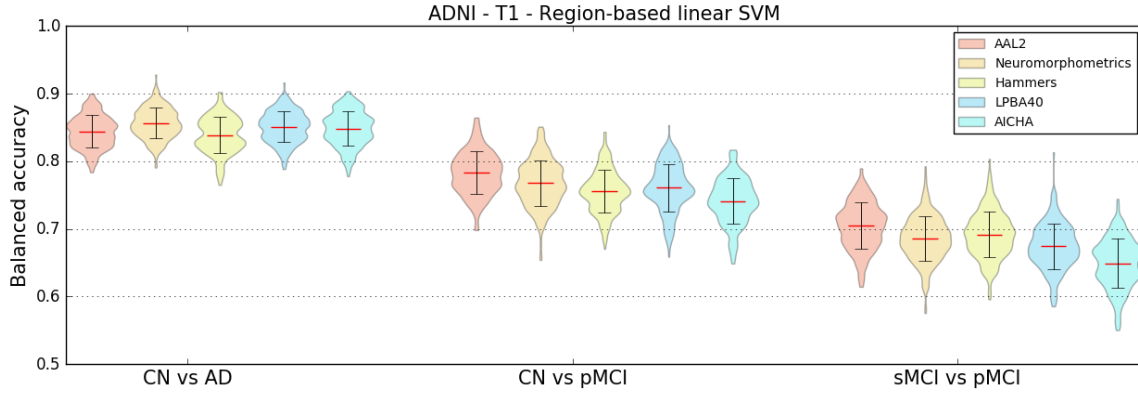


FIGURE 3.1: Influence of atlas. Distribution of the balanced accuracies obtained from the T1w MR images of ADNI participants using the reference classifier (linear SVM) and regional features from different atlases for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

Features from T1w MR images of ADNI participants were extracted using five different atlases: AAL2, AICHA, Hammers, LPBA40 and Neuromorphometrics. Three classification tasks were studied: CN vs AD, CN vs pMCI and sMCI vs pMCI.

As shown in Figure 3.1, no specific atlas provides the highest classification accuracy for all the tasks. For example, Neuromorphometrics and AICHA provide better results for CN vs AD, along with LPBA40, while AAL2 provides the highest balanced accuracy for CN vs pMCI and sMCI vs pMCI. The same analysis was performed on AIBL subjects and, similarly, no atlas consistently performed better than others across tasks. For the following region-based experiments, the AAL2 atlas was chosen as reference atlas as it leads to good classification accuracies and is widely used in the neuroimaging community. Again, all other results are available in the repository.

3.3.2.2 Influence of the smoothing

T1w MR images were either not smoothed or smoothed using Gaussian kernels with FWHMs of 4 mm, 8 mm and 12 mm. To determine the influence of different smoothing degrees on the classification accuracy, a linear SVM classifier using voxel features was chosen. Three classification tasks were studied: CN vs AD, CN vs pMCI and sMCI vs pMCI. The results in Figure 3.2 show that, for most classification tasks, the balanced accuracy does not vary to a great extent with the smoothing kernel size. The only variations are observed for the CN vs pMCI and sMCI vs pMCI tasks: the balanced accuracy increases slightly with the kernel size. The same analysis was run using the AIBL dataset. The mean balanced accuracy also increased slightly with the kernel size, but the standard deviations of the balanced accuracies are larger than for ADNI. As the degree of smoothing does not have a clear impact on the classification performance, we chose to present the subsequent results related to the voxel-based classification with a reference smoothing of 4 mm.

3.3.2.3 Influence of the type of features

We compared the balanced accuracies obtained for the voxel features with reference smoothing (Gaussian kernel of 4 mm FWHM) to the ones obtained for the regional features with

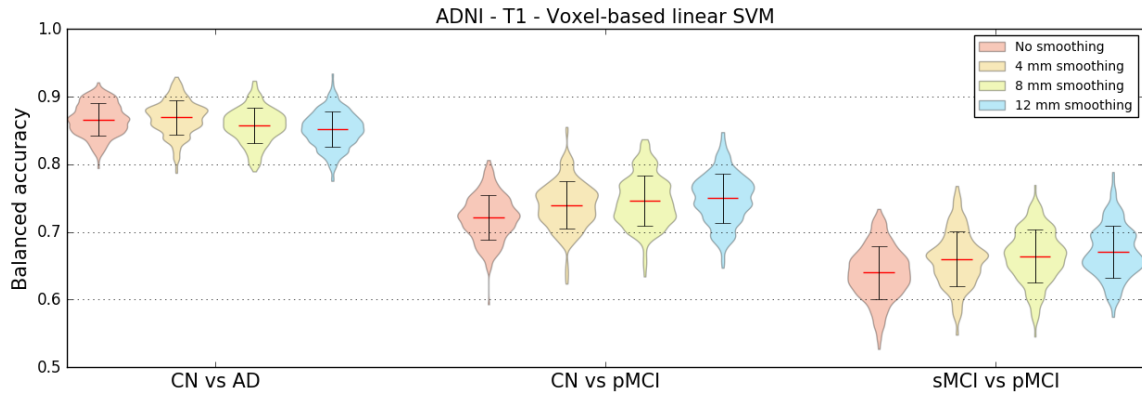


FIGURE 3.2: Influence of smoothing. Distribution of the balanced accuracy obtained from the T1w MR images of ADNI participants using the reference classifier (linear SVM) and voxel features with different degrees of smoothing for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

reference atlas (AAL2) when using linear SVM classifiers. These features were extracted from T1w MRI of ADNI participants. The same three classification tasks as before were evaluated.

The results, displayed in Table 3.2, do not show notable differences between the mean balanced accuracies obtained using voxel or regional features. In the case of the AIBL data set, the balanced accuracy is higher for the region-based classification (for AD vs CN: voxel-based $0.79 [\pm 0.059]$, region-based $0.86 [\pm 0.042]$), but we can observe that the corresponding standard deviations are high.

TABLE 3.2: Influence of feature types. Mean balanced accuracy and standard deviation obtained for three tasks (CN vs AD, CN vs pMCI and sMCI vs pMCI) using the reference classifier (linear SVM) with voxel (reference smoothing: 4 mm) and region (reference atlas: AAL2) features extracted from T1w MR images of ADNI subjects.

	Voxel-based (4 mm smoothing)	Region-based (AAL2 atlas)
CN vs AD	0.87 ± 0.026	0.84 ± 0.024
CN vs pMCI	0.74 ± 0.035	0.78 ± 0.031
sMCI vs pMCI	0.66 ± 0.040	0.70 ± 0.034

3.3.2.4 Influence of the classification method

Region-based experiments were carried out using three different classifiers to evaluate if there were variations in balanced accuracies depending on the chosen classifier. Regional features were extracted using the reference AAL2 atlas from T1w MR images of ADNI participants. The three previously defined classification tasks were performed.

The results displayed in Figure 3.3 show that both the linear SVM and logistic regression with L2 regularisation models lead to similar balanced accuracies, consistently higher than the one obtained with random forest for all the tasks tested.

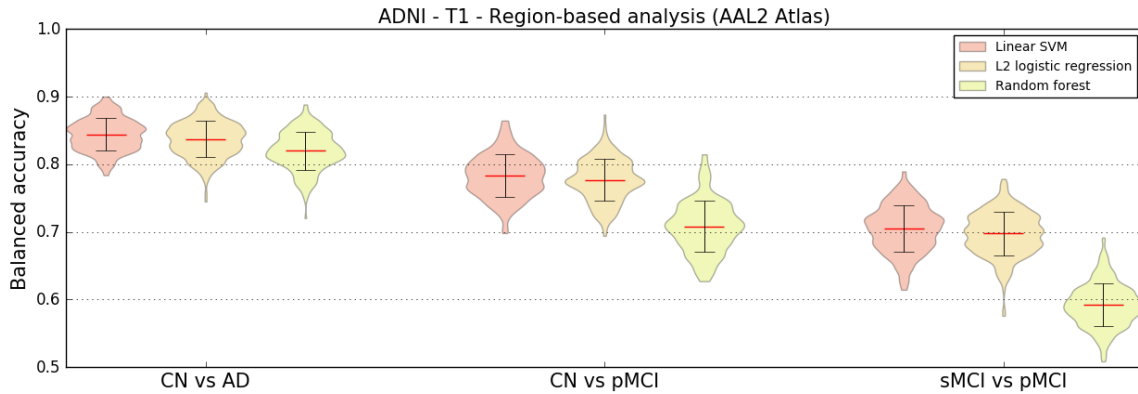


FIGURE 3.3: Influence of classification method. Distribution of the balanced accuracy obtained from the T1w MR images of ADNI participants using different region-based classifiers (reference atlas: AAL2) for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

3.3.2.5 Influence of the class imbalance

The tasks that we performed are done with unbalanced classes. Such class imbalance ranges from very mild (1.2 times more CN than AD for ADNI) to moderate (1.7 times more CN than pMCI and 2 times more sMCI than pMCI for ADNI) to very strong (6.1 times more CN than AD in AIBL). We aimed to assess if such class imbalance influenced the performance. To that purpose, we randomly sampled subgroups and performed experiments with 237 CN vs 237 AD, 167 pMCI vs 167 CN and 167 pMCI vs 167 pMCI for ADNI and 72 CN and 72 AD for AIBL. We ensured that the demographic and clinical characteristics of the balanced subsets did not differ from the original ones. Results are presented in Figure 3.4. For ADNI, the performance was similar to that obtained with the full population. For AIBL, the performance was substantially higher with balanced groups for the voxel-based features. It thus seems that a very strong class imbalance (as in the case of AIBL where the proportion is 6 to 1) leads to lower performance but that moderate class imbalance (up to 2 to 1 in ADNI) are adequately handled.

3.3.2.6 Influence of the data set

We also wanted to know how consistent were the results across data sets, and thus we compared the classification performance obtained from ADNI, AIBL and OASIS, for the task of differentiating control subjects from patients with Alzheimer’s disease. Voxel (4 mm smoothing) and regional (AAL2 atlas) features were extracted from T1w MR images and used with linear SVM classifiers. We tested two configurations: training and testing the classifiers on the same data set, and training a classifier on ADNI and testing it on AIBL and OASIS. Results are displayed in Table 3.3. The performance obtained on ADNI and AIBL is comparable and much higher than those obtained on OASIS. When training on ADNI and testing on AIBL or OASIS, the balanced accuracy was at least as high as when training and testing on AIBL or OASIS respectively, suggesting that classifiers trained on ADNI generalise well to the other data sets. In particular, training on ADNI substantially improved the classification performance on OASIS. We aimed to assess whether this was due to the larger number of subjects in ADNI. To that purpose, we performed the same

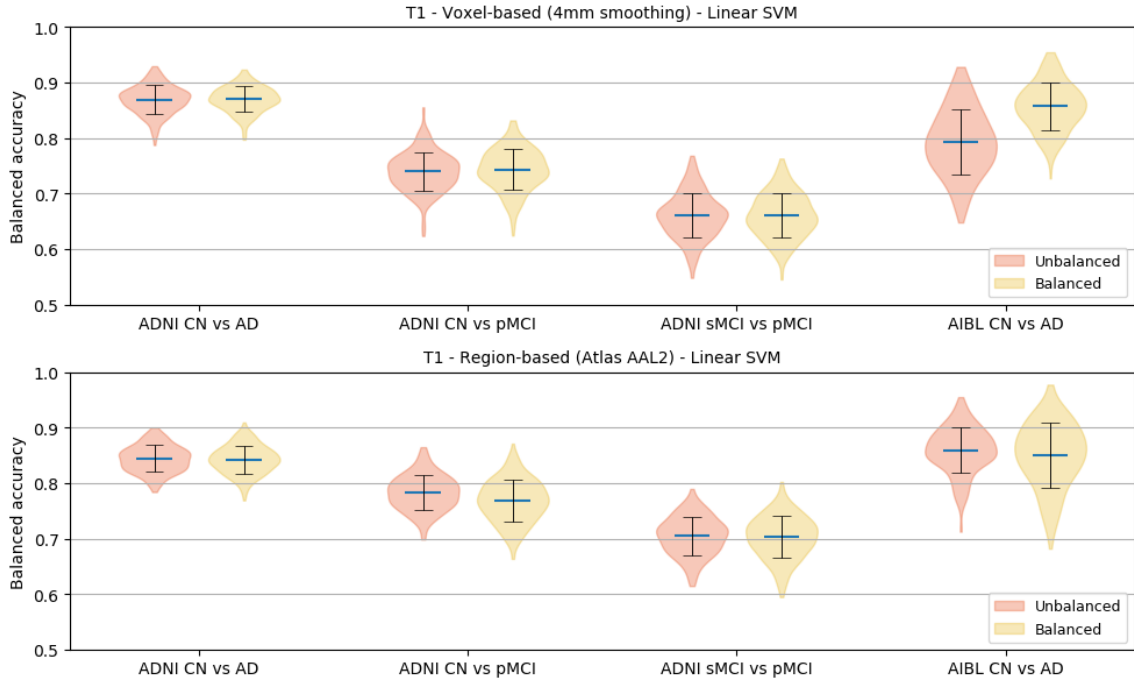


FIGURE 3.4: Influence of class imbalance. Distribution of the balanced accuracies obtained using voxel (reference smoothing: 4 mm) and regional (reference atlas: AAL2) features extracted from T1w MR images using the reference classifier (linear SVM) when training using unbalanced and balanced data sets. Four tasks were tested: CN vs AD, CN vs pMCI and sMCI vs pMCI for ADNI subjects, and CN vs AD for AIBL subjects.

experiments but with subsets of participants of equal size for each data set. We randomly sampled populations of 70 AD patients and 70 CN participants from each of the data sets, ensuring that the demographic and clinical characteristics of the subpopulations did not differ from the original ones. As can be seen from Table 3.3, using the subset, the improvement disappears for the voxel-based but remains for the regional features.

3.3.2.7 Influence of the training data set size

Learning curves were computed to assess how the performance of linear SVM classifiers varies depending on the size of the training data set. Using only ADNI participants, we tested two scenarios: voxel and region features extracted from T1w MR images. As cross-validation, 250 iterations were run where the data set was randomly split into a test data set (30% of the samples) and a training data set (70% of the samples). The maximum number of subjects used for training and testing for each of the different tasks is of 362 for CN vs AD, of 313 for CN vs pMCI and of 355 for sMCI vs pMCI. For each run, 10 classifiers were trained and evaluated on the same test set using from 10% to all of the training set (from 7% to up to 70% of the samples), increasing the number of samples used by 10% on each step. Therefore, the number of participants used for training ranged from 20 to 197 for CN, 24 to 239 for sMCI, 12 to 117 for pMCI and 17 to 166 for AD. We can observe from the learning curves in Figure 3.5 that, as expected, the balanced accuracy increases with the number of training samples.

Learning curves were also computed for the CN vs AD task when using larger data sets obtained by combining participants from ADNI and AIBL (balanced subset composed of 72

TABLE 3.3: Influence of data set. Average \pm SD of the balanced accuracy obtained for the reference linear SVM classifier when differentiating CN and AD subjects using voxel (4 mm smoothing) and regional (AAL2 atlas) features extracted from T1w MR images for three data sets: ADNI, AIBL and OASIS. Upper rows display results for the full population. Lower rows display results for subsets of equal size for each data set. The subsets were obtained by randomly sampling populations of 70 AD patients and 70 CN participants from each of the data sets. Note that for the ‘full data set’ experiment, a balanced subset of AIBL was used (i.e., 72 CN and 72 AD subjects). When the testing data set differs from the training data set, there is no CV and thus no empirical SD.

	Training data set	Testing data set	Voxel-based (4 mm smoothing)	Region-based (AAL2 atlas)
Full data set	ADNI	ADNI	0.87 ± 0.025	0.84 ± 0.024
	AIBL	AIBL	0.85 ± 0.003	0.86 ± 0.004
	ADNI	AIBL	0.87	0.88
	OASIS	OASIS	0.70 ± 0.058	0.71 ± 0.053
	ADNI	OASIS	0.76	0.76
Subset	ADNI	ADNI	0.85 ± 0.048	0.81 ± 0.06
	AIBL	AIBL	0.86 ± 0.048	0.85 ± 0.058
	ADNI	AIBL	0.86	0.87
	OASIS	OASIS	0.67 ± 0.063	0.64 ± 0.072
	ADNI	OASIS	0.67	0.7

CN subjects and 72 AD subjects) and from ADNI, AIBL and OASIS. Results are displayed in Figure 3.6. We observe that for an equivalent number of subjects, combining ADNI and AIBL or only using ADNI leads to a similar balanced accuracy. For regional features, the performance is slightly higher when combining ADNI and AIBL compared to when only using ADNI, but the difference is largely within the standard deviation. The balanced accuracy keeps increasing slightly as more subjects are used for training when combining ADNI and AIBL. However, when combining ADNI, AIBL and OASIS, the performance is worse than when only using ADNI or combining ADNI and AIBL, no matter the number of subjects. This is probably due to the fact that ADNI and AIBL follow the same diagnosis and acquisition protocols, which differ from those of OASIS.

3.3.3 Discussion

AD-ML is an open-source framework for the reproducible evaluation of AD classification methods that contains the following components: i) converters to normalize three publicly available data sets into BIDS; ii) standardised preprocessing and feature extraction pipelines for T1w MRI; iii) standard machine learning classification algorithms; iv) cross-validation procedures following recent best practices. We demonstrated its use for the assessment of different preprocessing options, features and classifiers on three public data sets.

We demonstrated the use of the framework on different classification tasks based on T1w MRI data. Through this, we aimed to provide a baseline performance to which advanced machine learning and feature extraction methods can be compared. The baseline performance is in line with the state-of-the-art results, which have been summarised in (Arbabshirani et al., 2017; Falahati et al., 2014; Rathore et al., 2017), where classification

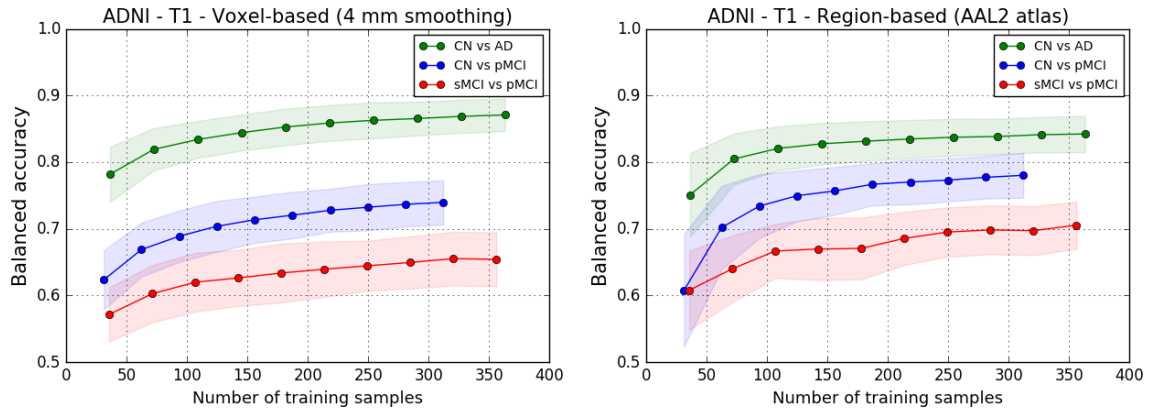


FIGURE 3.5: Influence of training data set size. Learning curves for the T1w MR images of ADNI participants using voxel features with 4 mm of smoothing (left) and regional features derived from the AAL2 atlas (right) for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

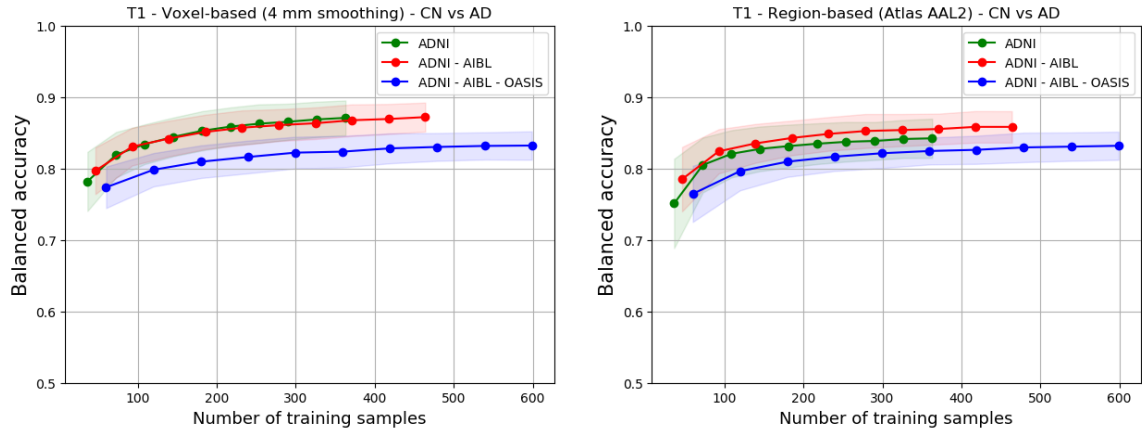


FIGURE 3.6: Influence of training set size when combining data sets. Learning curves for the voxel features with 4 mm of smoothing (left) and regional features derived from the AAL2 atlas (right) extracted from T1w MR images for the CN vs AD task when using subjects from ADNI only, from both ADNI and AIBL, and from ADNI, AIBL and OASIS. Note that a balanced subset of AIBL was used (i.e., 72 CN and 72 AD subjects).

accuracies typically range from 80% to 95% for CN vs AD, and from 60% to 80% for sMCI vs pMCI.

Diverse parameters and options are used as for preprocessing and feature extraction in AD machine learning studies. Their influence on classification performance is not clear and constitutes a problem for the comparability of classification methods. We assessed the effect of the choice of atlas, of degree of smoothing, and of the type of features (regions or voxels). We found no systematic effect of each of these different components on the performance. Some studies found an influence of the atlas on the classification performance (Ota et al., 2015; Ota et al., 2014). However, the number of subjects in this study was small. In (Chu et al., 2012), an improvement of 3% was found when using a combination of a few ROIs compared to using all the voxels. In our study, a much larger number of subjects and a strict validation process were used.

We compared three widely used classification methods: SVM, logistic regression with L2 regularisation and random forests. Our main finding was the underperformance of the

latter. This might be caused by the nature of brain imaging data that contains relatively homogeneous values, and which should show dependence across voxels or brain regions. These characteristics of the data could explain why techniques trying to find a smooth combination of features, such as those using L2 regularisation, are more suited for single modality classification problem. On the other hand, random forests or other ensemble methods could be useful when combining features from different modalities such as images, clinical data and cognitive scores, as done in (Moradi et al., 2015; Sørensen et al., 2018). In other papers comparing several standard classification algorithms such as SVM, linear discriminant analysis or Naive Bayes (Aguilar et al., 2013; Cabral et al., 2015; Sabuncu et al., 2014), results did not show differences between methods.

We also assessed the influence of class imbalance, which in our data sets ranges from very mild (1.2 times more CN than AD for ADNI) to moderate (1.7 times more CN than pMCI and 2 times more sMCI than pMCI for ADNI) to very strong (6.1 times more CN than AD in AIBL). In the case of voxel-based features, we found that a very strong class imbalance (as in the case of AIBL where the proportion is 6 to 1) leads to lower performance but that moderate class imbalance (up to 2 to 1 in ADNI) are adequately handled. On the other hand, there was no influence of class imbalance for regional features. This highlights that it may be beneficial to use balanced groups for training when there is a very strong class imbalance and when using very high dimensional features.

We assessed the influence of various components on classification performance: type of features, choice of atlas, smoothing, classifier. Other studies have assessed the influence of other components: different types of anatomical features including volume, cortical thickness and other surface characteristics (Gómez-Sancho et al., 2018; Schwarz et al., 2016; Westman et al., 2013), feature selection techniques (Tohka et al., 2016), normalisation to intracranial volume (Voevodskaya et al., 2014; Westman et al., 2013). Moreover, Tohka et al., 2016 compared LASSO and elastic-net to SVM and found that the former methods provide increased performance. Assessing the influence of these different components could also be done using our framework. In this particular work, we restricted the application of the framework to a set of components that were chosen for the following reasons. Voxel-based and regional features were both included because they are widely used. On the other hand, cortical measures based on Freesurfer were not included due to their computational cost. Smoothing is widely used for voxel-based analyses in the neuroimaging community and it seemed useful to assess its influence. Nevertheless, there is always some arbitrariness in such choices and it would be interesting to study other components with the framework.

Using multiple data sets is important to assess if the performance is robust to different populations, acquired in different conditions. A first component consisted in performing the same experiments on different data sets. We found that classification results were similar for ADNI and AIBL data sets, but much lower for OASIS. The lower performance for OASIS is likely due to the diagnosis criteria which are less rigorous (in OASIS, all participants with $CDR > 0$ are considered AD). It is also valuable to know how a classifier will perform when trained on one data set and tested on another one. The classifiers trained on ADNI data generalised well to AIBL and OASIS. Interestingly, for OASIS, the performance was substantially increased when training on ADNI compared to when training on OASIS. Such

improvement may arise from several factors: larger training set size, higher image quality or stricter diagnostic criteria. When using subsets of equal size, the improvement obtained for voxel-based features disappeared, suggesting that increased training set size is important, in particular when using very high dimensional features. On the other hand, for regional features, training on the ADNI subset improved performance compared to training on the OASIS subset, suggesting that other factors (image quality, stricter diagnostic criteria) contribute to the improvement. In general, we can say that classifiers are able to generalise across different data sets, as is also concluded in (Dukart et al., 2013; Sabuncu et al., 2014) particularly if they are obtained using large multicentric data sets with strict diagnostic criteria, as is the case for ADNI.

Unsurprisingly, increased training set size led to increased classification performance. This improvement of the results depending on the training set size has also been found in other studies such as (Abdulkadir et al., 2011; Chu et al., 2012; Franke et al., 2010). One can note that when combining multiple data sets, performance also increased with training set size. However, when combining OASIS together with ADNI and AIBL, the performance was lower than when using only AIBL and ADNI. This is consistent with the fact that performance for OASIS was systematically lower than those obtained on ADNI and AIBL. Again, this is likely due to diagnostic criteria which are less rigorous in OASIS. Interestingly, with the current number of samples available, the point where the results stop improving has not been reached. The performance of the classifier reaches a limit imposed by the number of images that have been provided for training, meaning that more data are necessary to find the top performance of a classifier. These results highlight the need for more publicly available data sets, on which most of the current research in the field relies.

3.4 AD-DL: Framework for the reproducible evaluation of deep learning classification experiments

As the most widely used architecture of deep learning, convolutional neural networks (CNN) have attracted a huge attention thanks to its great success in image classification (Krizhevsky et al., 2012). Contrary to conventional machine learning, deep learning allows the automatic abstraction of low-to-high level latent feature representations (e.g., lines, dots or edges for low level features, and objects or larger shapes for high level features). Thus, one can hypothesise that deep learning depends less on image preprocessing and requires less prior on other complex procedures, such as feature selection, resulting in a more objective and less bias-prone process (LeCun et al., 2015).

Numerous studies have proposed to assist diagnosis of AD by means of CNNs, but as for those implementing machine learning approaches, these studies are not directly comparable because they differ in terms of: i) sets of participants; ii) image preprocessing procedures, iii) cross-validation procedure and iv) reported evaluation metrics. It is thus impossible to determine which approach performs best. The generalisation ability of these approaches also remains unclear. In deep learning, the use of fully independent test sets is even more critical than in conventional machine learning, because of the very high flexibility with numerous possible model architecture and training hyperparameter choices. Assessing generalisation

to other studies is also critical to ensure that the characteristics of the considered study have not been overfitted. In previous works, the generalisation may be questionable due to inadequate validation procedures, the absence of an independent test set, or a test set chosen from the same study as the training and validation sets (Wen et al., 2020).

We thus extended our open-source framework for reproducible evaluation of AD classification to DL approaches by implementing a modular set of image preprocessing procedures, classification architectures and evaluation procedures dedicated to DL. We used this framework to rigorously assess the performance of different CNN architectures, representative of the literature. We studied the influence of key components on the classification accuracy, we compared the proposed CNNs to a conventional ML approach based on a linear SVM, and we assessed the generalisation ability of the CNN models within (training and testing on ADNI) and across datasets (training on ADNI and testing on AIBL or OASIS).

3.4.1 Methods

3.4.1.1 Preprocessing of T1w MRI

In principle, CNNs require only minimal preprocessing because of their ability to automatically extract low-to-high level features. However, in AD classification where data sets are relatively small and thus deep networks may be difficult to train, it remains unclear whether they can benefit from more extensive preprocessing. Moreover, previous studies have used varied preprocessing procedures but without systematically assessing their impact. Thus, in the current study, we compared two different procedures: ‘Minimal’ and ‘Extensive’. Both procedures included bias field correction, and (optional) intensity rescaling. In addition, the ‘Minimal’ processing included a linear registration while the ‘Extensive’ included non-linear registration and skull-stripping.

In brief, the ‘Minimal’ preprocessing procedure performs the following operations. The N4ITK method (Tustison et al., 2010) was used for bias field correction. Next, a linear (affine) registration was performed using ANTs (Avants et al., 2008) to register each image to the MNI space (ICBM 2009c nonlinear symmetric template) (Fonov et al., 2011; Fonov et al., 2009). To improve the computational efficiency, the registered images were further cropped to remove the background. The final image size is $169 \times 208 \times 179$ with 1 mm^3 isotropic voxels. Intensity rescaling, which was performed based on the min and max values, denoted as MinMax, was set to be optional to study its influence on the classification results.

In the ‘Extensive’ preprocessing procedure, bias field correction and non-linear registration were performed using the Unified Segmentation approach (Ashburner et al., 2005) available in SPM12⁶. Note that we do not use the tissue probability maps but only the non-linearly registered, bias corrected, MR images. Subsequently, we perform skull-stripping based on a brain mask drawn in MNI space. We chose this mask-based approach over direct image-based skull-stripping procedures because the latter did not prove robust on our data. This mask-based approach is less accurate but more robust. In addition, we performed intensity rescaling as in the ‘Minimal’ pipeline.

⁶<http://www.fil.ion.ucl.ac.uk/spm/software/spm12>

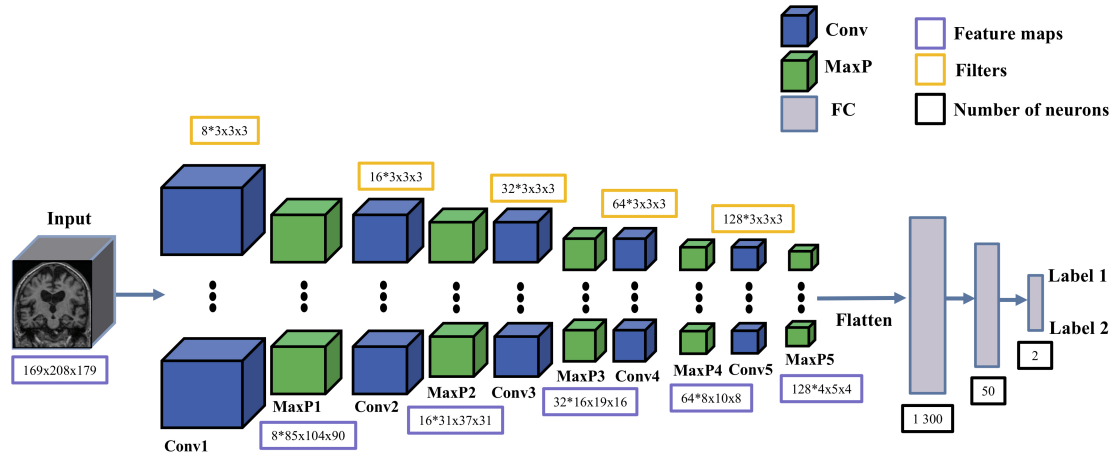


FIGURE 3.7: Architecture of the 3D subject-level CNN. For each convolutional block, we only display the convolutional and max pooling layers. Filters for each convolutional layer represent the number of filters * filter size. Feature maps of each convolutional block represent the number of feature maps * size of each feature map. Conv: convolutional layer; MaxP: max pooling layer; FC: fully connected layer.

3.4.1.2 Classification models

We considered the four classification approaches: i) 3D subject-level CNN, ii) 3D ROI-based CNN, iii) 3D patch-level CNN and iv) 2D slice-level CNN.

In the case of deep learning, one challenge is to find the ‘optimal’ model (i.e., global minimum), including the architecture hyperparameters (e.g., number of layers, dropout, batch normalisation) and the training hyperparameters (e.g., learning rate, weight decay). We first reviewed the architectures used in the literature among the studies in which no data leakage problem was found (see Table 1 in Wen et al., 2020). As there was no consensus, we used a heuristic strategy for each of the four approaches described in (Wen et al., 2020).

3D subject-level CNN For the 3D subject-level approach, the proposed CNN architecture is shown in Figure 3.7. The CNN consisted of 5 convolutional blocks and 3 FC layers. Each convolutional block was sequentially made of one convolutional layer, one batch normalisation layer, one ReLU and one max pooling layer.

3D ROI-based and 3D patch-level CNN For the 3D ROI-based and 3D patch-level approaches, the chosen CNN architecture, shown in Figure 3.8, consisted of 4 convolutional blocks (with the same structure as in the 3D subject-level) and 3 FC layers.

To extract the 3D patches, a sliding window ($50 \times 50 \times 50 \text{ mm}^3$) without overlap was used to convolve over the entire image, generating 36 patches for each image.

For the 3D ROI-based approach, we chose the hippocampus as a ROI, as done in previous studies. We used a cubic patch ($50 \times 50 \times 50 \text{ mm}^3$) enclosing the left (resp. right) hippocampus. The centre of this cubic patch was manually chosen based on the MNI template image (ICBM 2009c nonlinear symmetric template). We ensured visually that this cubic patch included all the hippocampus.

For the 3D patch-level approach, two different training strategies were considered. First, all extracted patches were fitted into a single CNN (denoting this approach as 3D patch-level single-CNN). Secondly, we used one CNN for each patch, resulting in finally 36 (number of patches) CNNs (denoting this approach as 3D patch-level multi-CNN).

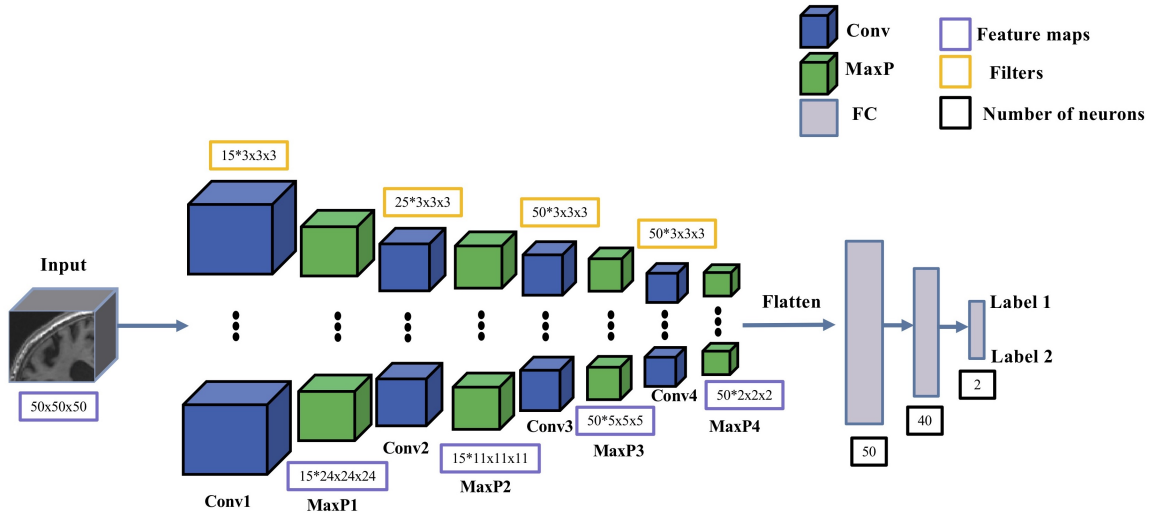


FIGURE 3.8: Architecture of the 3D ROI-based and 3D patch-level CNNs. For each convolutional block, we only display the convolutional and max pooling layers. Filters for each convolutional layer represent the number of filters * filter size. Feature maps of each convolutional block represent the number of feature maps * size of each feature map. Conv: convolutional layer; MaxP: max pooling layer; FC: fully connected layer.

2D slice-level CNN For the 2D slice-level approach, the ResNet pre-trained on ImageNet was adopted and fine-tuned. The architecture is shown in Figure 3.9. The architecture details of ResNet can be found in (He et al., 2016a). We added one FC layer on top of the ResNet. The last five convolutional layers and the last FC layer of the ResNet, as well as the added FC layer, were fine-tuned. The weight and bias of the other layers of the CNN were frozen during fine-tuning to avoid overfitting. For each subject, each sagittal slice was extracted and replicated into R, G and B channels respectively, in order to generate an RGB image. The first and last twenty slices were excluded due to the lack of information, which resulted in 129 RGB slices for each image.

Majority voting system For 3D patch-level, 3D ROI-based and 2D slice-level CNNs, we adopted a soft voting system (Raschka, 2015) to generate the subject-level decision. The subject-level decision is generated based on the decision for each slice (resp. for each patch for 3D patch-level / resp. for the left and right hippocampus for ROI-based). More precisely, it was computed based on the predicted probability p obtained after softmax normalisation of the outputs of all the slices/patches/ROIs from the same patient: $\hat{y} = \operatorname{argmax}_i w_j p_{ij}$, where w_j is the weight assigned to the j -th patch/slice/ROI. w_j reflects the importance of each slice/patch/ROI and is weighted by the normalised accuracy of the j -th slice/patch/ROI. For the evaluation on the test sets, the weights computed on the validation set were used.

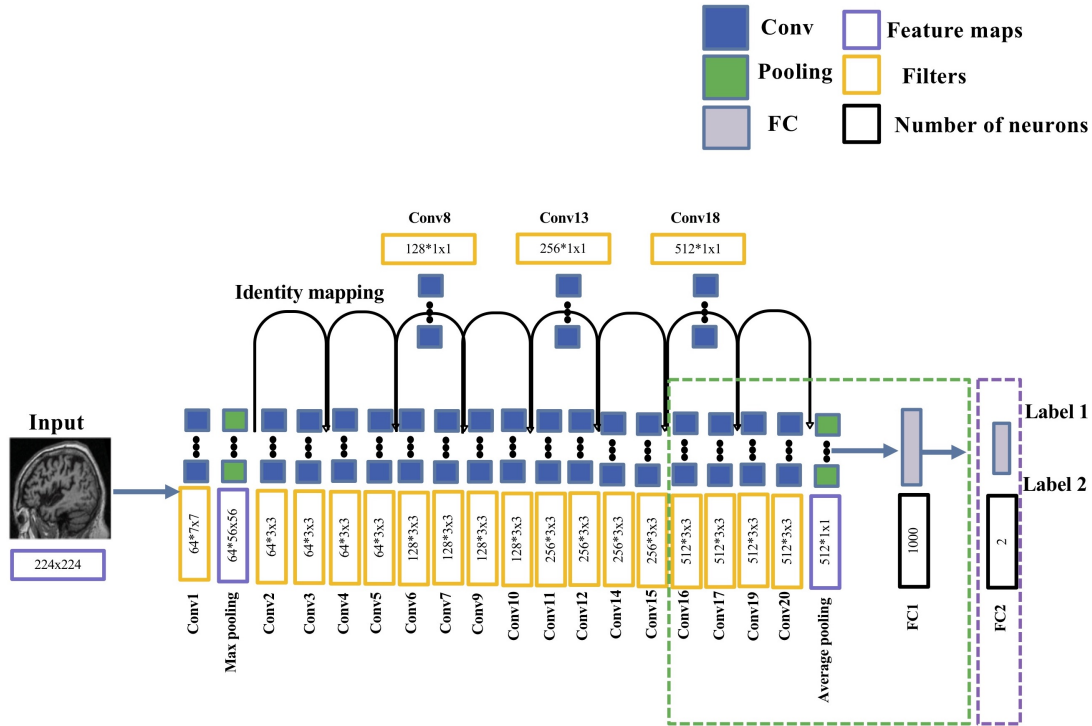


FIGURE 3.9: Architecture of the 2D slice-level CNN. An FC layer (FC2) was added on top of the ResNet. The last five convolutional layers and the last FC of ResNet (green dotted box) and the added FC layer (purple dotted box) were fine-tuned and the other layers were frozen during training. Filters for each convolutional layer represent the number of filters * filter size. Feature maps of each convolutional block represent the number of feature maps * size of each feature map. Conv: convolutional layer; FC: fully connected layer.

Note that the predicted probability p is not calibrated and should be interpreted with care as it is not reflective of the true underlying probability of the sample applied to CNNs (Kuhn et al., 2013; Guo et al., 2017).

For the 3D patch-level multi-CNN approach, the 36 CNNs were trained independently. In this case, the weaker classifiers' weight (balanced accuracy < 0.7) was set to be 0 with the consideration that the labels' probabilities of these classifiers could harm the majority voting system.

Comparison to a linear SVM on voxel-based features For comparison purposes, classification was also performed with a linear SVM classifier. We chose the linear SVM as we showed with the AD-ML framework that it obtained higher or at least comparable classification accuracy compared with other conventional models (logistic regression and random forest) (Samper-González et al., 2018). Moreover, given the very high-dimensionality of the input, a non-linear SVM, e.g., with a radial basis function kernel, may not be advantageous since it would only transport the data into an even higher dimensional space. The SVM took as input the modulated grey matter density maps non-linearly registered to the MNI space using the DARTEL method (Ashburner, 2007), as in AD-ML.

3.4.1.3 Transfer learning

Two different approaches were used for transfer learning: i) autoencoder (AE) pre-training for 3D CNNs; and ii) ResNet pre-trained on ImageNet for 2D CNNs.

Autoencoder pre-training An AE was designed based on the architecture of the classification CNN it initialises. The encoder part of the AE is composed of the same sequence of convolutional blocks as the corresponding CNN. Each block has one convolutional layer, one batch normalisation layer, one ReLU and one max pooling layer. The architecture of the decoder mirrored that of the encoder, except that the order of the convolution layer and the ReLU was swapped. Of note, the pre-training with AE and classification with CNNs in our experiments used the same training and validation data splits in order to avoid potential data leakage problems. Also, each AE was trained on all available data in the training sets. This means that all MCI, AD and CN subjects in the training data set were used to train the AE.

ImageNet pre-training For the 2D slice-level experiments, we investigated the possibility to transfer a ResNet pre-trained on ImageNet (He et al., 2016a) to our specific tasks. Next, the fine-tuning procedure was performed on some of the final layers (see Figure 3.9).

3.4.1.4 Classification tasks

We performed two tasks in our experiments. AD vs CN was used as baseline task to compare the results of our different frameworks. Then the best frameworks were selected to perform the prediction task sMCI vs pMCI: the weights and biases of the model learned on the source task (AD vs CN) were transferred to a new model fine-tuned on the target task (sMCI vs pMCI). For the SVM, the sMCI vs pMCI experiment was performed either by training directly on sMCI vs pMCI or by training on AD vs CN and applying the trained model to sMCI vs pMCI.

3.4.1.5 Evaluation strategy

Validation procedure Rigorous validation is essential to objectively assess the performance of a classification framework. This is particularly critical in the case of deep learning as one may easily overfit the validation data set when manually performing model selection and hyperparameter fine-tuning. An independent test set should be, at the very beginning, generated and concealed. It should not be touched until the cross-validation, based on the training and validation data sets, is finished and the final model is chosen. This test data set should be used only to assess the performance (i.e., generalisation) of a fully specified and trained classifier (Ripley, 1996; Sarle, 1997; Kriegeskorte et al., 2009). Considering this, we chose a classical split into training/validation/test sets. Training/validation sets were used in a cross-validation procedure for model selection while the test set was left untouched until the end of the peer-review process. Only the best performing model for each approach (3D subject-level, 3D patch-level, 3D ROI-based, 2D slice-level), as defined by the cross-validation on training/validation sets, was tested on the test set.

The ADNI test set consisted of 100 randomly chosen age- and sex-matched subjects for each diagnostic class (i.e., 100 CN subjects, 100 AD patients). The rest of the ADNI data was used as training/validation set. We ensured that age and sex distributions between training/validation and test sets were not significantly different. Two other test sets were composed of all subjects of OASIS and AIBL. The ADNI test set is used to assess model generalisation within the same data set (thereby assessing that the model has not overfitted the training/validation set). The AIBL test set is used to assess generalisation to another data set that has similar inclusion criteria and image acquisition parameters to those of the training set. The OASIS test is used to assess generalisation to a data set with different inclusion criteria and image acquisition parameters. As mentioned above, it is important to note that the diagnosis labels are not based on the same criteria in OASIS on the one hand and ADNI/AIBL on the other. Thus we do not hypothesise that the models trained on ADNI will generalise well to OASIS.

The model selection procedure, including model architecture selection and training hyperparameter fine-tuning, was performed using only the training/validation data set. For that purpose, a 5-fold cross-validation was performed, which resulted in one fold (20%) of the data for validation and the rest for training. Note that the 5-fold data split was performed only once for all the experiments with a fixed seed number (`random_state = 2`), thus guaranteeing that all the experiments used exactly the same subjects during cross-validation. Also, no overlap exists between the MCI subjects used for AE pre-training (using all available AD, CN and MCI) and the test data set of sMCI vs pMCI. Thus, the evaluation of the cross-task transfer learning (from AD vs CN to sMCI vs pMCI) is unbiased. Finally, for the linear SVM, the hyperparameter C controlling the amount of regularisation was chosen using an inner loop of 10-fold cross-validation (thereby performing a nested cross-validation).

Metrics We computed the following performance metrics: balanced accuracy (BA), AUC, accuracy, sensitivity and specificity. In the manuscript, for the sake of concision, we report only the BA but all other metrics are available on Zenodo under the DOI [10.5281/zenodo.3491003](https://doi.org/10.5281/zenodo.3491003).

3.4.2 Results of the cross-validation experiments

The different classification experiments and results (validation BA during 5-fold cross-validation) are detailed in Table 3.4.

3.4.2.1 3D subject-level

Our series of experiments started with the 3D subject-level CNN trained to perform the AD vs CN task (Table 3.4 A). We first assessed the influence of intensity rescaling. Without rescaling, the CNN did not perform better than chance ($BA = 0.50$) and there was an obvious generalisation gap (high training but low validation BA). With intensity rescaling, the BA improved to 0.80. Based on these results, intensity rescaling was used in all subsequent experiments.

With the experiments aimed at studying the influence of transfer learning (AE pre-training), we showed that the performance was slightly higher with AE pre-training (0.82)

TABLE 3.4: Results of the cross-validation experiments. For each model, we report the balanced accuracy for each of the five folds within square brackets and the average and standard-deviation across the folds. Note that this is not the standard-deviation of the estimator of balanced accuracy.

MinMax: for CNNs, intensity rescaling was performed based on min and max values, resulting in all values to be in the range of [0, 1]; SPM-based: the SPM-based grey matter maps are intrinsically rescaled; AE: autoencoder. For CNNs, sMCI vs pMCI tasks were performed as follows: the weights and biases of the model learnt on the source task (AD vs CN) were transferred to a new model fine-tuned on the target task (sMCI vs pMCI). For SVM, the sMCI vs pMCI was performed either training directly on sMCI vs pMCI or training on AD vs CN and applying the trained model to sMCI vs pMCI.

A. 3D subject-level CNN - AD vs CN

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Balanced accuracy
Baseline	Minimal	None	None	0.50 (0.00) [0.50, 0.50, 0.50, 0.50, 0.50]
		MinMax	None	0.80 (0.05) [0.76, 0.86, 0.81, 0.85, 0.74]
			AE pre-train	0.82 (0.05) [0.74, 0.90, 0.83, 0.77, 0.83]
Longitudinal	Minimal	MinMax	AE pre-train	0.85 (0.04) [0.88, 0.88, 0.84, 0.85, 0.78]
	Extensive	MinMax	AE pre-train	0.86 (0.06) [0.88, 0.94, 0.85, 0.85, 0.76]

B. 3D subject-level CNN - sMCI vs pMCI

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Balanced accuracy
Baseline	Minimal	MinMax	AE pre-train	0.73 (0.05) [0.73, 0.73, 0.63, 0.77, 0.76]
Longitudinal	Minimal	MinMax	AE pre-train	0.73 (0.03) [0.73, 0.73, 0.67, 0.76, 0.74]

C. 3D ROI-based CNN - AD vs CN

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Balanced accuracy
Baseline	Minimal	MinMax	AE pre-train	0.88 (0.03) [0.84, 0.89, 0.90, 0.89, 0.85]
Longitudinal	Minimal	MinMax	AE pre-train	0.86 (0.02) [0.83, 0.86, 0.86, 0.88, 0.86]

D. 3D ROI-based CNN - sMCI vs pMCI

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Balanced accuracy
Baseline	Minimal	MinMax	AE pre-train	0.77 (0.05) [0.81, 0.81, 0.67, 0.78, 0.76]
Longitudinal	Minimal	MinMax	AE pre-train	0.78 (0.07) [0.87, 0.73, 0.68, 0.82, 0.78]

E. 3D patch-based CNN - AD vs CN

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Training approach	Balanced accuracy
Baseline	Minimal	MinMax	AE pre-train	single-CNN	0.74 (0.08) [0.75, 0.84, 0.78, 0.75, 0.59]
				multi-CNN	0.81 (0.03) [0.82, 0.84, 0.83, 0.77, 0.79]
Longitudinal	Minimal	MinMax	AE pre-train	single-CNN	0.76 (0.04) [0.78, 0.77, 0.80, 0.78, 0.69]
				multi-CNN	0.83 (0.02) [0.83, 0.85, 0.84, 0.82, 0.79]

F. 3D patch-based CNN - sMCI vs pMCI

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Training approach	Balanced accuracy
Baseline	Minimal	MinMax	AE pre-train	multi-CNN	0.75 (0.04) [0.80, 0.72, 0.72, 0.79, 0.72]
Longitudinal	Minimal	MinMax	AE pre-train	multi-CNN	0.77 (0.04) [0.77, 0.75, 0.71, 0.82, 0.79]

G. 2D slice-based CNN - AD vs CN

Training data	Image preprocessing	Intensity rescaling	Transfer learning	Balanced accuracy
Baseline	Minimal	MinMax	AE pre-train	0.79 (0.04) [0.83, 0.83, 0.72, 0.82, 0.73]
Longitudinal	Minimal	MinMax	AE pre-train	0.74 (0.03) [0.76, 0.80, 0.74, 0.71, 0.69]

H. Linear SVM - AD vs CN

Training data	Image preprocessing	Intensity rescaling	Balanced accuracy
Baseline	DartelGM	SPM-based	0.88 (0.02) [0.92, 0.89, 0.85, 0.89, 0.84]
Longitudinal	DartelGM	SPM-based	0.87 (0.01) [0.86, 0.86, 0.88, 0.87, 0.85]

I. Linear SVM - sMCI vs MCI

Training data	Image preprocessing	Intensity rescaling	Training task	Balanced accuracy
Baseline	DartelGM	SPM-based	sMCI vs MCI	0.68 (0.02) [0.71, 0.68, 0.66, 0.67, 0.71]
			AD vs CN	0.70 (0.06) [0.66, 0.75, 0.70, 0.79, 0.63]
Longitudinal	DartelGM	SPM-based	sMCI vs pMCI	0.68 (0.06) [0.75, 0.77, 0.62, 0.62, 0.67]
			AD vs CN	0.70 (0.02) [0.68, 0.72, 0.67, 0.69, 0.73]

than without (0.80). Based on this, we decided to always use AE pre-training, even though the difference is small.

We then assessed the influence of the amount of training data, comparing training using only baseline data to those with longitudinal data. The performance was moderately higher with longitudinal data (0.85) compared with baseline data only (0.82). We choose to continue exploring the influence of this choice because the four different approaches have a very different number of learnt parameters and the sample size is intrinsically augmented in the 2D slice-level and 3D single-CNN patch-level approaches.

We also assessed the influence of the preprocessing comparing the ‘Extensive’ and ‘Minimal’ preprocessing procedures. The performance was almost equivalent with the ‘Minimal’ preprocessing (0.85) and with the ‘Extensive’ preprocessing (0.86). Hence in the following experiments we kept the ‘Minimal’ preprocessing.

After the AD vs CN task, we trained the 3D subject-level CNN to perform the sMCI vs pMCI task (Table 3.4 B). The BA was the same for baseline data and for longitudinal data (0.73).

3.4.2.2 3D ROI-based

The next series of experiments was performed with the 3D ROI-based network, which focuses on the hippocampi (Table 3.4 C and D). For AD vs CN, the BA was 0.88 for baseline data and 0.86 for longitudinal data. This is slightly higher than that of the subject-level approach. For sMCI vs pMCI, the BA was 0.77 for baseline data and 0.78 for longitudinal data. This is substantially higher than with the 3D subject-level approach.

3.4.2.3 3D patch-level

The next experiments still use patches as inputs, but this time covering the whole brain instead of a particular region (Table 3.4 E and F). For the single CNN approach on AD vs CN, the BA was 0.74 for baseline data and 0.76 for longitudinal data. For the multi CNN approach on AD vs CN, the BA was 0.81 for baseline data and 0.83 for longitudinal data, thereby outperforming the single CNN approach. For sMCI vs pMCI, the BA was 0.75 for baseline data and 0.77 for longitudinal data. The performance for both tasks is slightly lower than that of the 3D ROI-based approach.

3.4.2.4 2D slice-level

After experiments with 3D networks, we studied a 2D network (Table 3.4 G). In general, the performance of the 2D slice-level approach was lower than that of the 3D ROI-based, 3D patch-level multi CNN and 3D subject-level (when trained with longitudinal data) approaches but higher than that of the 3D patch-level single CNN approach. The use of longitudinal data for training did not improve the performance (0.79 for baseline data; 0.74 for longitudinal data).

3.4.2.5 Linear SVM

The final series of experiments studies the performance of a linear SVM (Table 3.4 H and I). For the AD vs CN task, the BA was 0.88 when trained with baseline data and 0.87 when trained with longitudinal data. For the sMCI vs pMCI task, when training from scratch, the BA was 0.68 when trained with baseline data and 0.68 when trained with longitudinal data. When using transfer learning from the task AD vs CN to the task sMCI vs pMCI, the BA was 0.70 (when trained with baseline data) and 0.70 (when trained with longitudinal data). The performance of the SVM on AD vs CN is thus higher than that of most deep learning models and comparable to the best ones. However, for the sMCI vs pMCI task, the BA of the SVM is lower than that of deep learning models.

3.4.3 Results on the test sets

Results on the three test sets (ADNI, OASIS and AIBL) are presented in Table 3.5. For each category of approach, we only applied the best models for both baseline (a single image per subject acquired at baseline) and longitudinal (multiple images per subject acquired at different visits) data.

3.4.3.1 3D subject-level

For AD vs CN, all models generalised well to the ADNI and AIBL test sets but not to the OASIS test set (losing over 0.15 points of BA). For sMCI vs pMCI, the models generalised relatively well to the ADNI test set but not to the AIBL test set (losing over 0.20 point). Note that the generalisation was better for longitudinal than for baseline.

3.4.3.2 3D ROI-based

For AD vs CN, the models generalised well to the ADNI test set, slightly worse to the AIBL test set (losing 0.04 to 0.05 point) and considerably worse for OASIS (losing from 0.13 to 0.19 point). For sMCI vs pMCI, there was a slight decrease in BA on the ADNI test set and a severe decrease for the AIBL test set. Note that on the ADNI test set, the performance of the 3D ROI-based is almost the same as that of the 3D subject-level (when using longitudinal data) while it was better on the validation set.

3.4.3.3 3D patch-based

For AD vs CN, the generalisation pattern was similar to that of the other models: good for ADNI and AIBL, poor for OASIS. For sMCI vs pMCI, the BA on the ADNI test set was 0.07 point lower than on the ADNI validation set. The BA on the AIBL test set was very poor.

3.4.3.4 2D slice-level

For AD vs CN, there was a slight decrease in performance on the ADNI test set (losing from 0 to 0.03 point) and the AIBL test set (losing from 0.01 to 0.03 point) and a considerable decrease on the OASIS test set (losing from 0.13 to 0.14 point).

TABLE 3.5: Results on three independent test sets. 3D subject-level CNNs were trained using intensity rescaling and our “Minimal” preprocessing, with a data split on the subject level and transfer learning (AE pretraining for AD vs CN tasks and cross-task transfer learning was applied for sMCI vs pMCI tasks). For each model, we first specify the validation balanced accuracy (mean and standard deviation across the five folds). Then, we report the balanced accuracy for each test set (ADNI, AIBL, OASIS), more specifically within square brackets we report the balanced accuracy for each of the trained models of the 5 folds of the validation set and then the average across the five folds.

Task: AD vs CN									
Classification architectures	Image preprocessing	Intensity rescaling	Training approach	Transfer learning	Training data	ADNI val.	ADNI test	AIBL test	OASIS test
3D subject-level CNN	Minimal	MinMax	single-CNN	AE pre-train	Baseline	0.82 (0.05)	0.82 [0.79, 0.85, 0.82, 0.81, 0.85]	0.83 [0.81, 0.85, 0.84, 0.78, 0.86]	0.67 [0.59, 0.69, 0.72, 0.64, 0.69]
					Longitudinal	0.85 (0.04)	0.85 [0.88, 0.84, 0.84, 0.84, 0.84]	0.86 [0.89, 0.85, 0.86, 0.85, 0.86]	0.68 [0.65, 0.70, 0.70, 0.71, 0.65]
3D ROI-based CNN	Minimal	MinMax	single-CNN	AE pre-train	Baseline	0.88 (0.03)	0.89 [0.87, 0.88, 0.90, 0.91, 0.89]	0.84 [0.83, 0.88, 0.84, 0.85, 0.83]	0.69 [0.62, 0.74, 0.70, 0.69, 0.71]
					Longitudinal	0.86 (0.02)	0.85 [0.87, 0.82, 0.87, 0.86, 0.87]	0.81 [0.79, 0.81, 0.79, 0.82, 0.85]	0.73 [0.71, 0.73, 0.72, 0.76, 0.71]
3D patch-level CNN	Minimal	MinMax	multi-CNN	AE pre-train	Baseline	0.81 (0.03)	0.81 [0.82, 0.81, 0.84, 0.80, 0.79]	0.81 [0.81, 0.75, 0.81, 0.84, 0.82]	0.64 [0.61, 0.65, 0.60, 0.69, 0.67]
					Longitudinal	0.83 (0.02)	0.86 [0.86, 0.86, 0.87, 0.85, 0.84]	0.80 [0.82, 0.78, 0.81, 0.81, 0.79]	0.71 [0.70, 0.70, 0.71, 0.71, 0.67]
2D slice-level CNN	Minimal	MinMax	single-CNN	ImageNet pre-train	Baseline	0.79 (0.04)	0.76 [0.76, 0.75, 0.77, 0.75, 0.78]	0.76 [0.74, 0.76, 0.78, 0.75, 0.75]	0.65 [0.67, 0.62, 0.64, 0.65, 0.69]
					Longitudinal	0.74 (0.03)	0.74 [0.81, 0.76, 0.70, 0.74, 0.72]	0.73 [0.72, 0.77, 0.72, 0.66, 0.79]	0.61 [0.62, 0.63, 0.64, 0.58, 0.60]
SVM	DartelGM	SPM-based	None	None	Baseline	0.88 (0.02)	0.88 [0.88, 0.87, 0.90, 0.90, 0.88]	0.88 [0.87, 0.90, 0.87, 0.89, 0.90]	0.70 [0.71, 0.71, 0.70, 0.68, 0.72]
					Longitudinal	0.87 (0.01)	0.87 [0.85, 0.84, 0.90, 0.89, 0.87]	0.87 [0.88, 0.86, 0.88, 0.87, 0.89]	0.71 [0.73, 0.68, 0.72, 0.70, 0.71]

Task: sMCI vs pMCI

Classification architectures	Image preprocessing	Intensity rescaling	Training approach	Transfer learning	Training data	ADNI val.	ADNI test	AIBL test
3D subject-level CNN	Minimal	MinMax	single-CNN	AE pre-train	Baseline	0.73 (0.05)	0.69 [0.68, 0.71, 0.64, 0.73, 0.67]	0.52 [0.51, 0.47, 0.55, 0.54, 0.55]
					Longitudinal	0.73 (0.03)	0.73 [0.75, 0.72, 0.72, 0.74, 0.72]	0.50 [0.48, 0.47, 0.54, 0.52, 0.51]
3D ROI-based CNN	Minimal	MinMax	single-CNN	AE pre-train	Baseline	0.77 (0.05)	0.74 [0.75, 0.72, 0.76, 0.75, 0.75]	0.60 [0.56, 0.56, 0.66, 0.62, 0.59]
					Longitudinal	0.78 (0.07)	0.74 [0.70, 0.73, 0.73, 0.75, 0.81]	0.57 [0.56, 0.53, 0.52, 0.66, 0.56]
3D patch-level CNN	Minimal	MinMax	multi-CNN	AE pre-train	Baseline	0.75 (0.04)	0.68 [0.71, 0.64, 0.64, 0.71, 0.69]	0.64 [0.63, 0.52, 0.67, 0.74, 0.63]
					Longitudinal	0.77 (0.04)	0.70 [0.70, 0.71, 0.69, 0.71, 0.69]	0.44 [0.45, 0.39, 0.55, 0.42, 0.39]
SVM	DartelGM	SPM-based	None	None	Baseline	0.70 (0.06)	0.75 [0.75, 0.75, 0.74, 0.76, 0.76]	0.60 [0.62, 0.54, 0.62, 0.59, 0.64]
					Longitudinal	0.70 (0.02)	0.76 [0.74, 0.75, 0.80, 0.77, 0.76]	0.68 [0.67, 0.66, 0.68, 0.67, 0.71]

MinMax: for CNNs, intensity rescaling was performed based on min and max values, resulting in all values to be in the range of [0, 1]; SPM-based: the SPM-based grey matter maps are intrinsically rescaled; AE: autoencoder.

3.4.3.5 Linear SVM

For AD vs CN, we observed the same pattern as for the other models: excellent generalisation to ADNI and AIBL but not to OASIS. For sMCI vs pMCI, the generalisation was excellent for ADNI but not for AIBL. Of note, the BA on the ADNI test set was even higher to that of the validation, reaching a level which is comparable to the best deep learning models.

3.4.4 Discussion

In this section, we rigorously compared different CNN approaches and studied the impact of key components on the performance. We hope that these results will provide a more objective assessment of the performance of CNNs for AD classification and constitute a solid baseline for future research.

The proposed framework was applied to images from three public data sets, ADNI, AIBL and OASIS. On the ADNI test data set, the diagnostic BA of CNNs ranged from 0.76 to 0.89 for the AD vs CN task and from 0.69 to 0.74 for the sMCI vs pMCI task. These results are in line with the state-of-the-art, where classification accuracy typically ranged from 0.76 to 0.91 for AD vs CN and 0.62 to 0.83 for sMCI vs pMCI (Wen et al., 2020). Nevertheless, the performance that we report is lower than that of the top-performing studies. This potentially comes from the fact that our test set was fully independent and was never used to choose the architectures or parameters. The proposed framework can be used to provide a baseline performance when developing new methods.

One interesting question is whether deep learning could perform better than conventional machine learning methods for AD classification. Here, we chose to compare CNN to a linear SVM. In the current study, the SVM was at least as good as the best CNNs for both the AD vs CN and the sMCI vs pMCI task. Note that we used a standard linear SVM with standard voxel-based features. It could be that more sophisticated conventional machine learning methods could provide even higher performance. Similarly, we do not claim that more sophisticated deep learning architectures would not outperform the SVM. However, this is not the case with the architectures that we tested, which are representative of the existing literature on AD classification. Besides, it is possible that CNNs will outperform SVM when larger public data sets will become available. Overall, a major result of the present paper is that, with the sample size which is available in ADNI, CNNs did not provide an increase in performance compared to SVM.

Unbiased evaluation of the performance is an essential task in machine learning. This is particularly critical for deep learning because of the extreme flexibility of the models and of the numerous architecture and training hyperparameters that can be chosen. In particular, it is crucial that such choices are not made using the test set. We chose a very strict validation strategy in that respect: the test sets were left untouched until the end of the peer-review process. This guarantees that only the final models, after all possible adjustments, are carried to the test set. Moreover, it is important to assess generalisation not only to unseen subjects but also to other studies in which image acquisitions or patient inclusion criteria can vary. In the present work, we used three test sets from the ADNI, AIBL and OASIS databases to assess different generalisation aspects.

We studied generalisation in three different settings: i) on a separate test set from ADNI, thus from the same study as those of the training set; ii) on AIBL, i.e., a different study but with similar inclusion criteria and imaging acquisitions; iii) on OASIS, i.e., a study with different inclusion criteria and imaging acquisitions. Overall, the models generalised well to ADNI (for both tasks) and to AIBL (for AD vs CN). On the other hand, we obtained a very poor generalisation to sMCI vs pMCI for AIBL. We hypothesise that it could be because pMCI and sMCI participants from AIBL are substantially older than those of ADNI, which is not the case for AD and CN participants. Nevertheless, note that the sample size for sMCI vs pMCI in AIBL is quite small (33 participants). Also, the generalisation to OASIS was poor. This may stem from the diagnosis criteria which are less rigorous (in OASIS, all participants with $CDR > 0$ are considered AD). Overall, these results bring important information. First, good generalisation to unseen, similar, subjects demonstrate that the models did not overfit the subjects at hand in the training/validation set. On the other hand, poor generalisation to different age, protocols and inclusion criteria show that trained models are too specific of these characteristics. Generalisation across different populations thus remains an unsolved problem and will require training on more representative data sets but maybe also new strategies to make training more robust to heterogeneity. This is critical for the future translation to clinical practice in which conditions are much less controlled than in research data sets like ADNI.

We studied the influence of several key choices on the performance. First, we studied the influence of AE pre-training and showed that it slightly improved the average over training from scratch. Three previous papers studied the impact of AE pre-training (Hosseini-Asl et al., 2016; Vu et al., 2018; Vu et al., 2017) and found that it improved the results. However, they are all suspected of data leakage. We thus conclude that, to date, it is not proven that AE pre-training leads to a significant increase in BA. A difficulty in AD classification using deep learning is the limited amount of data samples available for training. However, training with longitudinal instead of baseline data gave only a slight increase of BA in most approaches. The absence of a major improvement may be due to several factors. First, training with longitudinal data implies training with data from more advanced disease stages, since patients are seen at a later point in the disease course. This may have an adverse effect on the performance of the model when tested on baseline data, at which the patients are less advanced. Also, since the additional data come from the same patients, this does not provide a better coverage of inter-individual variability. We studied the impact of image preprocessing. First, as expected, we found that CNNs cannot be successfully trained without intensity rescaling. We then studied the influence of two different preprocessing procedures ('Minimal' and 'Extensive'). The 'Minimal' procedure is limited to an affine registration of the subject's image to a standard space, while for the 'Extensive' procedure non-linear registration and skull stripping are performed. They led to comparable results. In principle, this is not surprising as deep learning methods do not require extensive preprocessing. In the literature, varied types of preprocessing have been used. Some studies used non-linear registration (Bäckström et al., 2018; Lin et al., 2018; Liu et al., 2018c; Wang et al., 2018; Lian et al., 2018; Liu et al., 2018b; Basaia et al., 2019; Wang et al., 2019) while others used only linear (Hosseini Asl et al., 2018;

Aderghal et al., 2018; Li et al., 2018; Liu et al., 2018a; Shmulev et al., 2018; Aderghal et al., 2017a; Aderghal et al., 2017b) or no registration (Cheng et al., 2017). None of them compared these different preprocessings with the exception of (Bäckström et al., 2018) which compared preprocessing using FreeSurfer to no preprocessing. They found that training the network with the raw data resulted in a lower classification performance (drop in accuracy of 38 percent points) compared to the preprocessed data using FreeSurfer (Bäckström et al., 2018). However, FreeSurfer comprises a complex pipeline with many preprocessing steps so it is unclear, from their results, which part drives the superior performance, while we clearly demonstrated that the intensity rescaling is essential for the CNN training whereas there is no improvement in using a non-linear registration over a linear one. Finally, we found that, for the 3D patch-level framework, the multi-CNN approach gave better results than the single-CNN one. However, this may be mainly because the multi-CNN approach benefits from a thresholding system which excludes the worst patches, a system that was not present in the single-CNN approach. To test this hypothesis, we performed supplementary experiments in which the multi-CNN was trained without threshold and the single-CNN was trained using the same thresholding system as in the main experiments of the multi-CNN. We observed that the results of the multi-CNN and the single-CNN are comparable when they use the same thresholding system. These supplementary experiments suggest that, under similar conditions, the multi-CNN architecture does not always perform better than the single-CNN architecture. In light of this, it would seem preferable to choose a framework that offers a better compromise between performance and conceptual complexity, e.g., the ROI-based or the 3D subject-level approaches.

Our study has the following limitations. First, a large number of options exist when choosing the model architecture and training hyperparameters. Even though we did our best to make meaningful choices and test a relatively large number of possibilities, we cannot exclude that other choices could have led to better results. To overcome this limitation, we provided an open-source framework. Researchers can use it to propose and validate potentially better performing models. In particular, with this framework, researchers can easily try their own models without touching the test data sets. Secondly, the cross-validation procedures were performed only once. Of course, the training is not deterministic and one would ideally want to repeat the cross-validation to get a more robust estimate of the performance. However, we did not perform this due to limited computational resources. Finally, overfitting always exists in our experiments, even though different techniques have been tried (e.g., transfer learning, dropout or weight decay). This phenomenon occurs mainly due to the limited size of the data sets available for AD classification. It is likely that training with much larger data sets would result in higher performance.

3.5 Open-source contributions

With both the AD-ML and AD-DL frameworks, we aimed to contribute to make evaluation of machine learning approaches in AD more reproducible and more objective. Reproducibility is the ability to reproduce results based on the same data and experimental procedures. Calls to increase reproducibility have been made in different fields, including neuroimaging

(Poldrack et al., 2017) and machine learning (Ke et al., 2017). Reproducibility differs from replication, which is the ability to confirm results on independent data. Key elements of reproducible research include: data sharing, storing of data using community standards, fully automatic data manipulation, sharing of code. Our work can contribute to increase reproducibility of machine learning research for computer-aided diagnosis through different aspects. A first component is the fully automatic conversion of three public data sets into the community standard BIDS. Indeed, ADNI and AIBL cannot be redistributed. Through these tools, we hope to make it easy to reproduce experiments based on these data sets without redistributing them. In particular, we offer a huge saving of time to users compared to simply making public the list of subjects used. This is particularly true for complex multimodal data sets such as ADNI (with plenty of incomplete data, multiple instances of a given modality and complex metadata). The second key component is publicly available code for preprocessing, feature extraction and classification. These contributions are gathered in Clinica and ClinicaDL, two freely available software platforms for clinical neuroscience research studies that will be described in more details in the next sections.

We also hope to contribute to more objective evaluations. Objective evaluation of a new approach (classification algorithm, preprocessing pipeline or other) requires testing this specific component without changing the others. Our frameworks include standard approaches for the preprocessing of T1-weighted MRI (and PET) images, and standard classification tools. These constitute a set of baseline approaches against which new methods can easily be compared. Researchers working on novel methods can then straightforwardly replace a given part of the pipeline (e.g., classifier) with their own solution, and evaluate the added value of this specific new component over the baseline approach provided. We also propose tools for rigorous validation. For machine learning experiments, these include: i) large number of repeated random split to extensively assess the variability in performance; ii) reporting the full distribution of accuracies and standard deviation rather than only mean accuracies; iii) adequate nested CV for hyperparameter tuning (Varoquaux et al., 2017). Regarding deep learning experiments, our tools consists in preventing and detecting potential data leakage, i.e., the use of test data in any part of the training process.

3.5.1 Clinica: Software platform for neuroimaging studies

Clinica is an open-source software platform for reproducible clinical neuroimaging studies that aims to make neuroimaging research studies easier and pursues the community effort of reproducibility (Routier et al., 2021, www.clinica.run). The core of Clinica is a set of automatic pipelines for processing and analysis of multimodal neuroimaging data (currently, T1w MRI, diffusion MRI and PET data), as well as tools for statistics and machine learning. Clinica relies on tools written by the scientific community and provides converters of public neuroimaging data sets to BIDS, processing pipelines and organisation for processed files, statistical analysis, and machine learning algorithms. A schematic overview of Clinica can be found in Figure 3.10.

The target audience is mainly of two types. First, neuroscientists or clinicians conducting clinical neuroscience studies involving multimodal imaging, typically not experts in image processing for all of the involved imaging modalities or in statistical analysis. They will

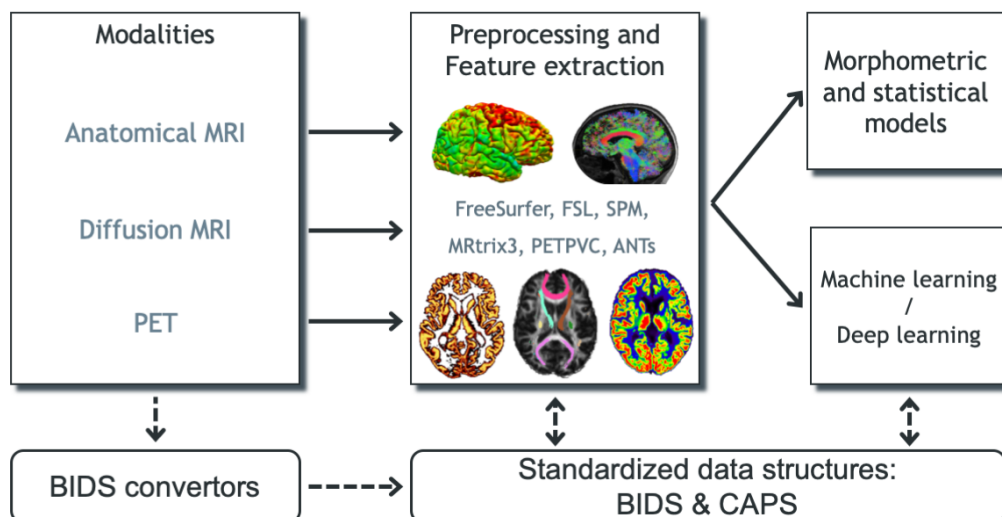


FIGURE 3.10: Overview of Clinica’s functionalities. Clinica provides processing pipelines for MRI and PET images that involve the combination of different software packages, and whose outputs can be used for statistical or machine learning analysis. Clinica expects data to follow the Brain Imaging Data Structure (BIDS) and provides tools to convert public neuroimaging data sets into the BIDS format. Output data are stored using the Clinica Processed Structure (CAPS).

benefit from a unified set of tools covering the complete set of steps involved in a study (from raw data to statistical analysis). Second, researchers developing advanced machine learning algorithms, typically not experts in brain image analysis. They will benefit from tools to convert public datasets into BIDS, fully automatic feature extraction methods, and baseline classification algorithms to which they could compare their results. Overall, we hope that Clinica will allow users to spend less time on data management and processing, to perform reproducible evaluations of their methods, and to easily share data and results within their institution and with external collaborators.

Clinica can take as inputs different neuroimaging modalities, currently anatomical MRI, diffusion MRI and PET and provides processing pipelines that involve the combination of different software packages. It currently relies on FreeSurfer (Fischl, 2012), FSL (Jenkinson et al., 2012), SPM (Ashburner, 2012), Advanced normalisation Tools (ANTs) (Avants et al., 2014), MRtrix3 (Tournier et al., 2019), and the PET Partial Volume Correction (PETPVC) toolbox (Thomas et al., 2016). The pipelines are written using Nipype (Gorgolewski et al., 2011). Features extracted with the different pipelines can be used as inputs to statistical analysis, which relies on SPM and SurfStat (Worsley et al., 2009), or machine learning analysis, which relies on scikit-learn (Pedregosa et al., 2011). The list of pipelines currently available in Clinica is presented in Figure 3.11.

Input neuroimaging data are expected to follow the BIDS data structure (Gorgolewski et al., 2016). Since this new standard has only recently been adopted by the community, not all public neuroimaging data sets are yet proposed in BIDS format. To facilitate the adoption of BIDS, Clinica curates several publicly available neuroimaging data sets and provides tools to convert them into the BIDS format. Processed data are organised following the Clinica Processed Structure (CAPS) format, which shares the same philosophy as BIDS. Finally, a set of tools is provided to handle input and output data generated by Clinica,


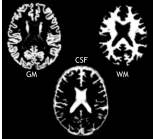
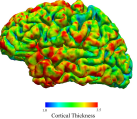
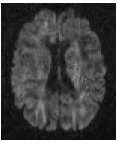
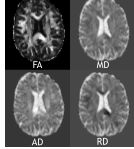
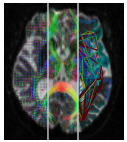
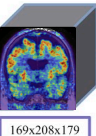
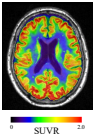
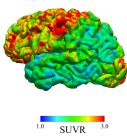

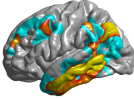
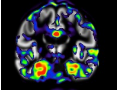

Anatomical MRI	<p>t1-linear Bias field correction, affine registration and cropping Dependencies: ANTs</p> <ul style="list-style-type: none"> • T1 MRI on ICBM 2009c nonlinear symmetric template • Used as input for deeplearning-prepare-data  <p>169x208x179</p>	<p>t1-volume Tissue segmentation (GM, WM, CSF), inter-subject registration using Dartel, spatial normalization to standard space (MNI) Dependencies: SPM, CAT12</p> <ul style="list-style-type: none"> • Voxel-based features (GM, WM, CSF) • Regional features (average GM) using atlases (currently AALZ, AICHA, Hammers, LPBA40, Neuromorphometrics) 	<p>t1-freesurfer t1-freesurfer-longitudinal Cortical surface extraction, segmentation of subcortical structures, cortical thickness estimation, spatial normalization to standard space (FsAverage) Dependencies: FreeSurfer</p> <ul style="list-style-type: none"> • Surface-based features (cortical thickness) • Regional features (average cortical thickness) using atlases (currently Desikan, Destrieux)  <p>Cortical Thickness</p>
Diffusion MRI	<p>dwi-preprocessing Correction of raw DWI data Dependencies: FSL, ANTs, Convert3D</p> <ul style="list-style-type: none"> • EPI correction using phase-difference map fieldmap or T1w ("fieldmap-less") • Prerequisite for dwi-dti or dwi-connectome pipelines 	<p>dwi-dti Extraction of DTI-based measures, normalization to standard space (MNI) Dependencies: FSL, ANTs, MRtrix3</p> <ul style="list-style-type: none"> • Voxel-based features (FA, MD, AD, RD) • Regional features (average FA, MD, AD, RD) using atlases (JHU DTI81, JHUTracts) 	<p>dwi-connectome Tractography & connectome Dependencies: FreeSurfer, FSL, MRtrix</p> <ul style="list-style-type: none"> • Probabilistic tractography • Structural connectome using atlases (currently Desikan, Destrieux) 
PET	<p>pet-linear Affine registration, intensity normalization and cropping Dependencies: ANTs</p> <ul style="list-style-type: none"> • PET on ICBM 2009c nonlinear symmetric template • Used as input for deeplearning-prepare-data  <p>169x208x179</p>	<p>pet-volume Registration to T1 MRI, partial volume correction, spatial normalization to standard space (MNI), intensity normalization Dependencies: SPM, PETPVC, CAT12</p> <ul style="list-style-type: none"> • Voxel-based features (e.g. FDG uptake, amyloid uptake) • Regional features (average FDG, amyloid uptake) using atlases (currently AALZ, AICHA, Hammers, LPBA40, Neuromorphometrics)  <p>SUVR</p>	<p>pet-surface pet-surface-longitudinal Registration to T1 MRI, intensity normalization, partial volume correction, projection to cortical surface, spatial normalization to standard space (FsAverage) Dependencies: FreeSurfer, FSL, SPM, PETPVC</p> <ul style="list-style-type: none"> • Surface-based features (e.g. FDG uptake, amyloid uptake) • Regional features (average cortical thickness) using atlases (currently Desikan, Destrieux)  <p>SUVR</p>
Statistics	<p>statistics-volume Voxel-based mass-univariate analysis with SPM Dependencies: SPM, Matlab</p> <ul style="list-style-type: none"> • Voxel-based features from t1-volume or pet-volume pipelines • Group comparison using GLM 	<p>statistics-surface Surface-based mass-univariate analysis with SurfStat Dependencies: Matlab</p> <ul style="list-style-type: none"> • Surface-based features from t1-freesurfer or pet-surface pipelines • Group comparison or correlations analysis using GLM 	
Machine Learning	<p>machinelearning-prepare-spatial-svm Preparation of T1 MRI and PET data for spatially regularized SVM Dependencies: None</p> <ul style="list-style-type: none"> • Regularization that accounts for the spatial and anatomical structure of neuroimaging data leading to a more regular and anatomically interpretable decision function. • Used as input for machine learning classification 	<p>(No command line interface) Classification based on machine learning Dependencies: None</p> <ul style="list-style-type: none"> • Voxel-based, surface-based or regional features • Classifications (SVM, ℓ_2 logistic regression, random forest) using cross-validations (K-fold, repeated K-fold, repeated hold-out) 	

FIGURE 3.11: List of the pipelines currently available in Clinica with their dependencies and outputs. GM: grey matter; CSF: cerebrospinal fluid; WM: white matter; FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; RD: radial diffusivity, SVM: Support Vector Machine; ICBM: International Consortium for Brain Mapping.

thus facilitating data management or connection to statistical or machine learning analysis.

The core of Clinica is written in Python and mainly relies on the Nipype framework (Gorgolewski et al., 2011) to create pipelines. Python dependencies also include NumPy (van der Walt et al., 2011), NiBabel (Brett et al., 2019), Pandas (McKinney, 2010), NIPY, SciPy (Jones et al., 2001), scikit-learn (Pedregosa et al., 2011), scikit-image (van der Walt et al., 2014) and Nilearn (Abraham et al., 2014). Clinica is provided to the end user in the form of a Python package distributed through Python Package Index (PyPI)⁷ and can simply be installed by typing `pip install clinica` through the terminal, within a virtual environment. The main usage of Clinica is through the command line, which is facilitated by the support of autocompletion. The commands are gathered into four main categories. The first category of command line (`clinica run`) allows the user to run the different pipelines on neuroimaging datasets following a BIDS or CAPS hierarchy. The `clinica convert` category allows the conversion of publicly available neuroimaging datasets into a BIDS hierarchy. To help with data management, the `clinica iotools` category comprises a set of tools that allows the user to handle BIDS and CAPS datasets, including generating lists of subjects or merging all tabular data into a single TSV file for analysis with external statistical software packages. Finally, the last category (`clinica generate`) is dedicated to developers and currently generates the skeleton for a new pipeline.

Clinica was first hosted on a GitLab instance managed by the Paris Brain Institute and was moved to GitHub in 2019 (<https://github.com/aramis-lab/clinica>). At the time of writing, it counts 164 stars and 52 forks. Since Clinica has been hosted on GitHub, about 200 issues have been opened, including 65 by external users. More than 140 discussions have been opened on the Google Group (<https://groups.google.com/g/clinica-user>). The journal article presenting Clinica (Routier et al., 2021) has been cited about fifty times. 23 citations correspond to articles published by members of the Paris Brain Institute who explicitly mention the use of the software, and 14 by external users.

In conclusion, Clinica is an open-source software platform that provides a comprehensive set of processing pipelines for different neuroimaging modalities. It builds upon existing standards and software tools developed by the community. It can make clinical neuroimaging studies easier to perform and more reproducible.

3.5.2 ClinicaDL: Software for reproducible neuroimaging processing with deep learning

ClinicaDL is an open-source software platform entirely written in Python that includes many functionalities, such as neuroimaging preprocessing, synthetic dataset generation, label definition, data split with similar demographics, architecture search, network training, performance evaluation and trained network interpretation. The three main objectives of ClinicaDL are to (1) help manipulate neuroimaging data sets, (2) prevent data leakage from biasing results and (3) reproduce deep learning experiments.

⁷<https://pypi.org/project/clinica>

First, ClinicaDL relies on standardised formats, the BIDS and CAPS, to organise raw and processed data, respectively. Though these formats were first introduced for neuroimaging data management, they can be easily extended to any kind of medical imaging data, as it would only require renaming and formatting files of a data set.

Secondly, ClinicaDL prevents data leakage as train and validation data characteristics are saved when the output structure is created. Then, when evaluating the performance of a trained model on a new data group, ClinicaDL checks that this data group is independent from the training and validation groups. However, this only works under the assumption that participants are always named in the same way across data groups.

Thirdly, ClinicaDL improves deep learning experiment reproducibility by sharing usable and tagged code, saving all parameters of the training set and data groups used for evaluation, and providing extensive documentation. However, though all these elements improve method reproducibility, reproducibility can still be easily broken. For example Crane, 2018 explained that using another GPU system may make the results irreproducible. Then it may not be possible for two different users to obtain the same results on different machines. However, one user may be interested in having a deterministic setting to correctly evaluate the impact of one particular property to improve their performance. Moreover, result reproducibility may also be broken by manual architecture search and the overuse of the same data set (Thompson et al., 2020). Indeed, research studies may be globally overfitting this data set and if one day another data set is released, performance of previous studies may collapse. This is why we implemented the random search method, although its very high computational cost may limit its reproducibility power. In conclusion, as reproducibility is a property which may be broken by many aspects of a study, we advise data scientists to refer to reproducibility checklists made available online⁸ to ensure that their work is (largely) reproducible.

ClinicaDL uses the PyTorch library as backbone (Paszke et al., 2019) and extends PyTorch for neuroimaging applications where the data set structure plays a key role in the organisation of the data and metadata. The software is publicly distributed as an easy-to-install package and is referenced in the Pypi package index⁹. Releases are performed on a periodic basis and the code follows the most standard current practices for software development.

ClinicaDL has been designed to be used via the command line interface, with separate sub-commands performing the main tasks, as defined in a classical machine learning pipeline: `extract`, `train`, `predict`. Other sub-commands are available in order to allow the user to structure the data sets, create synthetic data, look for hyperparameters and interpret trained networks. These functionalities are also available through the command line (`tsvtool`, `generate`, `random-search`, `interpret`).

The main functionalities of ClinicaDL cover all the steps needed for deep learning experiments, from data set management to the evaluation of results and network interpretation. ClinicaDL's workflow is illustrated in Figure 3.12. In addition to pre-implemented options,

⁸<https://miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf>

⁹<https://pypi.org/project/clinicadl>

the source code aims at being modular and the documentation helps users to easily implement their custom experiments. Technical details for each command can be found in the user documentation (<https://clinicadl.readthedocs.io>).

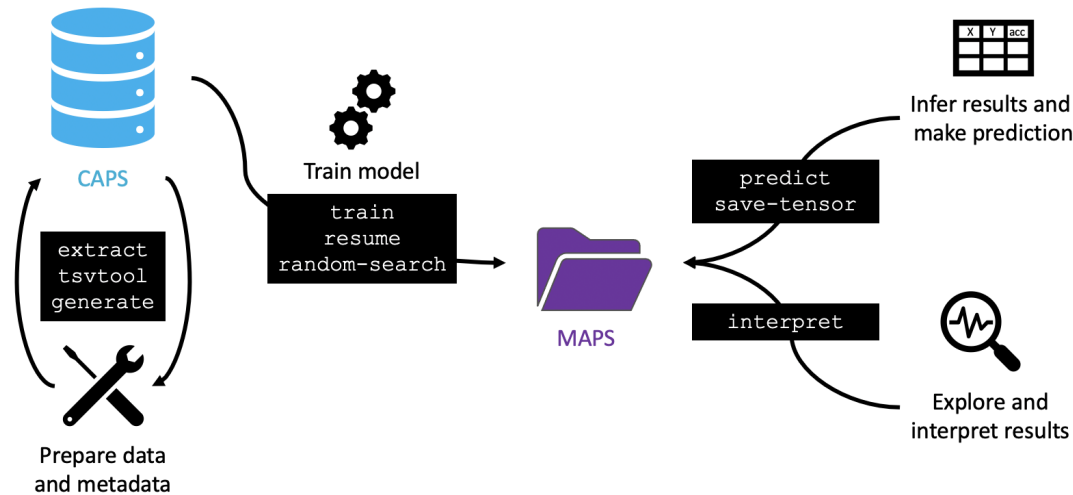


FIGURE 3.12: ClinicaDL main functionalities. `extract`, `tsvtool` and `generate` functionalities read and write in the Clinica Processed Structure (CAPS), which contains neuroimaging data preprocessed by Clinica pipelines. ClinicaDL writes its own output, the Model Analysis and Processing Structure (MAPS), which contains the results of the training phase as well as inference on new data or the results of interpretability methods.

ClinicaDL has been hosted on GitHub since 2020 (<https://github.com/aramis-lab/clinicadl>). At the time of writing, it counts 121 stars and 44 forks. About 130 issues have been opened, including 25 by external users. The journal article presenting ClinicaDL (Thibeau-Sutre et al., 2022b) has been cited nine times in eight months, until now only by members of the team.

Preprocessing of images ClinicaDL works preferably with images that have been previously preprocessed but one can also perform experiments with unprocessed images, the only requirement is to convert these images to the right format. Preprocessed images can be obtained using Clinica for different imaging modalities. For example, the `t1-linear` pipeline mainly performs bias field correction and spatial normalisation to the MNI space of T1-weighted MR images, while the `pet-linear` pipeline mainly performs spatial normalisation to the MNI space and intensity normalisation of PET images. As ClinicaDL and Clinica are fully compatible, outputs of the formerly mentioned pipelines can be introduced easily into `train` or `predict` functions of ClinicaDL.

ClinicaDL proposes a simple tool to transform NIfTI images into PyTorch tensors. The objective is to facilitate the training phase by decompressing the images beforehand (the NIfTI format usually provides compressed images). This functionality writes future input images for neural network training or inference formatted as tensors. The number and shape of these tensors depend on the mode chosen. Four possible uses of the image (modes) are currently implemented in ClinicaDL:

1. `image` uses the whole 3D image,

2. **patch** extracts 3D cubic patches with predefined size and stride to cover the whole image,
3. **roi** extracts specific 3D regions defined by binary masks generated by the user,
4. **slice** extracts 2D slices according to a neuroanatomical plane (sagittal, coronal or axial).

The tool will run through the entire CAPS/BIDS folder searching for an imaging modality specified by the user and will apply the conversion and extraction of corresponding images. It will also produce a configuration file summarising all the characteristics of the extraction procedure. The training procedure will then rely on this file to find the images needed for network training.

Generation of toy data sets ClinicaDL facilitates the generation of semi-synthetic data for evaluation and verification purposes. Such synthetic data sets have different purposes: they can be used to debug the functionalities of ClinicaDL and are used by the continuous integration workflow. Trivial data can also be used to ensure that a network is able to learn something from images with similar characteristics but simpler than the true ones that will be used. Finally, they can be used to simulate data sets with different properties (e.g., class imbalances).

Preparation of metadata To use the train and inference functionalities of the software or to analyse the data, inputs must be organised in the right way. A collection of tools to handle metadata of BIDS-formatted data sets is proposed with ClinicaDL. These tools are intended to provide the correct organisation of the data: get the labels used in classification tasks, split the data to define test, validation and train subsets, and analyse the population of interest. This set of commands is available through the command `clinica dl tsvtool`.

Random search Random search (Bergstra et al., 2012) is a procedure to find automatically the hyperparameters (architecture and other training hyperparameters) of a framework. It consists in randomly generating sets of hyperparameters to select the best set of hyperparameters as a result. This random generation is based on a hyperparameter space from which hyperparameter sets are sampled. In ClinicaDL, this hyperparameter space is described by a configuration file created by the user. The main advantage of the random search is its easy parallelisation, contrary to other optimisation methods that may require successive runs and be time consuming. On the other hand, it is computationally costly and it requires minimum knowledge regarding the subspace of hyperparameters that may work to limit the search and find satisfactory results. Moreover, although it can significantly improve the performance of a framework, it will not lead to the optimum, which is very hard to find.

Training networks The main functionality of ClinicaDL is to train neural networks to learn a task. These tasks can be:

1. **Classification** (of a categorical label, for example the diagnosis),
2. **Regression** (of a continuous label, for example the age),

3. Image reconstruction.

Segmentation is currently not handled by ClinicaDL. However, as the software is meant to be extensible, new tasks can be easily added by advanced users.

These tasks are highly dependent from the architecture. Some pre-built deep learning architectures for each task are available in ClinicaDL and their list and details can be displayed with the command `clinica dl train list_models`. However, an objective of the library is to allow the users to add and use their custom architectures easily. To this end, users can implement their custom networks by filling the abstract template, which includes specific methods that are used in ClinicaDL. The procedure of such addition is detailed in the documentation.

Performance evaluation ClinicaDL provides specific functions to easily perform inference with models previously trained with the tool. This functionality is available in a specific sub menu of the command line (`clinica dl predict`). For example, one may want to evaluate the performance of a trained model on a set of new samples. In this case, the command will load the best model, the input images (in a BIDS/CAPS-like format) and the list of subjects of the data group. Trained models are available within the Model Analysis and Processing Structure (MAPS) produced during the training and the other information can be either integrated into this structure or provided as a command line option. The results are written in the MAPS as pre-formatted reports with the metric values at different levels (e.g., image-level and patch-level) and the output values computed for each input image of the data group.

The metrics computed depend on the task learnt by the network. The regression and reconstruction tasks are associated with the mean squared error and mean absolute error, whereas the classification task is evaluated thanks to balanced accuracy, accuracy, sensitivity, specificity, positive and negative predictive values. Advanced users can add any new metric by following the procedure described in the advanced user guide. Moreover as the output values are computed for each input image individually, users can easily compute any metric of evaluation without modifying the source code.

Interpretation The most critical issue of deep learning methods is their lack of transparency. This is why some interpretability methods have been developed specifically for the field. These methods allow better understanding which patterns or zones of the images have been linked to the result produced by the network. Currently, only the gradient back-propagation method proposed in (Simonyan et al., 2014) is implemented in ClinicaDL. We plan to strengthen the content of this command in future releases.

In conclusion, with ClinicaDL we help deep learning users handling the three main issues encountered by non-specialists of the neuroimaging domain: (1) the data management and preprocessing of neuroimaging data sets, (2) the contamination of results by data leakage and (3) the lack of reproducibility of deep learning experiments.

3.6 Conclusion and perspectives

Machine learning is now widely used in the neuroimaging community, for cognitive neuroscience or computer-aided diagnosis applications. However, applying such approaches to neuroimaging data can be difficult for newcomers. Conversely, researchers in machine learning and deep learning are often interested in applying and validating their approaches to neuroimaging problems (e.g., diagnosis, prognosis). However, they often lack the necessary knowledge for handling and preprocessing neuroimaging data. Both Clinica and ClinicaDL are being developed to assist both types of users.

Reproducibility has been highlighted as a major challenge in many scientific fields including neuroimaging (Poldrack et al., 2017). This problem has also been highlighted in machine learning for healthcare in general (McDermott et al., 2021) and for brain diseases in particular (Samper-González et al., 2018; Wen et al., 2020). Clinica and ClinicaDL aim to make reproducible research easier to perform. To that purpose, they combine the use of a community standard for inputs, the definition of a standardised organisation for outputs, standardised ways to preprocess imaging data, detailed documentations, and extensive software testing. They extensively relies on community achievements such as the BIDS standard (Gorgolewski et al., 2016) and Nipype (Gorgolewski et al., 2011).

Clinica and ClinicaDL are under active development. For example, we aim to improve the reproducibility of Clinica by adding traceability features. Moreover, quality control of processed data is currently done using standard image viewers which is clearly suboptimal. We plan to add more advanced quality control of outputs in the spirit for instance of MindControl (Keshavan et al., 2018). This is important in order to ease the quality control of large data sets and to enforce the good practice of systematic quality control among the users. Implementing an integrated QC system is among our priorities for the development of Clinica. The aim is to provide a visual dashboard that would allow the user to: 1) easily control which pipelines have been executed and whether they exited without error; 2) systematically review snapshots of the major outputs of each pipeline (together with typical examples of how a correct output should look like); 3) flag incorrect outputs so that they can be excluded from further statistical analysis. ClinicaDL is convenient for research experiments but is not scalable enough and lacks tools for model deployment. We aim to integrate deep learning tools widely adopted by the community, such as MLflow and Pytorch Lightning, to improve model management and experiment tracking. This would make ClinicaDL more accessible and flexible by converging towards community standards.

Chapter 4

Computer-aided diagnosis of dementia from routine clinical data

This chapter results from the PhD work of Simona Bottani, who I co-supervised with Olivier Colliot. Corresponding publications:

- Bottani, S., **Burgos, N.**, Maire, A., Wild, A., Ströer, S., Dormont, D., Colliot, O. and the APPRIMAGE Study Group¹: ‘Automatic Quality Control of Brain T1-Weighted Magnetic Resonance Images for a Clinical Data Warehouse’, *Medical Image Analysis*, 75: 102219, 2022. [doi:10.1016/j.media.2021.102219](https://doi.org/10.1016/j.media.2021.102219) • [hal-03154792](https://hal.archives-ouvertes.fr/hal-03154792)
- Bottani, S., Thibeau-Sutre, E., Maire, A., Ströer, S., Dormont, D., Colliot, O., **Burgos, N.** and the APPRIMAGE Study Group¹: ‘Homogenisation of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-Net derived models’. In *SPIE Medical Imaging 2022*, 12032:576–582, 2022. [doi:10.1117/12.2608565](https://doi.org/10.1117/12.2608565) • [hal-03478798](https://hal.archives-ouvertes.fr/hal-03478798)
- Bottani, S., Thibeau-Sutre, E., Maire, A., Ströer, S., Dormont, D., Colliot, O., **Burgos, N.** and the APPRIMAGE Study Group¹: ‘Homogenisation of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation’. Submitted. [hal-03497645](https://hal.archives-ouvertes.fr/hal-03497645)
- Bottani, S., **Burgos, N.**, Maire, A., Saracino, D., Ströer, S., Dormont, D., and Colliot, O. and the APPRIMAGE Study Group¹: ‘Evaluation of MRI-based Machine Learning Approaches for Computer-Aided Diagnosis of Dementia in a Clinical Data Warehouse’. Under revision at *Medical Image Analysis*. [hal-03656136](https://hal.archives-ouvertes.fr/hal-03656136)

¹Members of the APPRIMAGE Study Group are listed in Appendix C.

4.1 Introduction

Dementia is a world-wide disease that is becoming more and more important due to population ageing. T1-weighted (T1w) brain magnetic resonance imaging (MRI) contributes to the positive diagnosis of dementia by displaying typical spatial patterns of brain atrophy. As we have seen in the previous chapters, computer-aided diagnosis (CAD) systems using T1w brain MRI data have been arising in the last years thanks to the development of machine learning (ML) and deep learning (DL) models.

CAD systems have mainly been developed using research data sets thanks to their ease of access (many can be downloaded from web platforms) and their ease of use, as they are acquired following a research protocol whose aim is to guarantee data quality and homogenisation. Several data sets originating from research studies such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI)², the Open Access Series Of Imaging Studies (OASIS)³, the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL)⁴ or the Frontotemporal lobar degeneration neuroimaging initiative (NIFD)⁵, that we used in previous chapters, are publicly available and contain various clinical and imaging data, including T1w MRI brain data. We have seen that they have pushed the research on ML and DL for CAD using T1w brain MRI. Even if all these data sets have proven extremely useful to propel methodological research on ML/DL applied to neurological diseases, they are far from the everyday clinical routine for two main reasons. First, they use only research images where quality of the data is guaranteed, which cannot be the case in clinical practice. Second, many of them aim to differentiate patients with a particular, well-characterised, disease, from healthy controls. Such homogeneous diagnostic classes are difficult to obtain in a clinical context, as well as totally healthy subjects.

To bring research advances to clinical practice, some works have developed CAD systems using clinical data sets (Morin et al., 2020; Chagué et al., 2021). Nevertheless, they involve small data sets. Moreover, the data comes from highly specialised centres that are not representative of the overall clinical practice (for instance rare dementias and early-onset cases are over-represented). Finally, they often restrict themselves to the diagnosis of patients with dementia. It is thus unclear what their specificity is when dealing with MRI from patients with other diagnoses. Some works focused on differential diagnosis, which is closer to what is done in clinical routine, but they still use a research data set. Ma et al., 2020 classified patients with Alzheimer’s disease and fronto-temporal dementia using ADNI and NIFD, Koikkalainen et al., 2016 trained a model for the classification among patients with Alzheimer’s disease, fronto-temporal dementia, Lewy bodies disease and vascular dementia using the Amsterdam Dementia Cohort, a research data set.

In this context, images from clinical data warehouses may be used to train and evaluate ML and DL models for the CAD of dementia systems. Representing best the everyday clinic life of a hospital, they are an important tool for the translation of research to the clinic. Such images are heterogeneous (i.e. different sites, MRI sequences not harmonised) and

²<http://adni.loni.usc.edu/>

³<https://www.oasis-brains.org/>

⁴<https://aibl.csiro.au/>

⁵<https://ida.loni.usc.edu/home/projectPage.jsp?project=NIFD>

they include a very wide range of diagnoses (including not only patients with dementia but also patients with other neurological or psychiatric diseases, as well as patients who received a brain MRI for another indication).

The end goal of this work was to experimentally study the performance of ML methods to classify dementia patients in a clinical data warehouse using T1w brain MRI. However, as we will see in the following sections, important steps of quality control and image homogenisation have been necessary before reaching this stage.

4.2 AP-HP clinical data warehouse

This work relies on T1w brain MR images and clinical data from the clinical data warehouse (CDW), in French *Entrepôt de Données de Santé (EDS)*, of the AP-HP (Assistance Publique – Hôpitaux de Paris). This CDW gathers data from millions of patients across 39 hospitals of the Greater Paris area. The data were made available by the data warehouse of the AP-HP and the study was approved by the Ethical and Scientific Board of the AP-HP. According to French regulation, consent was waived as these images were acquired as part of the routine clinical care of the patients.

4.2.1 Data set description

All the images acquired by the AP-HP centres are stored in a single central clinical PACS. The data warehouse team made a query on the central clinical PACS and copied the images to the so-called ‘research PACS’. Note that, in spite of its name, the research PACS is also within the hospital network. The images were then pseudonymised: the DICOM fields that contained information about the patient or the physician who performed the exam, such as their name or identifier were erased. For further anonymisation, the date of the exam and the date of birth were also erased from the DICOM fields. Nevertheless, this information was available from another database (but not for all patients). In this other database, to increase anonymisation, the date of the exam and the date of birth were also changed (they were shifted by a constant to keep the age information accurate). Note that data were accessed remotely and that all the analyses (including training and inference of deep learning models on GPUs) were performed within the hospital network, as exporting data outside of this network is not allowed.

The images were selected according to DICOM attributes. A first query on the PACS was performed to list the DICOM attributes corresponding to MRI. For all the MR images, we listed the ‘series descriptions’, ‘body parts examined’, and ‘study descriptions’ DICOM attributes. A neuroradiologist manually selected all the attribute values that may refer to 3D T1w brain MRI (e.g. ‘T1 EG 3D MPR’, ‘SAG 3D BRAVO’, ‘3D T1 EG MPRAGE’, ‘IRM cranio’, ‘Brain T1W/FFEGADO’). He selected 3736 relevant attribute values. In case of a doubt, the neuroradiologist kept the value to avoid discarding potential images of interest. Relevant attribute values were manually selected since some of the information present in the DICOM fields is filled manually by the radiology department or even by the radiographer who is performing the exam. Standardisation exists within a given hospital but our data came from 39 different hospitals, which all have different conventions. Even within

a hospital, there was still a large variability, probably because different MRI protocols for a head/brain examination exist and there was no specific effort to name the body part in a consistent way across them. It could also be that these had spelling errors or that they were not changed during an exam (resulting in the annotation of gadolinium injection even when it is not present or the opposite).

Among all the 3D T1w brain MRI of the AP-HP, a first batch of about 11,000 images was delivered by the data warehouse. We excluded all the images having less than 40 slices because they correspond to 2D brain images even if the corresponding DICOM attribute refer to 3D. For the present study, we randomly selected 5500 images, corresponding to 4177 patients. The images were acquired on various scanners from four manufacturers: Siemens Healthineers ($n = 3752$), GE Healthcare ($n = 1710$), Philips ($n = 33$) and Toshiba ($n = 5$). Among all the images, 3229 images were acquired with 3 Tesla machines and 2271 with 1.5 Tesla. From the 5500 images, age and gender information was known only for 4274 images, corresponding to 3169 patients. This is explained by the fact that, while images are stored on the PACS, socio-demographic and clinical data are stored using another software system that had been installed later in the different hospitals. Furthermore, age and sex in the DICOM header were erased during the pseudonymisation process. Among the 4274 images, we have 2297 women, 1968 men and 9 patients with unknown sex, with an average age of 55.15 ± 7.89 (min: 18, max: 95). Table 4.1 reports all the scanner models present in our data set with the corresponding magnetic field strength for the 5500 images and the corresponding age range and sex for the images for which this information is available.

4.2.2 Image preprocessing

The T1w MR images were converted from DICOM to NIfTI using `dicom2niix` (Li et al., 2016) and organised using the Brain Imaging Data Structure (BIDS) standard (Gorgolewski et al., 2016). Images with a voxel dimension smaller than 0.9 mm were resampled using a 3rd-order spline interpolation to obtain 1 mm isotropic voxels. To facilitate annotations, we applied the following pre-processing using the `t1-linear` pipeline of Clinica (Routier et al., 2021), which is a wrapper of the ANTs software (Avants et al., 2014). Bias field correction was applied using the N4ITK method (Tustison et al., 2010). An affine registration to MNI space was performed (Avants et al., 2008). The registered images were further rescaled based on the min and max intensity values ($y = (x - \min(x))/(\max(x) - \min(x))$, where x is the T1w brain MRI in the MNI space). Images were then cropped to remove background resulting in images of size $169 \times 208 \times 179$, with 1 mm isotropic voxels (Wen et al., 2020). One should note that we only aimed to obtain a rough alignment and intensity rescaling to facilitate annotation (see next section).

TABLE 4.1: Model name of all the scanners, grouped by manufacturer, with the corresponding magnetic field strength (T) and the number of images. Age (mean \pm std[range]) and sex (number of females [F] / males [M]) are reported when available for each model. As indicated in the text, from the 5500 images, age and gender information were available only for 4274 images. Thus, this information was left blank when it was available for none of the images of a given scanner model.

	Model Name	T	N images	Age (mean \pm std [range])	Sex (F/M)
Siemens	Aera	1.5	489	53.53 \pm 18.00 [18, 95]	223 / 142
	Amira	1.5	29	47.81 \pm 13.57 [19, 68]	6 / 10
	Avanto	1.5	603	52.79 \pm 15.39 [18, 88]	164 / 125
	Avanto_fit	1.5	81	56.06 \pm 16.64 [19, 88]	34 / 28
	Biograph mMR	3	12	-	-
	Espreo	1.5	1	-	-
	Magnetom Vida	3	3	-	-
	Magnetom Essenza	1.5	11	37.2 \pm 15.93 [22, 69]	1 / 9
	Sempre	1.5	3	45 \pm 0 [45]	1 / 0
	Skyra	3	1851	54.31 \pm 17.56 [18, 95]	708 / 692
	Spectra	3	23	55.13 \pm 18.87 [22, 66]	2 / 6
	Symphony	1.5	3	-	-
	Verio	3	643	55.65 \pm 17.75 [18, 92]	310 / 294
GE Healthcare	Discovery MR450	1.5	4	40.67 \pm 23.57 [24, 74]	1 / 2
	Discovery MR750(w)	3	675	55.52 \pm 17.49 [18, 93]	240 / 256
	Optima MR360	1.5	2	63 \pm 0 [63]	0 / 1
	Optima MR450w	1.5	284	59.80 \pm 18.0 [18, 95]	160 / 97
	Signa Architect	1.5	243	52.14 \pm 18.63 [19, 92]	128 / 99
	Signa Artist	1.5	4	88.0 \pm 1.41 [86, 89]	2 / 2
	Signa Excite	1.5	3	30.5 \pm 4.5 [26, 35]	2 / 0
	Signa Explorer	1.5	1	76 \pm 0 [76]	1 / 0
	Signa HDx(t)	1.5	489	61.53 \pm 18.34 [18, 94]	250 / 166
	Signa Pioneer	3	1	76 \pm 0 [76]	0 / 1
	Signa Voyager	1.5	1	-	-
Unknown	1.5	3	-	-	
Philips	Achieva	3	21	51.0 \pm 14.0 [27, 70]	5 / 2
	Ingenia	1.5	5	81.13 \pm 12.20 [64, 92]	1 / 2
	Intera	1.5	7	61 \pm 0 [61]	2 / 0
Toshiba	Titan	1.5	2	54.5 \pm 1.5 [53, 56]	2 / 0
	Vantage Elan	1.5	3	55.5 \pm 3.5 [52, 59]	1 / 1

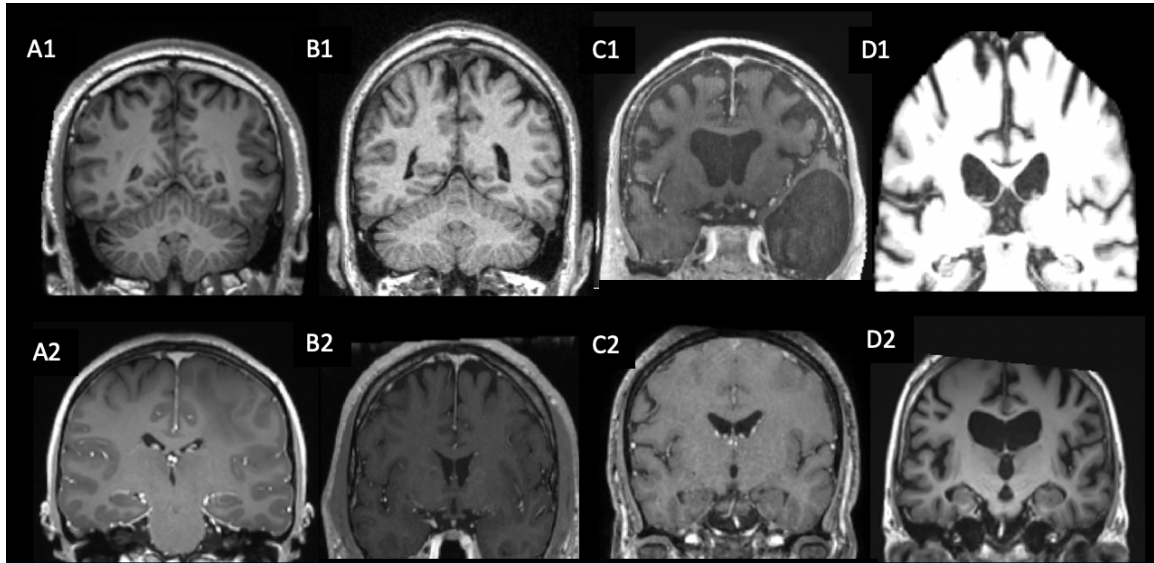


FIGURE 4.1: Examples of T1w brain images from the clinical data warehouse and the corresponding labels. A1: Image of good quality (tier 1), without gadolinium; A2: Good quality (tier 1), with gadolinium; B1: Medium quality (tier 2), without gadolinium (noise grade 1); B2: Medium quality (tier 2), with gadolinium (contrast grade 1); C1: Bad quality (tier 3), without gadolinium (contrast grade 2, motion grade 2); C2: Bad quality (tier 3), with gadolinium (contrast grade 2, motion grade 1); D1: Straight rejection (segmented); D2: Straight rejection (cropped).

4.3 Quality control of T1-weighted brain MRI from a clinical data warehouse

The quality of images acquired in a clinical routine context can greatly vary since the acquisition protocols are not standardised, scanners may not be recent and patients may have moved during the acquisition, see examples in Figure 4.1. All these factors can prevent CAD algorithms from working properly (Reuter et al., 2015; Gilmore et al., 2019). Quality control (QC) is thus a fundamental step before training and evaluating ML approaches on clinical routine data.

Manual QC takes time and is thus not always doable, especially in the context of ML-based CAD, where a large number of training samples is needed. Even if web-based systems facilitate annotation (Kim et al., 2019; Keshavan et al., 2018), the task remains unfeasible for very large data sets. In this context, automatic QC is thus needed. Many existing works extract image quality metrics, which requires an extensive preprocessing (Alfaro-Almagro et al., 2018; Esteban et al., 2017) that may fail on images of bad quality or train classifiers on images acquired following a well-defined research protocol (Sujit et al., 2019), which probably will not generalise to clinical data.

The objective of our work was to develop a method for the automatic QC of T1w brain MRI in large clinical data warehouses. The specific objectives were to: 1) discard images which are not proper T1w brain MRI; 2) identify images with gadolinium; 3) recognise images of bad, medium and good quality. We used 5000 images for training/validation and 500 for testing. To train/validate the models, the data were annotated by two trained raters. To that purpose, we introduced an original visual QC protocol that is applicable to clinical

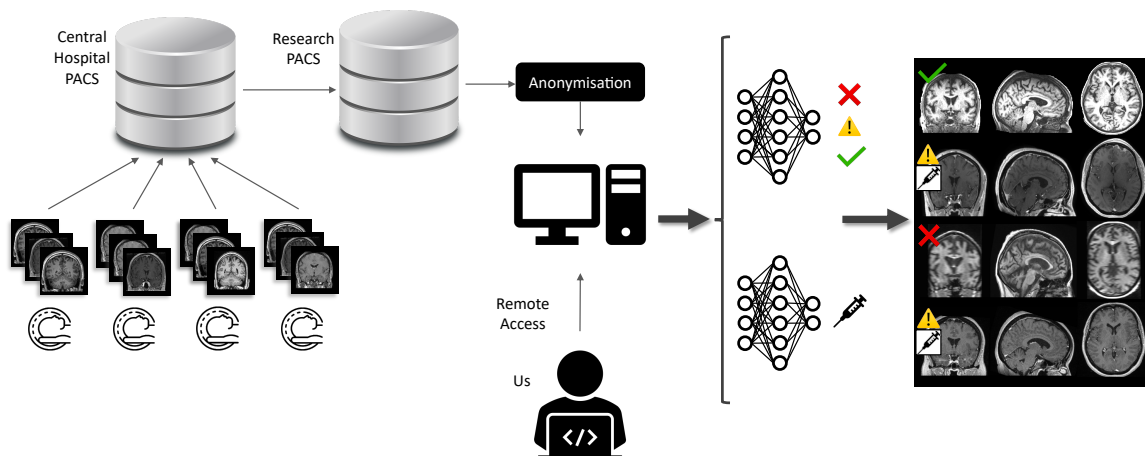


FIGURE 4.2: General workflow of the proposed QC framework. Images were acquired as part of the routine clinical care in different hospital sites and gathered in a central hospital PACS. Images relevant to our research project were copied to the research PACS and anonymised. They always remain within the hospital network that we accessed remotely. Thanks to the connection to the hospital IT network, we manually labelled the images before training and testing our deep learning models.

data warehouses. Figure 4.2 presents an overview of the proposed QC framework.

4.3.1 Manual labelling of the data set

Images were labelled by two trained raters and the annotation protocol was designed with the help of a radiologist.

4.3.1.1 Quality criteria

Five characteristics were manually annotated. The first two (straight rejection and gadolinium) are binary flags, while the other three (motion, contrast and noise) are assessed with a three-level grade.

- **Straight rejection (SR):** images not containing a T1w MRI of the whole brain (for instance images of segmented tissues or truncated images). Note that these images still have DICOM attributes corresponding to T1w brain MRI and thus were not removed through the selection step based on DICOM attributes.
- **Gadolinium:** presence of gadolinium-based contrast agent.
- **Motion 0:** no motion, 1: some motion but the structures of the brain are still distinguishable, 2: severe motion, the cortical and subcortical structures are difficult to distinguish.
- **Contrast 0:** good contrast, 1: medium contrast (grey matter and white matter are difficult to distinguish in some parts of the image), 2: bad contrast (grey matter and white matter are difficult to distinguish everywhere in the brain).
- **Noise 0:** no noise, 1: presence of noise that does not prevent identifying structures, 2: severe noise that does prevent identifying structures.

Gadolinium injection, motion, contrast and noise were noted for all the images which were not defined as SR. According to the grades given to the motion, contrast and noise characteristics, we determined three tiers corresponding to images of good, medium and bad quality. The tiers, along with the rules used to defined them, are described in Table 4.2.

TABLE 4.2: Description and determination rules of the proposed quality control tiers.

Tier	Description	Determination rule
Tier 1	3D T1w brain MRI of good quality	Grade 0 for motion, contrast and noise
Tier 2	3D T1w brain MRI of medium quality	At least one characteristic among motion, contrast and noise with grade 1 and none with grade 2
Tier 3	3D T1w brain MRI of bad quality	At least one characteristic among motion, contrast and noise with grade 2

4.3.1.2 Annotation set-up

Our aim was to annotate the largest possible number of images in an efficient manner while being restricted to the environment of the data warehouse which only included a Jupyter notebook and a command-line interface. We thus implemented a graphical interface in a Jupyter notebook. This interface displayed only the central axial, sagittal and coronal slices of the brain. Indeed, loading the whole 3D volume for inspecting all the slices in the data warehouse environment was unfeasible due to the above mentioned restrictions. Specifically, from the NIFTI format, we saved a screenshot of the central slice of each view (sagittal, coronal, axial) in PNG format. This allowed a fast loading of the image to annotate. Each image was labelled by two trained raters. The interface was flexible: it was possible to go back and label again an image, and after the labelling all the characteristics noted were displayed. The procedure was optimised to reduce the workload of the raters to a minimum. The implementation is available on a GitHub repository: https://github.com/SimonaBottani/Quality_Control_Interface.

4.3.1.3 Consensus label

The final label used to train and validate the automatic QC is a consensus between the two raters. If the users labelled different image characteristics, we determined a procedure to define a consensus label. We distinguished two types of disagreement: one regarding the SR status and the other one regarding the other characteristics based on which the tiers are assigned. When the two raters disagreed on the SR status, we manually set the consensus label: the two raters reviewed the images and decided together to keep the SR label or assign the alternative label. In case of disagreement regarding the other characteristics, the consensus was chosen as follows. The objective was to be as conservative as possible: we wanted to retain all the imperfections that may have been seen by one annotator and not by the other. For a given characteristic, the consensus grade was chosen as the maximum of the two grades of the observers. The tier was recomputed accordingly.

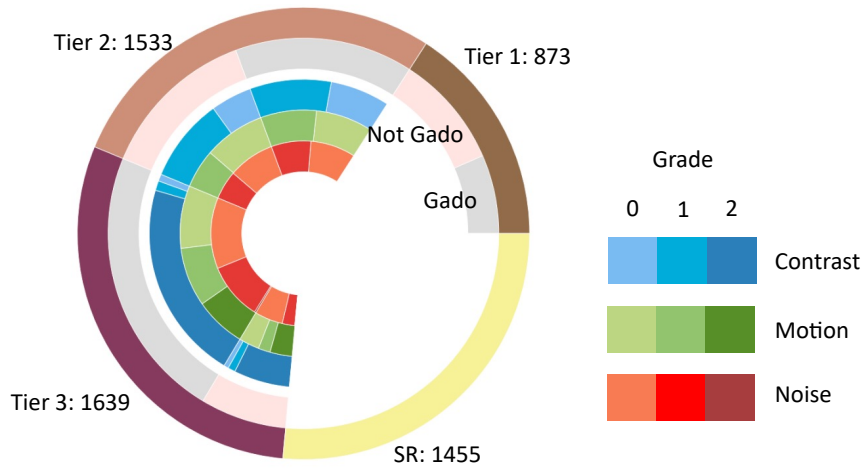


FIGURE 4.3: Distribution of the consensus labels for the whole data set of 5500 images. Outermost circle: images in SR and in the different tiers. For every tier, we divide between images with and without gadolinium injection. For each injection status we see the grade distribution of the contrast, motion and noise characteristics.

4.3.1.4 Results of the manual quality control

The inter-rater agreement was evaluated using the weighted Cohen’s kappa (Watson et al., 2010) between the two annotators for each of the characteristics. Results are presented in Table 4.3. The agreement is strong for the SR label and the gadolinium injection (0.88 and 0.89) and moderate for the other characteristics (from 0.68 to 0.79).

TABLE 4.3: Weighted Cohen’s kappa between the two annotators

Characteristics	Weighted Cohen’s kappa
SR (yes vs no)	0.88
Gadolinium injection (yes vs no)	0.89
Contrast (0 vs 1 vs 2)	0.79
Motion (0 vs 1 vs 2)	0.68
Noise (0 vs 1 vs 2)	0.70

The distribution of the consensus labels for the 5500 patients is shown in Figure 4.3. 26% of the images are labelled as SR, 16% as tier 1, 28% as tier 2, and 30% as tier 3. Figure 4.1 shows some representative examples of T1w brain images with the corresponding labels.

As expected, the proportion of images with gadolinium increased when the quality decreased (proportion of images with gadolinium: 41% in Tier 1, 53% in tier 2, 76% in tier 3; $p < 2.13e^{-8}$; χ^2 test). A vast majority of tier 3 images had a contrast of 2 (90%) and were with gadolinium (70%).

If we analyse the relationships between characteristics, we note that 73% of images with a grade 2 for motion have also a grade 2 for contrast. Unsurprisingly, a strong motion has

TABLE 4.4: Distribution of the manufacturers, field strength, sex and age according to QC grading (performed by the human raters) and on the overall population. We report the percentage of each manufacturer, field strength and sex, and the mean \pm standard deviation with the range for age. The analysis was restricted to the sub-population for which demographic information was available (4274 of 5500 images). Results with ** mean that the distributions between the overall population and a specific QC class were statistically significantly different (corrected $p < 0.05$).

	Manufacturer % Siemens, GE, Philips, Toshiba	Field strength % 1.5 T, 3 T	Age mean \pm std [range]	Sex % F, M
Tier 1 (n=702)	90%, 10%, 0%, 0%**	9%, 91% **	47.51 \pm 16.27 [18 - 88]	52%, 48%
Tier 2 (n=117)	78%, 22%, 0.2%, 0.01% **	44%, 56%	54.42 \pm 17.79 [18 - 95]	59%, 41%
Tier 3 (n=1323)	38%, 62%, 0%, 0.2%**	60%, 40% **	59.97 \pm 17.13 [18 - 85]	57%, 43%
SR (n=1132)	67%, 32%, 1%, 0%	28%, 72% **	54.95 \pm 18.01 [18 - 93]	47%, 53%
Total (n=4274)	65%, 35%, 0.2%, 0%	39%, 61%	55.15 \pm 17.89 [18 - 95]	53%, 46%

a severe impact on contrast. On the other hand, images with a grade 2 for contrast present a closer distribution of grade 0, 1 and 2 for motion (40%, 34%, and 26%, respectively).

We studied the influence of the age, sex, manufacturer and field strength for the SR images or the different tiers for which demographic information was available (4274 out of 5500). In Table 4.4, we report the percentage of each manufacturer, field strength and sex, and the mean, standard deviation and range for the age according to the QC grading performed by the human raters (SR, tier 1, tier 2 or tier 3). We compared the distribution of the four overall quality classes to the overall population using a χ^2 test for the manufacturer, field strength and sex, and with a t-test for the age. P-values were corrected for multiple comparisons using Bonferroni correction. We found statistically significant differences (corrected p-value < 0.05) for the manufacturer for tier 1, tier 2 and tier 3 and for the field strength for tier 1, tier 3 and SR. Specifically, in tier 1 and tier 2, there was a majority of Siemens machines (especially of 3T for tier 1), while in tier 3 there was a majority of GE Healthcare machines. In addition, the SR category contained many 3T images that are actually segmented images, as such processed images are usually available with the most recent machines (that come equipped with segmentation software). For age and sex, there was no significant difference.

DICOM attributes often contain information regarding the injection of gadolinium. However, it is well-known to radiologists that such information is often unreliable because it is manually entered by the MRI radiographer. We aimed to assess the extent to which such information was unreliable. We thus analysed the ‘study description’ and ‘series description’ DICOM attributes of the images to check if the presence of gadolinium injection was noted. We considered that it was noted if at least one of the words ‘gado’, ‘inj’ or ‘iv’ was present in the value of one of the attributes. Among the 2416 images that were manually annotated

as with gadolinium, 2033 images had the information in the DICOM attributes. Among the 1629 images that were manually annotated as without gadolinium, 987 were noted as images with gadolinium injection according to the DICOM attributes. Since our manual annotation of gadolinium injection is highly reproducible and was designed with the guidance of an experienced neuroradiologist, we conclude that, as expected, DICOM attributes do not provide reliable information regarding the presence of gadolinium. This highlights the importance of being able to detect it using an automatic QC tool.

4.3.2 Automatic quality control

We developed an automatic QC method based on convolutional neural networks (CNNs) trained to perform several classification tasks: 1) discard images which were not proper T1w brain MRI (SR: yes vs no); 2) identify images with gadolinium (gadolinium: yes vs no); 3) differentiate images of bad quality from images of medium and good quality (tier 3 vs tiers 2-1); 4) differentiate images of medium quality from images of good quality (tier 2 vs tier 1).

4.3.2.1 Network architecture

The network proposed was composed of five convolutional blocks and of three fully connected layers. The convolutional blocks were made of one convolutional layer, one batch normalisation layer, one ReLU and one max pooling. In the following, we refer to this architecture as Conv5_FC3. The models were trained using the cross entropy loss, which was weighted according to the proportion of images per class for each task. We used the Adam optimizer with a learning rate of 1e-4. We implemented early stopping and all the models were evaluated with a maximum of 50 epochs. The batch size was set to 2. The model with the lowest loss was saved as final model. Implementation was done using Pytorch. This architecture has previously been used and validated in the AD-DL framework (Wen et al., 2020) and is available in ClinicaDL (Thibeau-Sutre et al., 2022b).

We compared this network to more sophisticated CNN architectures. In particular, we implemented a modified 3D version of Google’s incarnation of the Inception architecture (Szegedy et al., 2016). In addition we also implemented a 3D ResNet (CNN with residual blocks) inspired from (Jónsson et al., 2019). Both the Inception and the ResNet models were trained using the cross entropy loss weighted according to the proportion of images per class, the Adam optimizer with a learning rate of 1e-4 and the batch size was set to 2. These two models have been used in (Couvry-Duchesne et al., 2020) to predict brain age from 3D T1w MRI. For that specific task, they achieved a higher performance than the 5-layer CNN mentioned above. Their implementation is openly available on GitHub <https://github.com/aramis-lab/pac2019> and all the parameters of the CNNs are listed in the supplementary materials of (Couvry-Duchesne et al., 2020).

4.3.2.2 Experiments

Before starting the experiments, we defined a test set by randomly selecting 500 images which respected the same distribution of tiers as the images in the training/validation set.

TABLE 4.5: Results of the CNN classifier for all the tasks. We report the BA of the annotators and for every metric of the CNN we report the mean and the empirical standard deviation across the five folds. BA: balanced accuracy; MCC: Matthews correlation coefficient; PPV: positive predictive values; NPV: negative predictive values

Metric	SR (yes vs no)	Gadolinium injection (yes vs no)	Tier 3 vs tiers 2-1	Tier 2 vs tier 1
BA annotators	97.13	96.10	91.56	88.27
BA classifiers	93.76 \pm 0.57	97.14 \pm 0.34	83.51 \pm 0.93	71.65 \pm 2.15
F1 score	94.85 \pm 0.41	97.04 \pm 0.31	84.07 \pm 1.02	74.10 \pm 1.35
MCC	85.71 \pm 1.11	94.00 \pm 0.64	67.38 \pm 2.13	42.10 \pm 3.25
Sensitivity	91.83 \pm 1.18	96.45 \pm 0.34	79.88 \pm 3.06	77.39 \pm 4.29
Specificity	95.69 \pm 0.53	97.82 \pm 0.62	87.14 \pm 3.14	65.92 \pm 7.47
PPV	86.44 \pm 1.43	98.33 \pm 0.46	81.93 \pm 3.36	83.20 \pm 2.31
NPV	97.51 \pm 0.35	95.39 \pm 0.42	85.83 \pm 1.49	57.78 \pm 2.63

We also verified that the distribution of the manufacturers and the different scanner models was respected. The remaining 5000 images were split into training and validation using a 5-fold cross validation (CV). The separation between training, validation and test sets was made at the patient level to avoid data leakage. For each of the four tasks considered (SR, gadolinium, tier 3 vs 2-1, tier 2 vs 1), the five models trained in the CV were evaluated on the test set. We also studied the influence of the size of the training set on the performance by computing learning curves. We compared the output of each classifier with the consensus label. To set the automatic QC results in perspective, we computed the balanced accuracy (BA) for the raters (defined as the average of the BAs between each rater and the consensus).

4.3.2.3 Results of the automatic quality control

Results obtained for the four tasks of interest by the proposed Conv5_FC3 classifier are presented in Table 4.5. We report the BA of the annotators for comparison. For the recognition of SR images, we used all the images available in the training/validation set ($n = 5000$); for the gadolinium and tier 3 vs tiers 2-1 tasks, the training/validation set does not include SR images ($n = 3770$); and for the tier 2 vs tier 1 task, the training/validation set does not include SR and tier 3 images ($n = 2182$).

Balanced accuracy for SR and gadolinium is excellent (94% and 97%). For SR, the CNN is slightly less good than the annotators. For gadolinium, the CNN is as good as the raters. For tier 3 vs 2-1, the classifier BA is good but lower than that of the annotators. For tier 2 vs 1, CNN BA is low (71%) and much lower than that of the raters (88%).

The influence of the size of the training set on the performance is shown in Figure 4.4. For SR, the performance increases with sample size, even if it is also good with few examples (90% for 500 images) because of the easiness of the task. For gadolinium, performance is very high regardless of the sample size. For tier 3 vs tiers 2-1, adding more training samples helps the classifier while this is not the case for tier 2 vs 1.

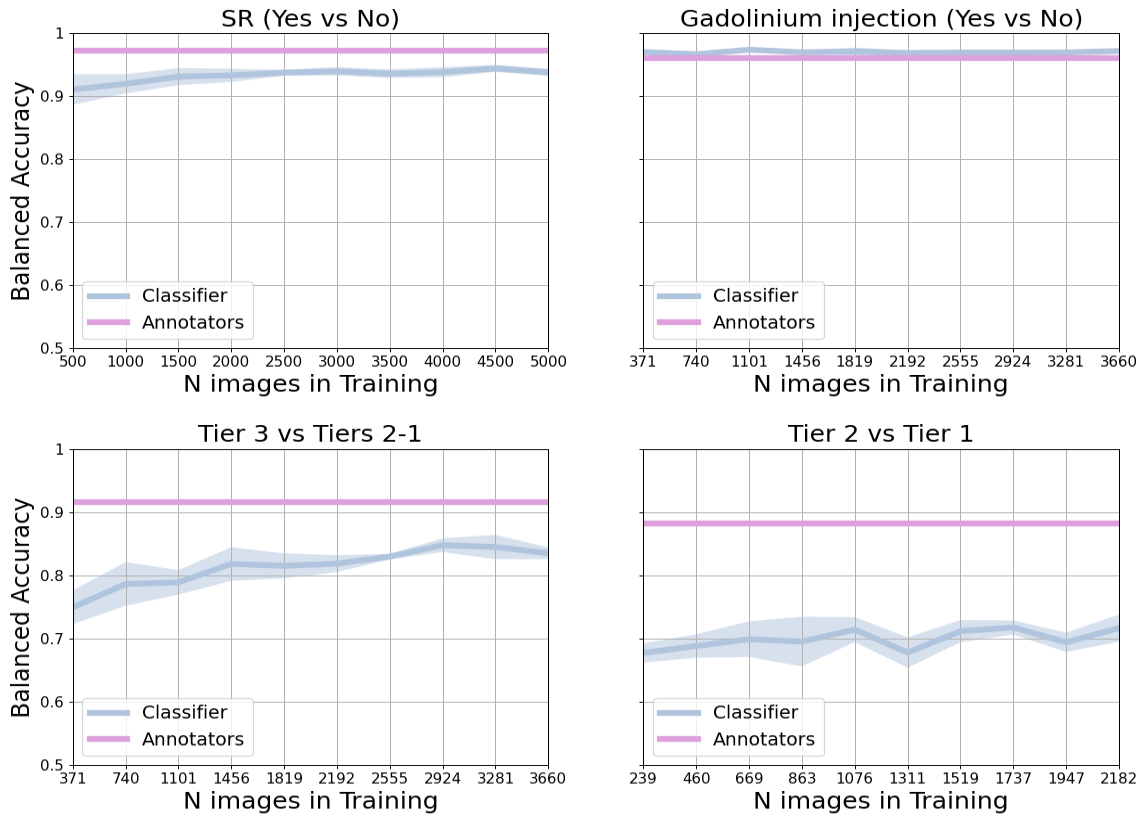


FIGURE 4.4: Learning curves for the SR (yes vs no), gadolinium injection (yes vs no), tier 3 vs tier 2-1 and tier 2 vs tier 1 tasks. Blue: balanced accuracy of the classifier across the five folds. Violet: balanced accuracy of the annotators on the testing set.

For tier 3 vs tiers 2-1 and tier 2 vs tier 1, we compared the proposed architecture, Conv5_FC3, with the Inception and ResNet architectures. For both tasks, the balanced accuracy obtained with the different networks is comparable: while for tier 3 vs tiers 2-1 it is slightly higher with the ResNet (85.82 ± 0.95) than the Conv5_FC3 (83.51 ± 0.93) and the Inception (82.40 ± 1.2), for tier 2 vs 1 it is slightly higher with the Conv5_FC3 (71.65 ± 2.15) than the ResNet (68.08 ± 1.6) or Inception (69.27 ± 2.05) architectures. For both tasks, the performance of the different classifiers were not statistically different (for tier 3 vs tiers 2-1: $p > 0.21$, McNemar's test; for tier 2 vs tier 1: $p > 0.12$, McNemar's test). All the metrics are reported in Table 4.6.

4.3.3 Conclusion

We developed a method for the automatic QC of T1w brain MRI for a large clinical data warehouse. Our approach allows discarding images which are of no interest (SR), recognising gadolinium injection and rating the overall image quality. To this aim, different CNN were trained and evaluated thanks to the manual annotation of 5500 images by two raters.

Manual annotation results showed that our protocol is reproducible across all tasks, even though agreement was less for more challenging characteristics. They also provide interesting information on the variability of image quality in a clinical routine data warehouse. As much as 25% are totally unusable (SR), and almost a third has a low quality (tier 3). We

TABLE 4.6: Results of three 3D CNN architectures (Conv5_FC3, Inception and ResNet) for the rating of the overall image quality. We report the mean and the empirical standard deviation across the five folds for all the metrics. BA: balanced accuracy; MCC: Matthews correlation coefficient; PPV: positive predictive values; NPV: negative predictive values

A. Tier 3 vs tiers 2-1

Metric	Conv5_FC3	Inception	ResNet
BA	83.51 \pm 0.93	82.41 \pm 1.28	85.82 \pm 0.95
Sensitivity	79.88 \pm 3.06	75.53 \pm 2.68	80.75 \pm 3.24
Specificity	87.14 \pm 3.14	89.29 \pm 3.45	90.89 \pm 2.22
F1 score	84.07 \pm 1.02	83.38 \pm 1.44	86.57 \pm 0.81
MCC	67.38 \pm 2.13	66.08 \pm 3.02	72.52 \pm 1.70
PPV	81.93 \pm 3.36	83.80 \pm 3.93	86.58 \pm 2.43
NPV	85.83 \pm 1.49	83.58 \pm 1.20	86.85 \pm 1.76

B. Tier 2 vs tier 1

Metric	Conv5_FC3	Inception	ResNet
BA	71.65 \pm 2.15	69.28 \pm 2.81	68.08 \pm 1.63
Sensitivity	77.39 \pm 4.29	76.86 \pm 4.76	82.35 \pm 2.90
Specificity	65.92 \pm 7.47	61.69 \pm 10.01	53.80 \pm 4.99
F1 score	74.10 \pm 1.35	72.28 \pm 1.13	72.94 \pm 1.18
MCC	42.10 \pm 3.25	37.74 \pm 4.10	37.13 \pm 2.73
PPV	83.20 \pm 2.32	81.51 \pm 3.08	79.40 \pm 1.34
NPV	57.78 \pm 2.63	55.49 \pm 1.70	58.77 \pm 2.40

also confirmed that gadolinium has a strong impact on image quality, hence the critical importance of detecting it accurately, the DICOM fields being unreliable in that regard. Note however that a limitation of this work is that annotations rely on three 2D slices, which means that some artefacts may have been missed. Nonetheless, if not visible on three central slices in different orientations, these potentially missed artefacts should have a minimal impact on subsequent image analyses.

For detecting straight reject, our CNN had excellent performance. Even though the task is relatively easy, this is very important to automatically discard images in a very large scale study. This was also the case for gadolinium, an important characteristic that strongly impacts the behaviour of many image analysis methods. We thus believe that these tools can be reliably used on the rest of this large data warehouse and already have an important practical impact for researchers in deep learning for medical imaging.

For detecting low quality data (tier 3), the performance was good even though lower than that of manual raters. On the other hand, it was substantially lower for differentiating between high and medium quality images. Such tools still seem useful for analysing the failure modes of CAD systems or other ML approaches, as such correlative work is still

doable with an imperfect tool. More work is nevertheless needed to use them for a strict rating of MRI quality.

4.4 Contrast-enhanced to non-contrast-enhanced image translation to exploit a clinical data warehouse of T1-weighted brain MRI

To perform differential diagnosis using classification algorithms, homogeneous features must be extracted from the images, no matter the disease, otherwise a link could be established between MRI sequence and pathology, which would create bias. Software tools such as SPM (Penny et al., 2011; Ashburner, 2012), ANTs (Avants et al., 2014) or FSL (Jenkinson et al., 2012) have been widely used for feature extraction but they were largely validated using structural T1w MRI without gadolinium only, to the best of our knowledge, and their good performance on images with gadolinium is thus not guaranteed. A solution could then be to convert contrast-enhanced T1w (T1w-ce) into non-contrast-enhanced T1w (T1w-nce) brain MRI before using such tools.

Deep learning has been widely used in the image translation domain to enhance image quality, such as for denoising (Hashimoto et al., 2019; Benou et al., 2017; Jiang et al., 2018; Yang et al., 2020) or super-resolution (Chen et al., 2018; Du et al., 2020; Kim et al., 2018; Pham et al., 2017; Zeng et al., 2018), but also for image harmonisation (Dewey et al., 2019). Closer to our objective, 3D U-Net like models have been developed for the synthesis of images with gadolinium from images without gadolinium (Bône et al., 2021; Kleesiek et al., 2019; Sun et al., 2020a).

Our objective in this work was to evaluate how image translation models could be used to exploit data from a clinical data warehouse by converting T1w-ce into T1w-nce images. This homogenisation step should enable a consistent extraction of features that would later be used for computer-aided diagnosis in a clinical setting. We thus developed and compared different deep learning models that rely on typical architectures used in the medical image translation domain to convert T1w-ce into T1w-nce images. In particular, we implemented 3D U-Net like models with the addition of residual connections, attention modules or transformer layers. We also used these 3D U-Net like models in a conditional GAN setting. We trained and tested our models using 307 pairs of T1w-nce and T1w-ce images coming from the AP-HP clinical data warehouse. We first assessed synthesis accuracy by comparing real and synthetic T1w-nce images using standard metrics. We tested our models both on images of good or medium quality and on images of bad quality to ensure that deep learning models could generate accurate T1w-nce images no matter the quality of the input T1w-ce images. We then compared the volumes of grey matter, white matter and cerebrospinal fluid obtained by segmenting the real T1w-nce, real T1w-ce and synthetic T1w-nce images using SPM (Ashburner et al., 2005) to verify that features extracted from synthetic T1w-nce were reliable.

4.4.1 Data set description

The data set used in this particular work is composed of 307 pairs of 3D T1w-ce and T1w-nce images that were extracted from the batch of 9941 images made available by the AP-HP data warehouse. We first selected all the images of low, medium and good quality, excluding images that were not proper T1w brain MRI (Bottani et al., 2022b), resulting in 7397 images. This selection was based on manual quality control for 5500 images and on automatic quality control for the remaining 4441 images (Bottani et al., 2022b). In the same way, the presence or absence of gadolinium-based contrast agent was manually noted for 5500 images, while it was obtained through the application of a CNN classifier for the remaining 4441 images. We then considered only patients having both a T1w-ce and a T1w-nce image at the same session, with a T1w-nce image of medium or good quality. Finally, to limit heterogeneity in the training data set, we visually checked all the images and excluded 52 image pairs that were potential outliers because of extremely large lesions. Among the selected images, 256 image pairs were of medium and good quality, and 51 image pairs had a T1w-ce of low quality and a T1w-nce of good or medium quality. In total the data set comprises 614 images: 534 images were acquired at 3 T and 80 at 1.5 T, 556 images were acquired with a Siemens machine (with seven different models) and 58 with a GE Healthcare machine (with five different models).

4.4.2 Network architecture

To generate T1w-nce from T1w-ce images, both 3D U-Net like models and conditional GANs were developed and compared. The code used to implement all the architectures and perform the experiments is openly available (https://github.com/SimonaBottani/image_synthesis).

4.4.2.1 3D U-Net like structures

We implemented three models derived from the 3D U-Net (Ronneberger et al., 2015; Çiçek et al., 2016): a 3D U-Net with the addition of residual connections (called *Res-U-Net*), a 3D U-Net with the addition of attention mechanisms (called *Att-U-Net*), a 3D U-Net with both transformer and convolutional layers (called *Trans-U-net*). The U-Net structure allows preserving the details present in the original images thanks to the skip connections (Ronneberger et al., 2015) and has shown good performance for image-to-image translation (Han, 2017; Shiri et al., 2019; Gong et al., 2018; Ladefoged et al., 2019; Spuhler et al., 2019; Yang et al., 2019; Neppl et al., 2019; Wolterink et al., 2017). Here we detail the three architectures, which are also shown in Figure 4.5.

Res-U-Net The *Res-U-Net* we implemented is based on the architecture first proposed by Milletari et al., 2016 and later used by Bône et al., 2021. The five descending blocks are composed of 3D convolutional layers followed by an instance normalisation block and a LeakyReLU (negative slope coefficient $\alpha = 0.2$). The four ascending blocks are composed of transposed convolutional layers followed by a ReLU. The final layer is composed of an upsample module (factor of 2), a 3D convolutional block and a hyperbolic tangent module.

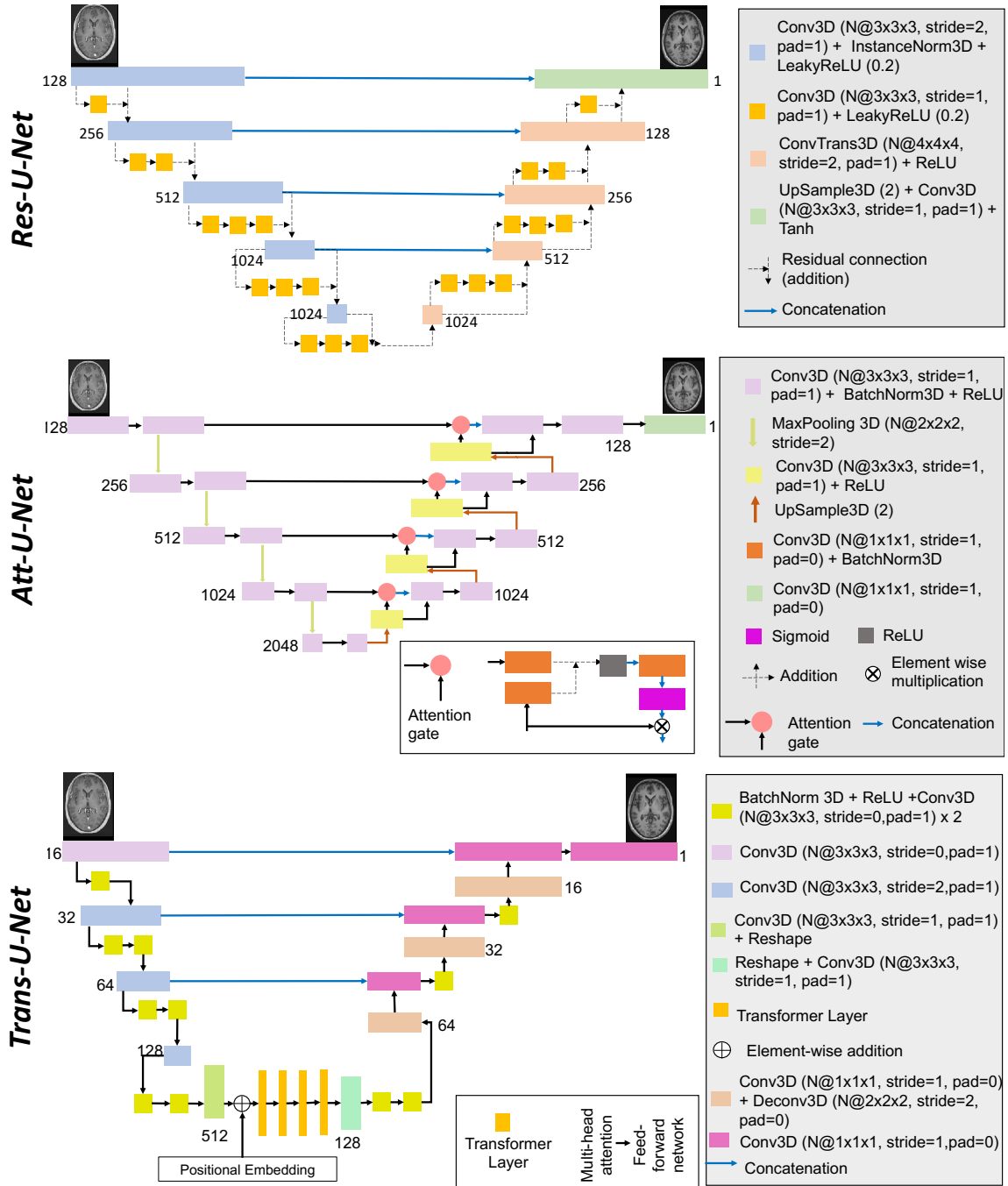


FIGURE 4.5: Architectures of the proposed 3D U-Net like models. The models take as input a real T1w-nce image of size $128 \times 128 \times 128$ and generate a synthetic T1w-nce of size $128 \times 128 \times 128$. *Res-U-Net*: images pass through five descending blocks, each one followed by a residual module, and then through four ascending blocks and one final layer. *Att-U-Net*: images pass through five descending blocks and then through four ascending blocks and one final layer. One of the inputs of each ascending block is the result of the attention gate. *Trans-U-Net*: images pass through four descending blocks, four transformer layers and four ascending layers. All the parameters such as kernel size, stride, padding, size of each feature map (N) are reported.

Each descending or ascending block is followed by a residual module, which can vary from one to three blocks composed of a 3D convolutional layer and a LeakyReLU ($\alpha = 0.2$). Residual blocks were introduced to avoid the problem of the vanishing gradients in the training of deep neural network (He et al., 2016b): they ease the training since they improve

the flow of the information within the network.

Att-U-Net We implemented the *Att-U-Net* relying on the work of Oktay et al., 2018. In this architecture, the five descending blocks are composed of two blocks with a 3D convolutional layer followed by a batch normalisation layer and a ReLU. They are followed by four ascending blocks. Each ascending block is composed of an upsample module (factor of 2), a 3D convolutional layer followed by a ReLU, an attention gate and two 3D convolutional layers followed by a ReLU. The attention gate is composed of two 3D convolutional layers, a ReLU, a convolutional layer and a sigmoid layer. Its objective is to identify only salient image regions: the input of the attention gate is multiplied (element-wise multiplication) by a factor (in the range 0–1) resulting from the training of all the blocks of the networks. In this way it discards parts of the images that are not relevant to the task at hand.

Trans-U-Net The *Trans-U-Net* was implemented by Wang et al., 2021 (who called the model *TransBTS*). They proposed a 3D U-Net like structure composed of both a CNN and a transformer. The CNN is used to produce an embedding of the input images not to lose local information across depth and space. The features extracted by the CNN are the input of the transformer whose aim is to model the global features. The descending blocks are composed of four different blocks, each being composed of a 3D convolutional layer and one, two or three blocks composed of a batch normalisation layer, a ReLU and another 3D convolutional layer. The model is then composed of four transformer layers, after a linear projection of the features. Each transformer layer is itself composed of a multi-head attention block and a feed forward network. The four ascending blocks are composed of a 3D convolutional layer and one or two blocks with a batch normalisation layer, a ReLU, a 3D convolutional layer followed by a 3D deconvolutional layer. The final layer is composed of a 3D convolutional layer and a soft-max layer.

For the three 3D U-Net like models we used the same training parameters. We used the Adam optimiser, the L_1 loss, a batch size of 2 and trained during 300 epochs. The model with the best loss, determined using the training set, was saved as final model. We relied on Pytorch for the implementation.

4.4.2.2 Conditional GANs

We propose three different conditional GANs (cGANs) models that differ in the architecture of the generators, which correspond to the three architectures presented above. The discriminator is the same for all the cGANs: it is a 3D patch CNN, first proposed by Isola et al., 2017 and used in the medical image translation domain (Wei et al., 2019; Choi et al., 2018). Its aim is to classify if each pair of patches contains two real images or a real and a fake image. The advantages of working with patches is that the discriminator focuses on the details of the images and the generator must improve them to fool the discriminator.

Our discriminator is composed of four blocks: the first three blocks are composed of a 3D convolutional layer followed by a LeakyReLU (negative slope coefficient $\alpha = 0.2$), and the last block is composed of a 3D convolutional layer and a 3D average pooling layer. From images of size $128 \times 128 \times 128$, we created eight patches of size $64 \times 64 \times 64$ with a stride of 50.

For the training of the discriminator we used the least-square-loss as proposed in Mao et al., 2017 to increase the stability, thus avoiding the problem of vanishing gradients that occurs with the usual cross-entropy loss. Stability of the training was also improved using soft labels: random numbers between 0 and 0.3 represented real images and random numbers between 0.7 and 1 represented fake images.

The total loss of the cGANs combines

- the loss of the generator composed of the sum of the L1 loss (i.e. pixel-wise absolute error) computed between the generated and true images, and the least-square loss computed between the predicted probabilities of the generated images and positive labels.
- the loss of the discriminator composed of the mean of the least-square loss computed between the predicted probabilities of the true images and positive labels and the least-square loss computed between the predicted probabilities of the generated images and negative labels.

At first, both the generators and discriminators were pre-trained separately. Regarding each generator, we reused the best model obtained previously. The discriminators were pre-trained for the recognition of real and fake patches (fake images were obtained from each pre-trained generator). The generators and discriminators were then trained together. The generator models with the best loss, determined using the training set, were saved as final models. Note that the batch size was set to 1 due to limited computing resources.

4.4.3 Experiments and validation measures

The experiments relied on 307 pairs of T1w-ce and T1w-nce images. We randomly selected 10% of the 256 image pairs of medium and good quality for testing (data set called $\text{Test}_{\text{good}}$), the other 230 image pairs being used for training. Only images of good and medium quality were used for training to ensure that the model focuses on the differences related to the presence or absence of gadolinium, and not to other factors. The remaining 51 image pairs with a T1w-ce of low quality and a T1w-nce of good or medium quality were used only for testing (data set called Test_{low}).

4.4.3.1 Synthesis accuracy

Image similarity was evaluated using the mean absolute error (MAE), peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) (Wang et al., 2004). The MAE is the mean of each absolute value of the difference between the true pixel and the generated pixel and PSNR is a function of the mean squared error: these two metrics allow a direct comparison between the synthetic image and the real one. The SSIM aims to measure quality by capturing the similarity of images, it is a weighted combination of the luminance, contrast and structure. For the MAE, the minimum value is 0 (the lower, the better), for PSNR the maximum value is infinite (the higher, the better) and for SSIM the maximum value is 1 (the higher, the better). We calculated these metrics both between the real and synthetic T1w-nce images and between the real T1w-nce and T1w-ce images (as reference). These metrics were calculated within the brain region. A brain mask was obtained for each subject

by skull-stripping the T1w-nce and T1w-ce images using HD-BET (Isensee et al., 2019) and computing the union of the two resulting brain masks.

4.4.3.2 Segmentation fidelity

Our goal is to obtain grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) segmentations from T1w-ce images using widely-used software tools that are consistent with segmentations obtained from T1w-nce images. We thus assessed segmentation consistency by analysing the tissue volumes resulting from the segmentations, which are important features when studying atrophy in the context of neurodegenerative diseases. We used the algorithm proposed in SPM (Ashburner et al., 2005) but these features can be obtained with commercial tools, such as NeuroreaderTM, volBrain, NeuroQuant or Inbrain, and used in a clinical setting (Morin et al., 2020; Zaki et al., 2022; Koikkalainen et al., 2016; Yu et al., 2021; Lee et al., 2021; Heckemann et al., 2008).

The volumes of the different tissues were obtained as follows. At first, synthetic T1w-nce images were resampled back to a size of $169 \times 208 \times 179$ using trilinear interpolation in Pytorch so that real and synthetic images have the same grid size. We processed the images using the `t1-volume-tissue-segmentation` pipeline of Clinica (Routier et al., 2021; Samper-González et al., 2018). This wrapper of the Unified Segmentation procedure implemented in SPM (Ashburner et al., 2005) simultaneously performs tissue segmentation, bias correction and spatial normalisation. Once the probability maps were obtained for each tissue, we computed the maximum probability to generate binary masks and we multiplied the number of voxels by the voxel dimension to obtain the volume of each tissue. We calculated both the relative absolute difference (rAD) and the relative difference (rD) for each tissue between the real T1w-ce or synthetic T1w-nce and the real T1w-nce as follows:

$$\text{rAD} = \frac{|V_t^I - V_t^J|}{TIV^I} \times TIV, \quad \text{rD} = \frac{V_t^I - V_t^J}{TIV^I} \times TIV,$$

where V_t^I is the volume of tissue t extracted from the real T1w-nce image I , V_t^J is the volume of tissue t extracted from image J , J being the synthetic T1w-nce or real T1w-ce image. TIV^I corresponds to the total intracranial volume (sum of the grey matter, white matter and cerebrospinal fluid volumes) obtained from the real T1w-nce image I and TIV corresponds to the average total intracranial volume computed across the two test sets. The multiplication by the average total intracranial volume (TIV) aims at obtaining volumes (in cm^3) rather than fractions of the TIV of each subject, which is easier to interpret. Since this is a multiplication by a constant, it has no impact on the results. To assess whether the tissue volumes presented a statistically significant difference in terms of rAD depending on the images they were obtained from, we performed paired t-tests using Bonferroni correction for multiple comparisons.

In addition, we compared the binary tissue maps extracted from the real T1w-ce or synthetic T1w-nce image to those extracted from the real T1w-nce using the Dice score.

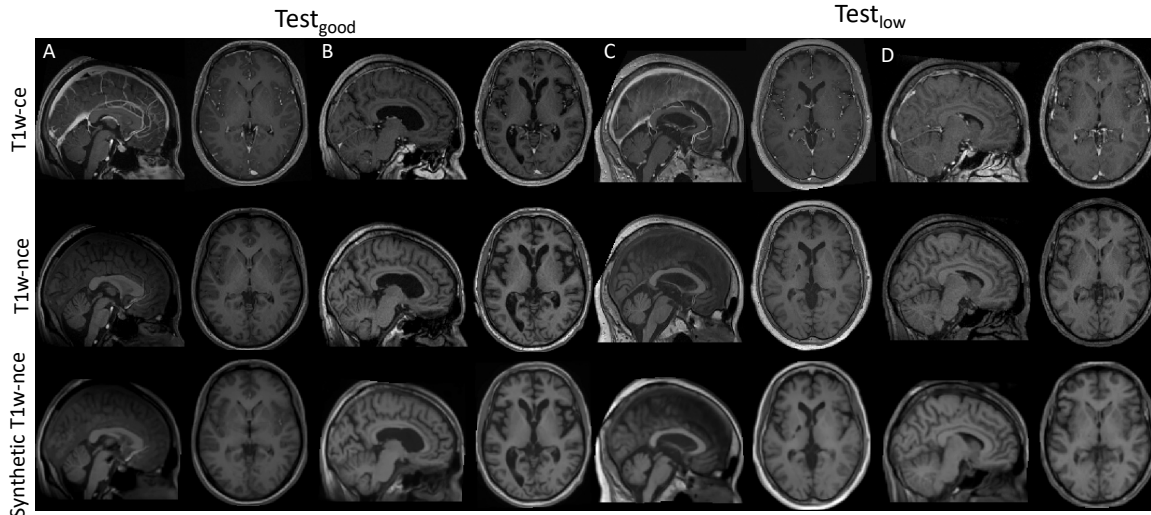


FIGURE 4.6: Examples of real T1w-ce (top), real T1w-nce (middle) and synthetic T1w-nce obtained with the *cGAN Att-U-Net* model (bottom) images in the sagittal and axial planes. Images of patients A and B belong to $\text{Test}_{\text{good}}$ (left) while images of patients C and D belong to Test_{low} (right).

4.4.4 Results

We report results for the proposed generator-only 3D U-Net like models and cGANs trained on 230 image pairs of good and medium quality, and tested on $\text{Test}_{\text{good}}$ and Test_{low} obtained from a clinical data set.

Examples of synthetic T1w-nce images obtained with the *cGAN Att-U-Net* model together with the real T1w-ce and T1w-nce images are displayed in Figure 4.6. Images of patients A and B belong to $\text{Test}_{\text{good}}$ while images of patients C and D belong to Test_{low} . We note the absence of contrast agent in the synthetic T1w-nce, while it is clearly visible in the sagittal slice of the T1w-ce (particularly visible for patients A and C) and that the anatomical structures are preserved between the synthetic and real T1w-nce, even in the case of a disease (as for patient B). We also note that contrast between grey and white matter is preserved in the synthetic T1w-nce (particularly visible for patients B and D). For Test_{low} , the contrast seems improved in the synthetic compared with the real T1w-ce image (especially for patient D). This results is not surprising as the networks were trained with images of medium or good quality, which will have on average a better contrast of than images of low quality.

4.4.4.1 Synthesis accuracy

Table 4.7 reports the image similarity metrics obtained for the two test sets within the brain region. We computed these metrics to assess the similarity between real and synthetic T1w-nce images, but also between T1w-nce and T1w-ce images to set a baseline. We observe that, for all models, the similarity is higher between real and synthetic T1w-nce images than between T1w-nce and T1w-ce images according to all three metrics on both test sets. The differences observed in terms of MAE, PSNR and SSIM between the baseline and each image translation approach are statistically significant (corrected p-value < 0.05 according to a paired t-test corrected for multiple comparisons using the Bonferroni correction).

TABLE 4.7: MAE, PSNR and SSIM obtained on the two independent test sets with various image quality. For each metric, we report the average and standard deviation across the corresponding test set. We compute the metrics for both T1w-ce and synthetic T1w-nce in relation to the real T1w-nce, and so within the brain region.

Test set	Compared images	Model	MAE (%)	PSNR (dB)	SSIM
Test _{good}	T1w-nce / T1w-ce	-	4.14 ± 1.59	23.03 ± 2.83	0.90 ± 0.05
	T1w-nce / Synthetic T1w-nce	<i>Res-U-Net</i>	3.06 ± 1.50	26.89 ± 4.30	0.95 ± 0.04
		<i>Att-U-Net</i>	2.73 ± 1.69	29.07 ± 4.53	0.96 ± 0.05
		<i>Trans-U-Net</i>	2.80 ± 1.42	28.00 ± 4.13	0.96 ± 0.04
		<i>cGAN Res-U-Net</i>	3.47 ± 1.59	23.89 ± 4.30	0.95 ± 0.04
		<i>cGAN Att-U-Net</i>	2.69 ± 1.68	28.89 ± 4.44	0.97 ± 0.05
		<i>cGAN Trans-U-Net</i>	2.86 ± 1.59	28.00 ± 4.32	0.96 ± 0.04
Test _{low}	T1w-nce / T1w-ce	-	3.71 ± 1.99	24.20 ± 3.85	0.91 ± 0.06
	T1w-nce / Synthetic T1w-nce	<i>Res-U-Net</i>	2.93 ± 1.77	26.71 ± 4.32	0.95 ± 0.05
		<i>Att-U-Net</i>	2.89 ± 1.85	27.15 ± 4.57	0.95 ± 0.05
		<i>Trans-U-Net</i>	2.98 ± 1.89	26.71 ± 4.38	0.94 ± 0.05
		<i>cGAN Res-U-Net</i>	3.20 ± 1.96	26.20 ± 4.42	0.93 ± 0.05
		<i>cGAN Att-U-Net</i>	2.86 ± 1.83	27.12 ± 4.50	0.95 ± 0.05
		<i>cGAN Trans-U-Net</i>	2.97 ± 1.83	26.68 ± 4.40	0.94 ± 0.05

Among the generator-only 3D U-Net like models, the *Att-U-Net* performed slightly better than the others, both for Test_{good} (mean MAE: 2.73%, PSNR: 29.07 dB, SSIM: 0.96) and Test_{low} (mean MAE: 2.89%, PSNR: 27.18 dB, SSIM: 0.95). The performance of the cGANs was comparable to their counterparts composed only of the generator. *cGAN Att-U-Net* had a lower MAE for both test sets (mean MAE: 2.69% for Test_{good} and mean MAE: 2.86% for Test_{low}). There was no statistically significant difference observed, no matter the synthesis accuracy measure, between *cGAN Att-U-Net*, the best performing model according to the MAE, and the other approaches for both test sets (corrected p-value > 0.05). For further validation we kept only the generator-only *Att-U-Net* and *cGAN Att-U-Net*.

4.4.4.2 Segmentation fidelity

Examples of probability grey matter maps obtained from T1w-ce, T1w-nce and synthetic T1w-nce images are displayed in Figure 4.7. Compared with the T1w-ce images, the grey matter maps obtained from the synthetic T1w-nce better resembles that extracted from the T1w-nce, especially for Test_{low}.

Absolute volume differences (rAD) obtained between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the generator-only *Att-U-Net* model and the *cGAN Att-U-Net*) for GM, WM and CSF are reported in Table 4.8. For both test sets and all tissues, the absolute volume differences are smaller between T1w-nce and synthetic T1w-nce images than between T1w-nce and T1w-ce images for the two models. Using the generator-only *Att-U-Net* on Test_{good}, absolute volume differences of GM and CSF between T1w-nce/T1w-ce and T1w-nce/Synthetic T1w-nce are statistically significantly different (corrected p-value < 0.01 according to a paired t-test corrected for multiple comparisons using the Bonferroni correction), while on Test_{low} absolute volume differences of all the tissues are statistically significantly different (corrected p-value < 0.01). Using the *cGAN Att-U-Net* model, absolute volume differences of all the tissues are statistically

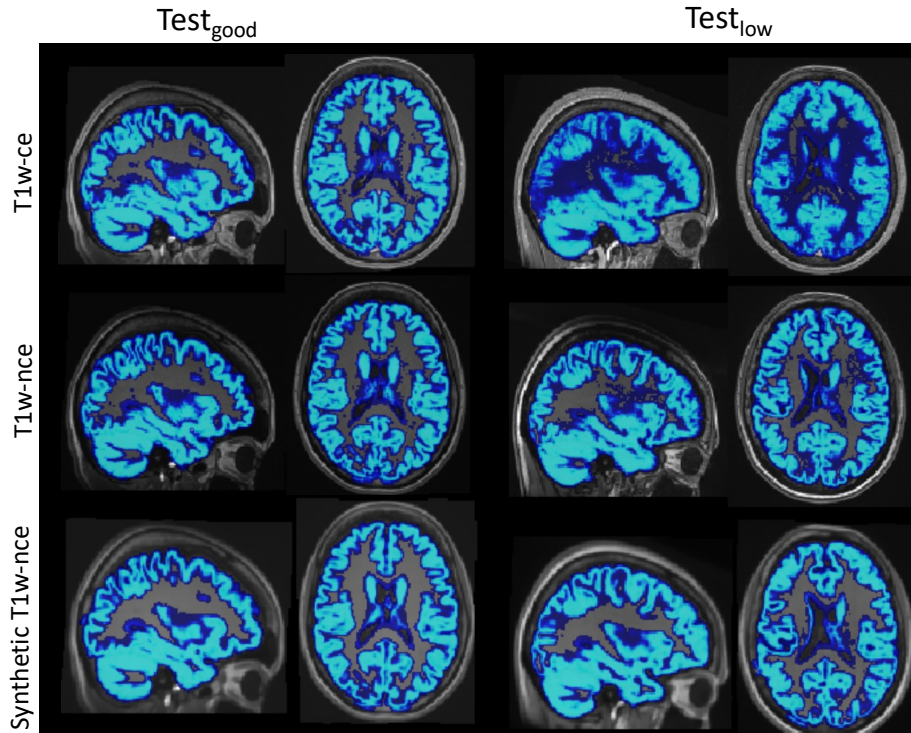


FIGURE 4.7: Example of the probability grey matter maps obtained from T1w-ce (top), T1w-nce (middle) and synthetic T1w-nce (*cGAN Att-U-Net* model) images from $\text{Test}_{\text{good}}$ (left) and Test_{low} (right).

TABLE 4.8: Absolute volume difference (mean \pm standard deviation in cm^3) between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the generator-only *Att-U-Net* and *cGAN Att-U-Net* models) for the grey matter, white matter and cerebrospinal fluid (CSF). * indicates that the absolute volume difference between T1w-nce and synthetic T1w-nce images is statistically significantly different from that of the baseline (corrected p-value < 0.01) according to a paired t-test corrected for multiple comparisons using the Bonferroni correction.

	Compared images	Model	$\text{Test}_{\text{good}}$ [cm^3]	Test_{low} [cm^3]
Grey matter	T1w-nce / T1w-ce	-	26.68 ± 15.92	49.63 ± 49.38
	T1w-nce / Synthetic T1w-nce	<i>Att-U-Net</i>	10.36 ± 6.98 *	19.61 ± 29.54 *
		<i>cGAN Att-U-Net</i>	9.24 ± 6.10 *	19.67 ± 28.32 *
White matter	T1w-nce / T1w-ce	-	10.81 ± 3.71	25.36 ± 27.73
	T1w-nce / Synthetic T1w-nce	<i>Att-U-Net</i>	7.79 ± 5.87	13.95 ± 24.74 *
		<i>cGAN Att-U-Net</i>	6.40 ± 4.43 *	14.49 ± 21.06 *
CSF	T1w-nce / T1w-ce	-	61.62 ± 34.61	69.55 ± 37.77
	T1w-nce / Synthetic T1w-nce	<i>Att-U-Net</i>	13.37 ± 10.18 *	12.25 ± 7.72 *
		<i>cGAN Att-U-Net</i>	18.27 ± 17.20 *	17.10 ± 18.45 *

significantly different (corrected p-value < 0.01) for both test sets. This means that there is an advantage in using synthetic T1w-nce images rather than T1w-ce images, no matter the model used for the synthesis: segmentation of GM, CSF and WM is more reliable since closer to the segmentation of the tissues in the real T1w-nce.

Volume differences (rD) computed between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the generator-only *Att-U-Net* and

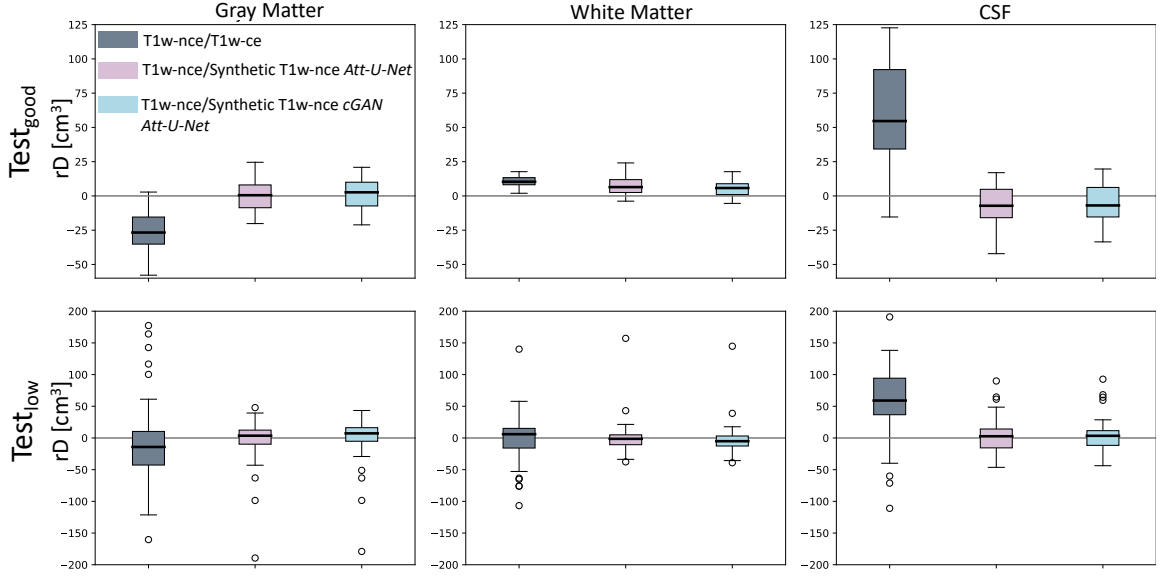


FIGURE 4.8: Volume differences (rD) in cm^3 between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the generator-only *Att-U-Net* and the *cGAN Att-U-Net* models) for grey matter (left), white matter (middle) and cerebrospinal fluid (CSF, right) for both $Test_{good}$ (top) and $Test_{low}$ (bottom).

cGAN Att-U-Net) for GM, WM and CSF are reported in Figure 4.8. We observe that volumes extracted from T1w-ce images tend to be over-estimated (GM) or under-estimated (CSF) and that most of these biases disappear when tissues are extracted from synthetic T1w-nce images (mean rD closer to 0).

TABLE 4.9: Dice scores obtained when comparing the grey matter, white matter and cerebrospinal fluid (CSF) segmentations between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the generator-only *Att-U-Net* and the *cGAN Att-U-Net*).

	Compared images	Model	$Test_{good}$	$Test_{low}$
grey matter	T1w-nce / T1w-ce	-	0.88 ± 0.02	0.77 ± 0.12
	T1w-nce / Synthetic T1w-ce	<i>Att-U-Net</i>	0.87 ± 0.02	0.81 ± 0.07
		<i>cGAN Att-U-Net</i>	0.87 ± 0.02	0.81 ± 0.07
White matter	T1w-nce / T1w-ce	-	0.93 ± 0.01	0.85 ± 0.10
	T1w-nce / Synthetic T1w-ce	<i>Att-U-Net</i>	0.90 ± 0.02	0.86 ± 0.04
		<i>cGAN Att-U-Net</i>	0.91 ± 0.02	0.86 ± 0.03
CSF	T1w-nce / T1w-ce	-	0.63 ± 0.10	0.62 ± 0.10
	T1w-nce / Synthetic T1w-ce	<i>Att-U-Net</i>	0.80 ± 0.05	0.78 ± 0.07
		<i>cGAN Att-U-Net</i>	0.80 ± 0.05	0.78 ± 0.07

The Dice scores obtained when comparing the GM, WM and CSF segmentations between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the generator-only *Att-U-Net* and the *cGAN Att-U-Net*) are displayed in Table 4.9. We observe that for both grey and white matter, the Dice scores are similar between T1w-nce

and T1w-ce or synthetic T1w-nce images, while for CSF higher Dice scores are obtained using synthetic T1w-nce images.

4.4.5 Conclusion

A source of heterogeneity among clinical data sets is the fact that they contain a mix of images acquired with and without gadolinium-based contrast agent. In our case, among the 7397 proper T1w brain images made available by the AP-HP data warehouse out of a batch of 9941 images, 59% of the images were contrast-enhanced (Bottani et al., 2022b). As a first step towards the homogenisation of this data set, we thus proposed a framework to convert T1w-ce images into T1w-nce images using deep learning models (generator-only U-Net models and conditional GANs).

We first assessed the similarity between real and synthetic T1w-nce images and between real T1w-nce and T1w-ce images using three similarity metrics, MAE, PSNR and SSIM. We showed that the similarity between real and synthetic T1w-nce images was higher than the similarity between real T1w-nce and T1w-ce images according to all the metrics, no matter the models used nor the quality of the input image. The synthesis accuracy obtained with the models evaluated was of the same order as the one reached in recent works on non-contrast-enhanced to contrast-enhanced image translation (Bône et al., 2021; Kleesiek et al., 2019). The performance of all the models was equivalent (no statistically significant difference observed), meaning that all were able to synthesise T1w-nce images. Slightly better performance was reached with the addition of attention modules (generator-only *Att-U-Net* and *cGAN Att-U-Net* models), and these models were thus further evaluated.

In the second step of the validation, we assessed the similarity of features extracted from the different images available using a widely adopted segmentation framework, SPM (Penny et al., 2011; Ashburner et al., 2005). We showed that the absolute volume differences of GM, WM and CSF were larger between real T1w-nce and T1w-ce images than between real and synthetic T1w-nce images (statistically significant difference most of the times). This confirms the hypothesis that gadolinium-based contrast agent may alter the contrast between the different brain tissues, making features extracted from such images with standard segmentation tools, here SPM (Penny et al., 2011; Ashburner et al., 2005), unreliable. At the same time, we validated the suitability of the synthetic images since their segmentation was consistent with those obtained from real T1w-nce images as the volume differences were small. In particular we see that for both test sets, volume differences are statistically significantly different (corrected p-value < 0.01 according to a paired t-test corrected for multiple comparisons using the Bonferroni correction) for GM which is the main feature when studying atrophy in neurodegenerative diseases. The fact that the relative differences between the volumes extracted from the real and synthetic T1w-nce images are relatively close to zero show that the tissue volumes are not systematically under- or over-estimated when extracted from the synthetic images.

Overall, these results demonstrate the ability of deep learning methods to help exploit data from the AP-HP clinical data warehouse.

4.5 Computer-aided diagnosis of dementia in a clinical data warehouse

As mentioned at the beginning of this chapter, the end goal of this work is to experimentally study the performance of ML methods to classify dementia patients in a CDW using T1w brain MRI. Patients with dementia were defined using ICD-10 codes assigned during the hospitalisation period. The ML model was a linear SVM using grey matter maps as features. It was then compared to several deep learning models. We compared the performance obtained on a research data set to that obtained on our clinical data set. We studied how results in a clinical data set may be biased by the characteristics of the training data set (in particular by the injection of gadolinium and the presence of images of different quality). To improve the classification, three different solutions were assessed: applying an image translation approach to change the appearance of images for which gadolinium was injected, using images of good quality or training the models using only research data.

4.5.1 Materials

To compare the performance of CAD systems to detect dementia in a research and a clinical setting, two data sets were used.

4.5.1.1 Research data set

Our research data set was composed of subjects from the ADNI database. We considered subjects from ADNI 1/2/Go/3 diagnosed as cognitive normal (CN) or Alzheimer's disease (AD) at baseline and only kept subjects whose diagnosis did not change over time. This resulted in 800 subjects with a T1w MR image at the first session including imaging data (CN: 410 subjects, 54.87 % F, age 73.20 ± 6.15 in range [55.1, 89.6]; AD: 390 subjects, 44.0 % F, age 74.88 ± 7.76 in range [55.1, 90.1]). Two hundred subjects (100 CN and 100 AD) composed the independent test set and the remaining subjects (310 CN and 290 AD) were used for the training/validation of the models using a 5-fold cross-validation (CV).

4.5.1.2 Clinical routine data set

The clinical data set comes from the AP-HP data warehouse.

Imaging and clinical data collection Images from this clinical data warehouse are very heterogeneous: they include 3D T1w brain MR images of patients with a wide range of ages (from 18 to more than 90 years old) and diseases, acquired with different scanners (more than 30 different models). Images relevant to our research project (i.e. 3D T1w brain MR images of patients aged more than 18 years old) were copied to the research PACS where they were pseudonymised. The selection process to obtain images of interest was described in Section 4.2.

At the same time, clinical data corresponding to the patients of our query are stored in a database managed by the ORBIS clinical information system. Clinical data gather all the information connected to the patients, i.e. date of birth, sex, diagnostic codes, medications,

biological tests, electronic health reports. As explained in (Daniel et al., 2020), ORBIS has been installed progressively in the AP-HP hospitals since 2009. Among all the patients aged more than 18 years old who undertook a 3D T1w brain MRI examination at AP-HP (~130,000 patients), only ~25% were registered in ORBIS. Among them, 23,688 patients were hospitalised. Note that for non-hospitalised patients, only sociodemographic data (sex and age) are available and not clinical data. As for the imaging data, the data warehouse provided the pseudonymised clinical data.

For our work, we were interested in two sociodemographic items (age and sex) as well as one clinical item (diagnostic codes). Codes from the 10th revision of international classification of diseases (ICD-10) (World Health Organization et al., 2007) were used to associate a diagnosis to each T1w brain MRI. Images were labelled according to the ICD-10 codes assigned to the visit corresponding to the acquisition of the image. We defined a visit as a period of plus or minus three months from the acquisition date of the image. As clinical data can be entered by the medical staff at different moments during hospitalisation, this time window ensures that all pieces of information regarding brain disorders related to the need of a brain MRI exam are collected.

In conclusion, the initial clinical data set of interest was composed of 23,688 patients, which corresponds to 32,348 visits and 43,418 3D T1w brain MR images.

Definition of the different diagnostic categories from ICD-10 codes On average, 60 ICD-10 codes were assigned to each visit. Since we did not know the reason of a patient’s hospitalisation (which may be different from the reason why they were prescribed an MRI examination), we considered principal diagnoses, secondary diagnoses and comorbidities at the same level. First, we identified all the ICD-10 codes that could refer to dementia (denoted as D). Note that we use the term ‘dementia’ in a broad sense, i.e. we consider mild cognitive impairment as belonging to this category. Then, we divided the remaining codes into two groups: ICD-10 codes referring to diseases (for instance cancer, demyelinating diseases, stroke, hydrocephalus) that lead to lesions altering T1w brain MRI (referred to as ‘no dementia but with lesions’ - NDL) and ICD-10 codes corresponding to diseases that, in principle, do not lead to lesions altering T1w brain MRI (referred to as ‘no dementia and no lesions’ - NDNL). We considered two different classification tasks in which dementia patients had to be differentiated from these two classes (NDL and NDNL), which have very different characteristics.

In Table 4.10, we list the three classes mentioned above (D, NDL, NDNL). For each of them, we provide a brief description and a list of all the associated ICD-10 codes. Sixteen diseases were associated to the category dementia. Four families of diseases were associated to the NDL category (which are defined by grouping different ICD-10 codes). The NDNL category corresponded to all the other codes. According to the standard structure of the ICD-10 codes, we considered just the first letter and the first two numbers, indicating the category, to identify the diseases belonging to the NDL category. The third number, indicating the aetiology, was used to identify the diseases corresponding to the dementia category as we wanted to be more specific.

TABLE 4.10: Description of the three categories of interest with the corresponding ICD-10 codes. Details about dementia codes: ‘/’ indicates that the two codes refer to the same diagnosis, ‘+’ means that the diagnosis of dementia is defined by the presence of both codes.

Category	ICD-10 codes
D: Dementia associated to a neurodegenerative disease or a vascular disease that causes atrophy visible on T1w MRI.	<ul style="list-style-type: none"> • Dementia in AD with early onset (F00.0/G30.0) • Dementia in AD with late onset (F00.1/G30.1) • Dementia in AD, atypical or mixed type (F00.2/G30.8) • Dementia in AD, unspecified (F00.9/G30.9) • Dementia in Pick disease (F02.0/G31.0) • Dementia in Creutzfeldt-Jakob disease (F02.1/A81.0) • Dementia in Huntington disease (F02.2 + G10) • Vascular dementia of acute onset (F01.0) • Multi-infarct dementia (F01.1) • Subcortical vascular dementia (F01.2) • Mixed cortical & subcortical vascular dementia (F01.3) • Other vascular dementia (F01.8) • Vascular dementia, unspecified (F01.9) • Mild cognitive disorder (F06.7) • Dementia in Parkinson’s disease (F02.3 + G20) • Lewy bodies dementia (G02.8 + G31.8)
NDL: No dementia but diagnosis that suggests presence of lesions that modify the anatomical structure of the brain visible on T1w MRI.	<ul style="list-style-type: none"> • Cancer (C70, C71, C72, D32, D33, D42) • Demyelination (G35, G36, G37) • Stroke (G45, G46) • Hydrocephalus (G91)
NDNL: No dementia and no diagnosis suggesting the presence of lesions on T1w brain MRI.	All the other codes

Selection of patients belonging to the dementia category Dementia is the principal category we consider since our aim is to study how well this category can be distinguished from the others. We thus started by selecting patients labelled as dementia. In the workflow displayed in Figure 4.9 we report the different choices made to create this population. For each step, we report the number of patients, visits and images.

Starting from 2441 patients with at least one ICD-10 code in the dementia category, corresponding to 2671 visits and 3633 images (considering only 3D T1w brain MRI), the final population is composed of 1255 patients, corresponding to 1255 visits and 1415 images. We first excluded patients that had multiple ICD-10 codes belonging to the dementia category at the same visit to have a unique label per visit. We then excluded patients with an ICD-10 code belonging to the NDL category with the aim that lesions visible on T1w brain MRI originate only from dementia. Patients were further excluded if the ICD-10 code in the dementia category was changing over time (i.e. over the different visits) as this may be due to an error in coding. Patients aged more than 90 years old were excluded because there were very few patients above this age across the different diagnostic groups (and thus it was not possible to find patients with the same age/sex). Patients labelled F06.7 (mild cognitive disorder) aged less than 45 years old were excluded because the diagnosis may correspond

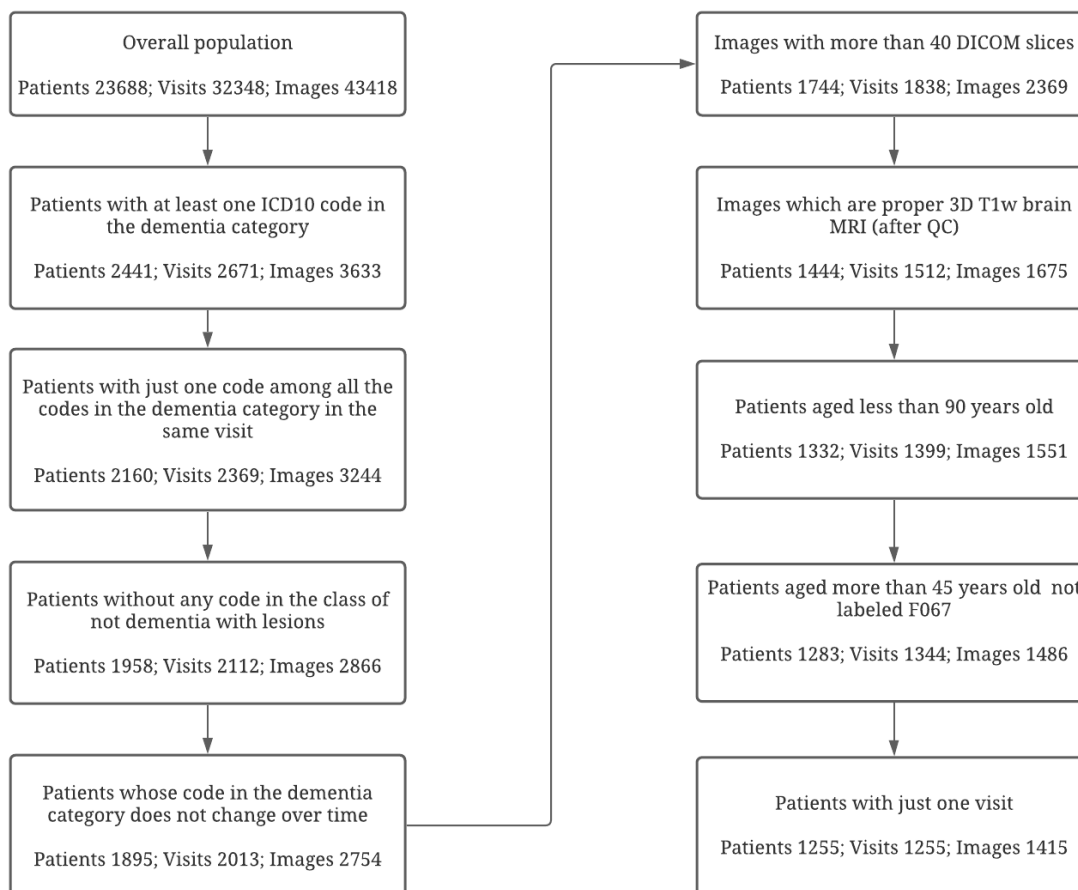


FIGURE 4.9: Workflow describing the selection of patients belonging to the dementia category. For each selection step, we report the corresponding number of patients, visits and images.

to a transient mild cognitive impairment and not to a prodromal stage of dementia. Some images were also excluded after the pre-processing step: if they had less than 40 DICOM slices or if they were labelled as straight reject by the quality control step.

Selection of patients belonging to the no dementia with lesions (NDL) and no dementia and no lesions (NDNL) categories The aim of this work is to assess whether patients with dementia can be distinguished from patients with other brain diseases, no matter if these diseases result in the presence (NDL category) or absence (NDNL category) of lesions visible on T1w brain MRI. To define the cohorts for the NDL and NDNL categories, we matched each patient belonging to the dementia category with a patient in the NDL category and with a patient in the NDNL category that had the same age and sex.

We first created the NDL cohort, which is composed of patients with one of these four diseases potentially leading to brain lesions visible on the T1w MRI: cancer, stroke, demyelination and hydrocephalus (see Table 4.10). We selected all the patients having at least one ICD-10 code in this category, resulting in 3843 patients corresponding to 6598 visits and 9615 images. We then matched these patients with those composing the dementia cohort following several criteria. For each patient with dementia:

- We selected all the patients with the same age and the same sex having at least one code in the NDL category.
- We excluded all the patients having different NDL codes at the same session.
- We considered only one visit for each patient when there were multiple visits available with the same diagnosis. The visit was selected randomly.
- Among all the patients with one visit matching these criteria, we randomly selected one of them.

We iterated this selection process twice since some images were discarded after the pre-processing steps (i.e. images with fewer than 40 DICOM slices or flagged as straight reject at the quality control step). In total we matched 808 patients (corresponding to 808 visits and 978 images).

The NDNL class is composed of all the patients having no code in the dementia nor NDL categories. For each patient with dementia:

- We selected all the patients with the same age and the same sex having no ICD-10 code in the dementia nor NDL categories.
- In case of multiple visits for a patient, we randomly selected one of them.
- Among all the patients with one visit matching these criteria, we randomly selected one of them.

We iterated this selection process twice since some images were discarded after the pre-processing steps. In total we matched 1144 patients (corresponding to 1144 visits and 1343 images).

Final cohorts The final cohorts were created by taking the intersection of the NDL patients matching with dementia patients and of the NDNL patients matching with dementia patients. This resulted in three cohorts each of 756 patients for a total number of 2268 patients (corresponding to 2268 visits and 2823 images). Note that this number of 756 patients is lower than the initial number of patients in the dementia class because some of them could not be matched for age and sex with a patient of the two other classes.

In Table 4.11 we report the number of subjects, visits and images for each category. In addition, we report the percentage of females and the average age of the patients as well as the percentage of images with and without injection of gadolinium, and of images of medium or good quality (tier 2-1). The presence of gadolinium and the quality of the images were determined through the automatic approach described in Section 4.3.

Training, validation and testing subsets Before starting the experiments, we defined a test set by randomly selecting 20% of the patients of the dementia class and the corresponding matched patients of the other two classes (NDL and NDNL). While for the training/validation set, if there were several images at the same visit all were kept to increase the number of training samples, for the test set, we selected only one image per visit (the selection was made randomly). This resulted in a test set composed of 152 patients/images for each of the three classes (D, NDL, NDNL). The training/validation set was composed of 604 patients and 719 images for D, 604 patients and 799 images for NDL, 604 patients and 756 images for NDNL.

TABLE 4.11: For each category, we report the number of patients and images, the age, the percentage of females, of images in Tier 2-1 (i.e. images of medium and good quality) and the percentage of images with gadolinium-based contrast agent. Results with ** mean that the distributions between the overall population and a specific category were statistically significantly different (χ^2 test corrected for multiple comparisons using the Bonferroni procedure, corrected p-value <0.05). Age and sex were computed at the patient level, while the tiers and the gadolinium injection were computed at the image level.

Category	N patients	N images	Age (mean \pm std [range])	Sex (%F)	%Tier 2-1	With gadolinium
D	756	887	71.17 \pm 11.58 [18,90]	50.34%	57.72%**	24.80%**
NDL	756	997	71.17 \pm 11.58 [18,90]	50.34%	52.25%	63.59%**
NDNL	756	939	71.17 \pm 11.58 [18,90]	50.34%	36.42%**	66.13%**
Total	2268	2823	71.17 \pm 11.58 [18,90]	50.34%	48.71%	52.24%

We respected the same distribution of image quality and presence of gadolinium between the test and the training/validation sets. We also checked that the distribution of the ICD-10 codes between the test and the training/validation sets among the dementia and NDL categories was the same.

For each task, the images of the training/validation set were further split using a 5-fold CV. The splits were the same for all the experiments and the distribution of image quality and presence of gadolinium respected the overall distribution.

Training subsets To study potential biases related to the presence of gadolinium or the quality of the images, we created different training subsets:

- $T_{\text{no gado}}^{172}$ includes only matched dementia, NDL and NDNL patients with images acquired without gadolinium injection. This results in a training subset of 172 patients per class.
- $T_{\text{tier } 1/2}^{181}$ includes only matched dementia, NDL and NDNL patients with images of medium or good quality (tier 2-1). This results in a training subset of 181 patients per class.
- T^{172} includes 172 patients per class with the same distribution of image quality and gadolinium injection than the overall data set.
- $T_{\text{no gado, tier } 1/2}^{88}$ includes only matched dementia, NDL and NDNL patients with images of medium or good quality acquired without gadolinium injection. This results in a training subset of 88 patients per class.
- $T_{\text{tier } 1/2}^{88}$ includes 88 patients per class of only images of good or medium quality.
- T^{88} includes 88 subjects per class with the same distribution of image quality and gadolinium injection than the overall data set.

4.5.2 Methods

4.5.2.1 Image processing

The T1w MR images were pre-processed as described in Section 4.2.2 (DICOM to BIDS conversion followed by bias field correction and affine registration to the MNI space, as implemented in the `t1-linear` pipeline of Clinica). This pre-processing was used to assess the quality of the images with the automatic approach described in Section 4.3.

A second pre-processing consisted in applying the `t1-volume-tissue-segmentation` pipeline of Clinica (Routier et al., 2021; Samper-González et al., 2018) to obtain probability grey matter maps. This wrapper of the Unified Segmentation procedure implemented in SPM (Ashburner et al., 2005) simultaneously performs tissue segmentation, bias correction and spatial normalisation. This results in probability grey matter maps in the MNI space that have a size of $121 \times 145 \times 121$ voxels.

To attenuate a potential bias due to the presence or absence of gadolinium, all the images pre-processed with the `t1-linear` pipeline went through the *Att-U-Net* described in Section 4.4 that translates contrast-enhanced images into non-contrast-enhanced images. To prevent introducing a potential bias because of differences in smoothness between the real and synthetic images, all the images were fed to the network no matter the initial presence or absence of gadolinium. The synthetic images were then pre-processed with the `t1-volume-tissue-segmentation` pipeline, as done for the real images.

4.5.2.2 Machine learning models used for classification

Linear support vector machine A linear support vector machine (SVM) using probability grey matter maps as features was used for the binary classification tasks. We followed the implementation of Samper-González et al., 2018 using Scikit-learn (Pedregosa et al., 2011). We optimised the penalty parameter C of the error term. The optimal value of C was chosen using nested CV, with an inner k-fold ($k=10$). For each fold of the outer CV, the value of C that led to the highest balanced accuracy in the inner k-fold was selected.

Convolutional neural networks We used three different 3D CNNs for the binary classification tasks to have a comparison with the linear SVM. Note that the input of the CNNs are the images pre-processed with `t1-linear` as this procedure was validated in (Wen et al., 2020).

The three 3D CNN architectures considered are the same as the ones used for the automatic QC: Conv5_FC3, ResNet, InceptionNet. The first is composed of five convolutional layers and three fully connected layers as implemented in (Wen et al., 2020; Thibeau-Sutre et al., 2022b), the ResNet contains residual blocks inspired from (Jónsson et al., 2019) and the InceptionNet is a modified version of the Inception architecture implemented by (Szegedy et al., 2016).

The models were trained using the cross entropy loss. We used the Adam optimizer with a learning rate of 10^{-5} for the ResNet and of 10^{-4} for the InceptionNet and Conv5_FC3 architectures. We implemented early stopping and all the models were evaluated with a maximum of 50 epochs. The batch size was set to 2. The model with the lowest loss,

determined on the validation set, was saved as final model. Implementation was done using Pytorch through the ClinicaDL platform (Thibeau-Sutre et al., 2022b).

4.5.3 Results

We first classified AD vs CN subjects using the ADNI data set to obtain baseline results on a research data set. Then we performed two tasks using the clinical data sets: dementia vs no dementia with lesions (D vs NDL) and dementia vs no dementia no lesions (D vs NDNL).

4.5.3.1 Performance in a research data set

Results for classification of AD vs CN on ADNI are reported in Table 4.12. The best balanced accuracy was reached using the linear SVM with grey matter maps as input (86.4%), followed by the ResNet (85.3%), the Conv5_FC3 (84.1%), and the InceptionNet (82.1%) using minimally pre-processed T1w MR images as input. These results are in line with the literature (Samper-González et al., 2018; Wen et al., 2020). As training linear SVMs is less computationally expensive than CNNs and since the objective of our work is not to compare different machine learning approaches, for the subsequent experiments we will only report results obtained with the linear SVM.

TABLE 4.12: Dementia classification performance (AD vs CN) on the research data set (ADNI). Results were obtained with different machine learning models: a linear SVM using as input grey matter maps and three CNN models (Conv5_FC3, ResNet and InceptionNet) using as input minimally pre-processed T1w MR images). We present results on the independent test set using the average performance of the five models corresponding to the five folds.

AD vs CN

Metric	SVM	Conv5_FC3	ResNet	InceptionNet
Balanced accuracy	86.80	84.10	85.30	82.10
Sensitivity	82.80	79.80	83.00	75.80
Specificity	90.80	88.40	87.60	88.40

4.5.3.2 Performance in the clinical data set

Classification results on the clinical data set (for both D vs NDNL and D vs NDL) using all the training samples available are reported in Table 4.13. We observed an important drop in balanced accuracy compared with that obtained on the research data set: 68.8% for D vs NDNL and 73.1% for D vs NDL compared with 86.4% for AD vs CN in ADNI. This may due to the heterogeneity of the classes in the clinical data set, where many diagnoses coexist, but also to differences in image characteristics.

Influence of gadolinium injection and image quality on the classification performance As shown in Table 4.11, the proportions of images with and without gadolinium

TABLE 4.13: Dementia classification performance (D vs NDNL and D vs NDL) in the clinical data set. Results were obtained by training a linear SVM on grey matter maps.

Metric	D vs NDNL	D vs NDL
Balanced accuracy	68.75	73.09
Sensitivity	66.97	75.92
Specificity	70.53	70.26

injection and of medium/good vs low quality differ in the dementia, NDL and NDNL categories. In the dementia class, 25% images were acquired with gadolinium injection. In NDL and in NDNL, this proportion is around 65%. In the dementia and NDL categories, the majority of the images are of medium/good quality (58% and 52%, respectively), while in the NDNL category only 36% of images are of medium/good quality. Since these acquisition characteristics are correlated with the diagnostic class, it is possible that the classifier uses this information characteristic, thereby biasing the performance upwards, a phenomenon often referred to as the Clever Hans effect (Lapuschkin et al., 2019).

To test this hypothesis, we used the training subsets $T_{\text{no gado}}^{172}$, $T_{\text{tier } 1/2}^{181}$ and T^{172} . The order of magnitude of patients per class among the training subsets is equivalent, meaning that differences observed in the classification score should not depend on the training sample size but on the characteristics of the training subset. We assume that if gadolinium or image quality has no impact, the performance will not vary when using the different training subsets. On the other hand, if results differ between training subsets, this will be the sign of a Clever Hans effect. Results of these experiments are displayed in Table 4.14. Note that the test set never changed across all the experiments of the work: it is composed of 152 patients/images per class. The balanced accuracy when using T^{172} was substantially higher than when using $T_{\text{no gado}}^{172}$ or $T_{\text{tier } 1/2}^{181}$. This indicates that results are biased by the presence of gadolinium and by the differences in image quality. The classifiers exploit these characteristics to determine the diagnosis.

The training subset $T_{\text{no gado}}^{172}$ still contains images of different quality and $T_{\text{tier } 1/2}^{181}$ images with and without gadolinium. The classifier may thus still be exploiting biases in the image characteristics. To evaluate the performance of the classifier using a training data set without any of these two potential biases, we used the training subset called $T_{\text{no gado, tier } 1/2}^{88}$ and compared it with using the training subset T^{88} having the same training size. Results are reported in Table 4.15. For both tasks, there was a dramatic drop in balanced accuracy, down from 70% to random (50%). Therefore, the classifier is only using the Clever Hans effect and not relevant diagnostic information. In other words, when it cannot exploit biases in image characteristics, the trained classifier is not better than a random classifier.

Classification performance obtained after gadolinium removal using image translation We assessed whether the deep learning-based image translation approach we developed to remove the visual effect of gadolinium from contrast-enhanced T1w MR images could reduce the classification bias due to gadolinium injection. We created a training

TABLE 4.14: Influence of gadolinium injection and image quality on the classification performance. Results were obtained for the D vs NDNL and D vs NDL classification tasks by training a linear SVM on grey matter maps from different clinical data subsets ($T_{\text{no gado}}^{172}$, $T_{\text{tier 1/2}}^{181}$ and T^{172}).

A. D vs NDNL

Metric	$T_{\text{no gado}}^{172}$	$T_{\text{tier 1/2}}^{181}$	T^{172}
Balanced accuracy	60.33	61.32	68.16
Sensitivity	52.76	79.87	73.95
Specificity	67.89	42.76	62.37

B. D vs NDL

Metric	$T_{\text{no gado}}^{172}$	$T_{\text{tier 1/2}}^{181}$	T^{172}
Balanced accuracy	69.74	64.61	72.30
Sensitivity	85.13	45.53	66.45
Specificity	54.34	83.68	78.16

TABLE 4.15: Joint influence of gadolinium injection and image quality on the classification performance. Results were obtained for the D vs NDNL and D vs NDL classification tasks by training a linear SVM on grey matter maps from two clinical data subsets ($T_{\text{no gado, tier 1/2}}^{88}$ and T^{88}).

A. D vs NDNL

Metric	$T_{\text{no gado, tier 1/2}}^{88}$	T^{88}
Balanced accuracy	51.51	69.47
Sensitivity	6.71	71.97
Specificity	96.32	66.97

B. D vs NDL

Metric	$T_{\text{no gado, tier 1/2}}^{88}$	T^{88}
Balanced accuracy	50.00	73.03
Sensitivity	40.00	66.58
Specificity	60.00	79.47

subset composed of 88 synthetic images obtained from images of medium/good quality acquired with and without gadolinium injection that all went through the gadolinium removal *Att-U-Net*. If the gadolinium is successfully removed, training with this subset should be equivalent to training with the $T_{\text{no gado, tier 1/2}}^{88}$ subset that includes only images without gadolinium. Results of these experiments are reported in Table 4.16. The balanced accuracy is equivalent in both cases, meaning that the effect of gadolinium has been removed using the synthetic images. Nevertheless, it is not better than chance indicating, again, that the classifier cannot learn image characteristics which are relevant to the diagnostic classification.

However, it is possible that the low performance is due to the small size of the training set. We therefore used the image translation method to build a larger clinical data set composed only of images of medium/good quality and where the visual appearance has been removed, this data set was denoted Synthetic $T_{\text{tier 1/2}}^{181}$. Using this training set, we assessed both the linear SVM (using grey matter maps) and the ResNet (with minimally pre-processed T1w MRI as input). Results appear in Table 4.17. We found increased performance using this synthetic, larger, training set. The ResNet obtained a slightly higher performance than

the SVM. It is thus possible that homogenizing the data set using image translation allows removing bias and increasing classification performance. Nevertheless, we cannot directly demonstrate this in the absence of a training set of the same size containing only images without gadolinium and of higher quality. It is thus possible that visually imperceptible differences still exist between the images that were initially acquired with gadolinium and those without, and that the classifiers exploit these differences.

TABLE 4.16: Classification performance obtained after gadolinium removal using image translation, training on a set of 88 patients. Results were obtained for the D vs NDNL and D vs NDL classification tasks with a linear SVM using as input grey matter maps and trained on three clinical data subsets ($T_{\text{tier } 1/2}^{88}$, $T_{\text{tier } 1/2}^{\text{Synthetic}}$, $T_{\text{no gado, tier } 1/2}^{88}$).

A. D vs NDNL

Metric	$T_{\text{tier } 1/2}^{88}$	Synthetic $T_{\text{tier } 1/2}^{88}$	$T_{\text{no gado, tier } 1/2}^{88}$
Balanced accuracy	60.26	51.71	51.51
Sensitivity	58.68	75.66	6.71
Specificity	61.84	27.76	96.32

B. D vs NDL

Metric	$T_{\text{tier } 1/2}^{88}$	Synthetic $T_{\text{tier } 1/2}^{88}$	$T_{\text{no gado, tier } 1/2}^{88}$
Balanced accuracy	68.29	54.08	50.00
Sensitivity	69.34	52.50	40.00
Specificity	67.24	55.66	60.00

TABLE 4.17: Classification performance obtained after gadolinium removal using image translation, training on a set of 181 patients. Results were obtained a linear SVM with probability grey matter maps or a ResNet with minimally pre-processed T1w MR images.

A. D vs NDNL

Metric	SVM	ResNet
Balanced accuracy	61.91	63.22
Sensitivity	81.32	52.24
Specificity	42.50	74.21

B. D vs NDL

Metric	SVM	ResNet
Balanced accuracy	64.61	67.50
Sensitivity	45.53	64.47
Specificity	83.68	70.53

Classification performance when training on a research data set and testing on the clinical data set

Another way to ensure that gadolinium or poor image quality is not exploited by the classifier is to train using the research data set (ADNI contains only images without gadolinium and of good quality). We both trained a linear SVM and a ResNet. Results appear in Table 4.18. No matter the task, the linear SVM trained on research data led to a slightly higher balanced accuracy than the ResNet. Note that the accuracy was also slightly higher than when training with synthetic data (Table 4.17). In any case, one should keep in mind that these classification performance are too low to be acceptable in clinical practice.

TABLE 4.18: Classification performance when training on a research data set and testing on a clinical data set. Results were obtained for the D vs NDNL and D vs NDL classification tasks using a linear SVM with probability grey matter maps or a ResNet with minimally pre-processed T1w MR images.

A. D vs NDNL			B. D vs NDL		
Metric	SVM	ResNet	Metric	SVM	ResNet
Balanced accuracy	64.08	61.84	Balanced accuracy	69.47	61.78
Sensitivity	62.76	60.92	Sensitivity	62.76	60.92
Specificity	65.39	62.76	Specificity	76.18	62.63

4.5.4 Discussion

In this work, we studied the performance of machine learning approaches for computer-aided detection of dementia based on T1w MRI using a real-life clinical routine cohort coming from an hospital data warehouse. To the best of our knowledge, this is the first of this kind since previous works have used either research data sets or clinical data from specialised centres that have been carefully selected and are thus not representative of daily clinical routine. We demonstrated that the classifiers trained on clinical routine data are highly biased by image acquisition specificities such as image quality or injection of gadolinium. When such biases are removed, the performance is very poor. Models trained on research data perform better but their accuracy remains unacceptably low for clinical use.

As a research topic, machine learning for diagnosis of Alzheimer’s disease is now almost 15 year old (Klöppel et al., 2008; Vemuri et al., 2008; Gerardin et al., 2009; Fan et al., 2008). While high performance has been consistently reported, most of these works use research data sets for training and validation (Samper-González et al., 2018; Falahati et al., 2014; Manera et al., 2021; Bron et al., 2021). There are a few papers using clinical routine data sets but they cannot be considered representative of daily clinical routine as they come from a single or a handful of highly specialised centers and carefully select data using strict criteria regarding data quality (Morin et al., 2020; Platero et al., 2019; Sohn et al., 2015; Klöppel et al., 2015). It is thus unclear how such methods would perform on real-life clinical MRI and ultimately translate to the clinic.

The main results of our work are three-fold: i) the performance of such CAD methods is considerably lower on clinical routine data compared to research data sets; ii) on clinical routine data, classifiers were heavily biased by irrelevant characteristics and when such biases were removed, the performance was particularly low; iii) training on research data and testing on clinical data allowed reaching slightly higher accuracies but the overall performance remained low. More specifically, when both training and testing on research data, we obtained high classification performance (around 87% balanced accuracy) which is in line with the literature. When training/testing on clinical data, the performance dropped by more than 15 percent points and, more importantly, was heavily biased by irrelevant characteristics. When such confounders were removed, the performance was around the chance level. Training on the research data set and testing on the clinical routine data set allowed removing this source of bias but the performance remained poor (decrease of at least 19 percent points of balanced accuracy). Thus, classifiers that lead to high classification

performance in a research framework do not necessarily generalise to clinical data set. Part of this drop in accuracy could be explained by an increase in the difficulty of the classification task between the research and clinical setups. In the research setup, the AD and CN classes are quite homogeneous, while in the clinical setup, the D, NDL and NDNL classes are much more heterogeneous as each category corresponds to several diagnoses. However, this may not be the only factor leading to this performance difference and more analyses were performed to dissect these results.

In the clinical routine data set, there was a clear correlation between the diagnostic groups on the one hand and image quality and presence of gadolinium on the other hand (65% of images with gadolinium in NDL and NDNL and 25% in D; 37% of images of medium or good quality in NDNL, and 55% in D and NDL). We hypothesised that models trained on such data could exploit this bias. To assess this, we trained different models changing the characteristics of the training subsets: we used training subsets having only images without gadolinium ($T_{\text{no gado}}^{172}$) or images of medium/good quality ($T_{\text{tier } 1/2}^{181}$) or both ($T_{\text{no gado, tier } 1/2}^{88}$) and we compared their performance with a training subset of the same sample size but having the same proportions of images with gadolinium and of low quality than the whole data set (T^{172} and T^{88}). We showed that the performance of the classifier was heavily biased by these image characteristics, a phenomenon known as the Clever Hans effect (Lapuschkin et al., 2019). Such phenomenon has been previously described in different medical image computing applications (Lapuschkin et al., 2019; Wallis et al., 2022).

The primary aim of this work was not to find the most efficient machine learning algorithm but to evaluate the performance of well-known methods. For this reason, most of the experiments were conducted using a simple linear SVM using grey matter maps as inputs. To justify this choice, we first confirmed that its performance on research data was in line with the literature and comparable to more advanced deep learning classification methods (namely ResNet and Inception). At the end of the study, we also confirmed that a deep learning method did not reach substantially higher performance on clinical routine data. Of course, it does not mean that sophisticated deep learning methods could not achieve higher performance using larger, unbiased, clinical routine data sets but it was not the case in our work.

We aimed to remove the bias coming from gadolinium injection by applying an image translation *Att-U-Net* model that we proposed (Bottani et al., 2022c). On the smaller set of 88 patients, its performance was close to chance and similar to that of a classifier trained on images without gadolinium and of good/medium quality. When using a larger data set of synthetic images, we obtained higher accuracies. This potentially indicates that the use of image translation allows removing some of the biases while improving performance. Nevertheless, we cannot strictly assert this because there may be residual, visually imperceptible differences between images that were acquired with gadolinium and those without. Overall, this stresses the importance of developing image homogenisation techniques for training unbiased classifiers.

Our study has the following limitations. Unlike in research studies, the diagnosis may not be trustworthy as it is assigned using ICD-10 codes, which could be a source of bias. Indeed, in the French healthcare system, they are assigned during hospitalisation by the

clinical department for the billing of the expenses. In addition, ICD-10 codes do not undergo quality control and it is likely that mistakes occur when entering the codes. These limitations of the diagnostic labels may hamper the performance of the classifiers. In order to have more reliable diagnostic labels, it would be necessary to use information from medical reports. This could be done by medical experts but this is time consuming and may not scale up to large populations. Another option is to use natural language processing but it may also lead to errors. Other limitations concern the training data set we have used: due to the choices done we have reduced the sample size. Further evaluations should be done to assess whether the performance of the classifiers could improve according to the present work by adding more subjects in the training. Finally, we have limited our experimental settings to the use of a linear SVM or CNN models, but more improvements could be done using other models or other CNN architectures with different hyper-parameters.

Overall, our results highlight the challenges for translation of CAD systems from research to clinical routine. A major result of this study is uncovering the strong influence of biases coming from image heterogeneity. We specifically studied the case of gadolinium injection and image quality but other sources of biases such as image resolution, sequence parameters or scanner type could exist. They could in turn induce Clever Hans effects on the CAD systems if they are correlated with the diagnosis of interest. This highlights the need for automatic quality control tools in order to identify the various sources of biases as well as for homogenisation tools that could remove these biases.

4.6 Perspectives

This work constitutes, to the best of our knowledge, the first evaluation of machine learning and deep learning algorithms for the computer-assisted diagnosis of dementia on a large set of images acquired in clinical routine. It demonstrates that translation of such algorithms from research to clinical practice requires great efforts, which are too often overlooked. On the bright side, it opens the way to many research areas. We have seen all along the chapter the importance of quality control and feature homogenisation to guarantee unbiased diagnostic predictions. A first area of research would consist in improving these two elements.

The automatic QC approach we proposed allows accurately discarding images that are of no interest, recognising gadolinium injection and detecting low quality images, but could better differentiate high and medium quality images. Instead of predicting the overall image quality, we could train networks to independently rate image characteristics (motion, contrast and noise). In this case, it would be possible to simulate various degrees of motion, contrast and noise on high quality research data to pre-train networks before training on the clinical data. The detection of motion, contrast and noise artefacts could be followed by an artefact reduction step.

A source of heterogeneity among clinical data sets is the fact that they contain a mix of images acquired with and without gadolinium-based contrast agent. As a first step towards the homogenisation of such data set, we proposed a framework to convert T1w-ce images into T1w-nce images using deep learning models. However, heterogeneity within a clinical data set can arise from other sources, such as the use of different MRI scanner machines

or different acquisition parameters. Future works should study their influence and propose models to achieve a more general homogenisation, for example as proposed in (Cackowski et al., 2021).

Images are not the only data whose quality can bias CAD systems. As mentioned at the end of Section 4.5, the ICD-10 codes used to define the diagnostic labels may not be trustworthy as in France they are used for billing. Two strategies then open up. One would consist in thoroughly evaluating this potential bias and propose solutions to reduce it. Another would consist in developing CAD systems relying on other types of data that would be more reliable, for instance results of lab tests.

In terms of imaging modalities, the current work only exploited T1w MRI but others, particularly T2-weighted fluid-attenuated inversion recovery MRI, would be relevant to detect the presence of white matter hyperintensities, which are characteristic of certain diseases such as vascular dementia.

General conclusions and perspectives

Anomaly detection for the computer-aided diagnosis of dementia

In clinical practice, many imaging modalities such as PET are analysed visually, but the sensitivity and specificity of this approach greatly depend on the observer's experience. We proposed an anomaly detection framework for computer-aided diagnosis from PET images. Subject-specific abnormality maps were obtained by comparing the subject's PET image to a model of healthy PET appearance that is specific to the subject under investigation (Burgos et al., 2021b).

This work targets an important clinical application, that of being able to identify pathological areas of the brain in an unsupervised way, i.e., that is agnostic about the disease studied. The approach has successfully been applied in the context of dementia (Alzheimer's disease, frontotemporal dementia and primary progressive aphasia) but abnormality maps could be a useful tool for other neurological disorders such as epilepsy. PET has become widely used in the presurgical workup of drug-resistant epilepsy and the abnormality maps generated for such patients could help better localise the epileptogenic zones.

The current approach relies on traditional medical imaging computing methods (registration, modality fusion) but we believe that deep generative approaches such as variational autoencoders would be beneficial in detecting subtler anomalies, and in a more computationally efficient manner. This is the topic of a current PhD thesis (Ravi Hassanaly).

Finally, until now, the approach has only been applied to the detection of anomalies in PET images. The approach could be extended to other imaging modalities, such as anatomical MRI or diffusion weighted imaging (for instance fractional anisotropy or mean diffusivity maps).

Interpretable computer-aided diagnosis of dementia

The anomaly detection approach described above provides a visual feedback by construction. It is however not the case of all computer-aided diagnosis tools, such as deep learning classifiers, and interpretability methods are needed to validate them and ensure their reliability (Thibeau-Sutre et al., 2022a). We extended an existing framework to visualise the training of CNNs on quantitative imaging data in the context of Alzheimer's disease classification (Thibeau-Sutre et al., 2020). After assessing the robustness of the visualisation method, we studied the stability of the CNN training between different re-runs. We observed thanks to the visualisation method that the CNN training is not stable and mainly depends on the initialisation and training process. Our conclusion is that, currently, combining a CNN

classifier with an interpretability method does not constitute a robust tool for individual computer-aided diagnosis.

The fact that they appear as a ‘black-box’ to clinicians is not the only limitation of machine learning and deep learning approaches used to predict a diagnosis. Current classification algorithms, usually developed for two-class problems, are too rigid to be successfully applied in a clinical setting where patients may have several pathologies or a pathology that has never been seen during training, and where more data become available every day. Search and retrieval of brain images is a promising strategy to overcome the ‘black-box’ effect and handle multiple pathologies, thus facilitating adoption by clinicians. New decision support systems could be based on methods that can, for a given patient, retrieve similar clinical cases from existing databases in an efficient manner. Similarity between cases could be obtained by comparing the abnormality maps generated for each patient.

Reproducible computer-aided diagnosis of dementia

As other fields of science, machine learning-based computer-aided diagnosis also faces a reproducibility crisis. For instance, in Alzheimer’s disease, multiple works have been published using the public dataset ADNI but they are difficult to reproduce and an objective comparison between approaches is almost impossible. We proposed frameworks for reproducible neuroimaging-based computer-aided diagnosis, which main components are: 1) automatic conversion of public data sets into a standard format; 2) feature extraction pipelines; 3) baseline classification approaches; 4) procedures to enforce good practices for validation. We first introduced this framework for classification from anatomical MRI and PET data (Samper-González et al., 2018). The corresponding code has been implemented in Clinica, an open-source software platform for neuroimaging studies (Routier et al., 2021). We further extended our work to deep learning approaches (Wen et al., 2020). This has led to the development of a new software platform, ClinicaDL, devoted to reproducible deep learning for neuroimaging (Thibeau-Sutre et al., 2022b).

ClinicaDL is a software platform dedicated to deep learning for neuroimaging and has for ambition to offer its users tools to reproduce their experiments as reliably as possible. This will be made possible by implementing tools to track and manage both the data and model parameters, to automatically deploy the trained models and to provide a long-term storage. Such software developments will make our research more reproducible, which is essential to build trust in the tools we propose.

Computer-aided diagnosis of dementia from routine clinical data

Most machine learning-based computer-aided diagnosis approaches have been designed and validated using research data sets. It is thus not clear how they would generalise to clinical routine and ultimately what is their medical value. We are currently exploiting real-life data from hospital data warehouses (specifically the data warehouse of the 39 Greater Paris hospitals - AP-HP, comprising millions of patients). The first challenges we have dealt with

are the quality and heterogeneity of the imaging data. We built a deep learning system for automatic quality control of anatomical T1-weighted MRI (Bottani et al., 2022b) and developed approaches for data homogenisation (Bottani et al., 2022d; Bottani et al., 2021). We then studied the performance of computer-aided diagnosis algorithms on clinical routine data and showed that the performance was considerably lower than on a research data set (Bottani et al., 2022a). These are foundational works that will be instrumental for our future work on computer-aided diagnosis.

Translation of computer-aided diagnosis tools from research to clinical practice is not straightforward. Our work with anatomical MR images from the AP-HP data warehouse demonstrated the importance of quality control and data set homogenisation. Future work will focus on these two aspects before targeting computer-aided diagnosis as such. This is the topic of a current PhD thesis (Sophie Loizillon).

*
* *

In conclusion, by identifying brain anomalies at the scale of the individual, exploring how to make deep neural networks interpretable, developing reproducible analysis pipelines and validating our approaches on clinical data, we progress towards the application of systems for the computer-aided diagnosis of dementia to the clinic, but there is still a long way to go.

CV and publications

Ninon Burgos

CNRS RESEARCHER • MEDICAL IMAGE COMPUTING

ARAMIS Lab, Institut du Cerveau / Paris Brain Institute • ICM
Hôpital de la Pitié-Salpêtrière, 47 boulevard de l'hôpital, 75013 Paris, France

☎ +33 (0)1 57 27 43 48 • ✉ ninon.burgos@cnrs.fr • 🌐 ninonburgos.com • 📱 ninonburgos • 🗉 Ninon Burgos

Education

PhD in Medical and Biomedical Imaging

UNIVERSITY COLLEGE LONDON

London, UK

2016

MSc in Biomedical Engineering

IMPERIAL COLLEGE LONDON

London, UK

2012

Diplôme d'Ingénieur

ÉCOLE NATIONALE SUPÉRIEURE D'ÉLECTRONIQUE ET DE SES APPLICATIONS (ENSEA)

Cergy, France

2012

Academic Positions

PR[AI]RIE junior fellow

PARIS ARTIFICIAL INTELLIGENCE RESEARCH INSTITUTE (PR[AI]RIE)

Paris, France

2019 – present

CNRS researcher (Faculty position, equivalent to Associate Professor with tenure and no strict teaching duties)

INSTITUT DU CERVEAU – ARAMIS LAB (SORBONNE UNIVERSITÉ, CNRS UMR 7225, INRIA, INSERM U1127, AP-HP)

Paris, France

2018 – present

Postdoctoral researcher

INRIA/INSTITUT DU CERVEAU ET DE LA MOELLE ÉPINIÈRE – ARAMIS LAB (INSERM U1127, CNRS UMR 7225, SORBONNE UNIVERSITÉ)

Paris, France

2017 – 2018

- Advisor: Olivier Colliot
- Project: Differential diagnosis of dementia through identification of abnormality patterns in multimodal brain imaging

Postdoctoral researcher

CENTRE FOR MEDICAL IMAGE COMPUTING, UNIVERSITY COLLEGE LONDON

London, UK

2016

- Advisor: M. Jorge Cardoso
- Project: Towards automatic MR-based radiotherapy treatment planning

Research assistant

CENTRE FOR MEDICAL IMAGE COMPUTING, UNIVERSITY COLLEGE LONDON

London, UK

2021 – 2016

- Advisors: Prof. Sébastien Ourselin, Prof. Brian Hutton, Dr M. Jorge Cardoso
- Thesis: Image synthesis for the attenuation correction and analysis of PET/MR data

Publications

28	International journal articles	9 as first/last author, 8 as second/second-last author
1	Edited book	with David Svoboda
4	Edited conference proceedings	2 as main editor
4	Book chapters	2 as main author
19	Conferences with full-length peer-reviewed proceedings	13 as first/last author, 2 as second author
36	Conference abstracts	9 as first/last author, 10 as second/second-last author

Funding

2019 –	PR[AI]RIE Springboard Chair , 'Investissements d'avenir' programme from the French government under management of Agence Nationale de la Recherche (ANR-19-P3IA-0001)	185k €
2017 –	PRESTIGE Postdoctoral Research Fellowship , Campus France and the Marie Curie Actions—COFUND of the European Union's Seventh Framework Programme	30k €
2016	CMIC Pump-priming Award , Six months of funding received from the CMIC-EPSRC platform grant (EP/M020533/1) to explore a new field of research	

Society Memberships

International society for optics and photonics (SPIE)

Since 2021

Medical Image Computing and Computer Assisted Intervention (MICCAI) Society

2013, 2015 – present

Organization for Human Brain Mapping (OHBM)

2018 – 2020, 2022

Honours & Awards

2019	ERCIM Cor Baayen Young Researcher Award , Awarded each year to a promising young researcher in computer science or applied mathematics in Europe	
2017	Galileo Galilei Award 2017 , Best publication in the European Journal of Medical Physics - Physica Medica in 2017	
2016	Marie Curie Fellow and PRESTIGE Fellow , Campus France and the Marie Curie Actions—COFUND of the European Union's Seventh Framework Programme	
2016	Student travel award , International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)—attribution based on the quality of the paper (acceptance rate below 35%)	Athens, Greece
2016	Highlighted presentation , International Conference on the use of Computers in Radiation Therapy (ICCR)	London, UK
2015	Student travel award , MICCAI	Munich, Germany
2015	Best oral presentation runner-up award , 4th Conference on PET/MR and SPECT/MR (PSMR)	Elba, Italy
2013	Student travel award , MICCAI	Nagoya, Japan

Invited Presentations

INVITED PRESENTATIONS AT INTERNATIONAL EVENTS

AI4Health Winter School

Virtual

'INTRODUCTION TO DEEP LEARNING FOR MEDICAL IMAGING: FROM CONVOLUTION TO GENERATIVE ADVERSARIAL NETWORKS'

Jan 2022

Mathematics and Image Analysis - MIA'21

Virtual

'IMPROVING THE INTERPRETABILITY OF COMPUTER-ASSISTED ANALYSIS TOOLS IN NEUROIMAGING'

Jan 2021

Annual Congress of the European Association of Nuclear Medicine

Vienna, Austria

'MR-BASED ATTENUATION CORRECTION FOR BRAIN STUDIES' (CONTINUING MEDICAL EDUCATION SESSION)

Oct 2017

INVITED PRESENTATIONS AT NATIONAL EVENTS & WORKSHOPS

Journée Scientifique PLBS - ImAgerie & IA

Lille, France

'IA ET NEUROIMAGERIE: LE RÈGNE DES RÉSEAUX DE NEURONES CONVOLUTIFS'

Nov 2022

Congrès des Jeunes Chercheuses et Chercheurs en Mathématiques Appliquées

Palaiseau, France

'IA POUR L'IMAGERIE MÉDICALE : DE L'ACQUISITION DES IMAGES À LA PRISE DE DÉCISION CLINIQUE'

Oct 2021

Registering Medical Images

Paris, France

'ON THE INTERPLAY BETWEEN MEDICAL IMAGE REGISTRATION AND SYNTHESIS'

Oct 2021

3e colloque sur l'imagerie médicale à l'heure de l'intelligence artificielle

Paris Brain Institute, France

'COMPUTER-AIDED DIAGNOSIS FROM NEUROIMAGES: A FRAMEWORK FOR OBJECTIVE & REPRODUCIBLE EXPERIMENTS'

Oct 2020

MaDICS Symposium

Rennes, France

'REPRODUCIBLE EVALUATION OF METHODS FOR THE DIAGNOSIS AND PROGNOSIS OF ALZHEIMER'S DISEASE'

June 2019

Neuro OpenScience Workshop

Paris Brain Institute, France

'COLLABORATIVE NEUROIMAGING TOOLS'

Jan 2019

Workshop on Machine Learning in Radiology

Lausanne University Hospital, Switzerland

'REPRODUCIBLE EVALUATION OF CLASSIFICATION METHODS IN ALZHEIMER'S DISEASE'

Nov 2018

Young Researchers' Futures Meeting 2016

London, UK

'JOINT SEGMENTATION AND CT SYNTHESIS IN THE PELVIC REGION FOR MRI-ONLY RADIOTHERAPY TREATMENT PLANNING'

Sept 2016

Data processing challenges in PET-MR

London, UK

MULTI-ATLAS CT & ATTENUATION MAP SYNTHESIS FOR HYBRID PET-MR SCANNERS

Jan 2015

Experts' MR brain attenuation correction workshop

Copenhagen, Denmark

'CT & ATTENUATION MAP SYNTHESIS IN THE BRAIN REGION FOR HYBRID PET-MR SCANNERS'

Oct 2014

INVITED SEMINARS

BME Paris Seminars « Open Your Mind »

'WHY SHOULD RESEARCHERS SPEND TIME WRITING GOOD CODE?'

Arts et Métiers, Paris

September 2022

Master 2 Mathématiques pour les Sciences du Vivant

'IMPROVING THE INTERPRETABILITY OF COMPUTER-ASSISTED ANALYSIS TOOLS IN NEUROIMAGING'

Virtual

January 2022

Séminaire Médecine et Humanités de l'ENS

'AI FOR THE LIFE SCIENCES' (VIDEO)

ENS, Paris

November 2021

Bioinfo seminars of the Labex Memolife

'REPRODUCIBLE COMPUTER-AIDED DIAGNOSIS OF ALZHEIMER'S DISEASE USING DEEP LEARNING'

Virtual

April 2021

iBrain seminars

'TOWARDS THE INDIVIDUAL COMPUTER-ASSISTED ANALYSIS OF BRAIN IMAGES'

Université de Tours, France

Nov 2019

ARAMIS Lab seminars

'IMAGE SYNTHESIS FOR THE ATTENUATION CORRECTION AND ANALYSIS OF PET/MR DATA'

Paris, France

Sept 2016

Institute of Nuclear Medicine seminars

'ATTENUATION MAP SYNTHESIS FOR HYBRID PET-MR SCANNERS: A CLINICAL PERSPECTIVE'

University College London Hospitals, UK

May 2015

Supervision of Research Activities

PHD THESES

Sophie Loizillon

Co-supervision with Olivier Colliot

'Deep learning for assisting diagnosis of neurological diseases using a very large-scale clinical data warehouse' [E.2]

Oct 2021 – present

Ravi Hassanaly

'Deep generative models for the detection of anomalies in the brain' [E.1]

Primary supervision

Nov 2020 – present

Simona Bottani

'Machine learning for neuroimage processing using a very large-scale clinical data warehouse' [A.2, E.5, F.5, H.1]

Co-supervision with Olivier Colliot

Oct 2018 – March 2022

Elina Thibeau-Sutre

'Reproducible and interpretable deep learning for the diagnosis, prognosis and subtyping of Alzheimer's disease from neuroimaging data' [A.1, A.11, E.3 E.4, E.6, F.1, F.2, F.8, A.3]

Co-supervision with Didier Dormont and Olivier Colliot

Sept 2018 – Dec 2021

Jorge Samper-González

'Learning from multimodal data for classification and prediction of Alzheimer's disease' [A.14, E.7, E.9, F.10, F.18, F.20]

Co-supervision with Olivier Colliot

Jan 2017 – Dec 2019

MASTER THESES

Maëlys Solal

'Deep learning for anomaly detection in neuroimages for the computer-aided diagnosis of dementia'

Primary supervision

Oct 2022 – June 2023

Arnaud Berenbaum

'Automatic classification of brain PET/CT scans with deep learning' [H.5]

Co-supervision with Aurélie Kas and Olivier Colliot

Mar 2021 – Sept 2021

Ravi Hassanaly

'Pseudo-healthy image synthesis for the detection of anomalies in the brain, a deep learning approach'

Primary supervision

Apr 2020 – Sept 2020

Pablo Rey

'Individual analysis of diffusion weighted imaging data'

Primary supervision

June 2018 – Aug 2018

ENGINEERS

Camille Brianceau

Developer of ClinicaDL, a software for reproducible neuroimaging processing with deep learning

July 2022 – present

Matthieu Joulot

Developer of Clinica, focusing on dataset converters and diffusion MRI pipelines

June 2021 – present

Ghislain Vaillant

Developer of the web service ClinicaCloud

May 2021 – present

Omar El Rifai

Lead developer of Clinica, a software platform for clinical neuroimaging research studies [F.3]

Mar 2021 – Oct 2022

Adam Wild

Developer of software tools to process massive medical imaging datasets [A.2]

Jan 2019 – June 2020

Alexandre Routier

Lead developer of Clinica, a software platform for clinical neuroimaging research studies [A.7, F.6, F.9, F.17]

Nov 2018 – Oct 2020

Arnaud Marcoux

Developer of software tools to process multimodal medical images (PET and MRI) [A.15, F.16]

Feb 2017 – Feb 2020

Software Development & Management

- | | | |
|------------------|---|---|
| Clinica | <ul style="list-style-type: none">Open-source software platform for clinical neuroimaging research studiesRole: Management of the project and of the developers | www.clinica.run
github.com/aramis-lab/clinica |
| ClinicaDL | <ul style="list-style-type: none">Open-source deep learning software for reproducible neuroimaging processingRole: Management of the project and of the developers | github.com/aramis-lab/clinicaDL |
| NiftySeg | <ul style="list-style-type: none">Open-source image segmentation and parcellation softwareRole: Contributor of novel algorithms for image synthesis | github.com/KCL-BMEIS/NiftySeg |
| NiftyWeb | <ul style="list-style-type: none">Web service tool for the fully automated synthesis of CT from MRI imagesRole: Creator of the pCT web service tool | niftyweb.cs.ucl.ac.uk/program.php?p=PCT |

Transfer of Technology

Transfer to clinical research

The image synthesis method that I developed during my PhD for the attenuation correction of PET/MR data is currently integrated into the image processing pipeline of several dementia studies at the Dementia Research Centre (UCL Institute of Neurology), such as Insight 46—a neuroscience sub-study of the MRC National Survey for Health and Development, involving 1000 PET/MR acquisitions [A.21, A.22, A.24, A.25, A.28, A.29].

Transfer to industry

The attenuation correction method raised the interest of Oncovision, a company dedicated to the development, manufacturing and distribution of medical image devices, resulting in the signature of a commercial agreement.

Other Professional Activities

EDITORSHIP

- | | |
|--------------------|---|
| Book | Burgos, N., Svoboda, D., eds.: Biomedical Image Synthesis Simulation: Methods and Applications, MICCAI Book series, Elsevier, 2022. 10.1016/C2020-0-01250-8 |
| Conferences | MIDL Technical Committee (2022), SASHIMI Programme Chair (2019, 2020) and Co-Chair (2018, 2021) |

REVIEW ([Web of Science profile](#))

- | | |
|-----------------------------|---|
| Journals (selection) | IEEE Transactions on Medical Imaging; Medical Image Analysis; IEEE Transactions on Pattern Analysis and Machine Intelligence; IEEE Transactions on Image Processing; PLOS ONE; Scientific Reports; Artificial Intelligence Review; Communications Biology; NeuroImage; Frontiers in Neuroscience; Medical Physics; Neurocomputing; Neuroinformatics; Journal of Nuclear Medicine; International Journal of Radiation Oncology, Biology, Physics; Journal of Alzheimer's Disease |
| Conferences | MICCAI (2016, 2020–2022), ISBI (2018, 2020–2023), MIDL (2018, 2020), SASHIMI (2018–2022), OHBM (2019, 2020, 2022) |
| Grants | ERC Advanced Grants (2020), Luxembourg National Research Fund (2020), National Science Centre Poland (2020), DIM ELICIT (2021), Alzheimer's Society (2021), ANR JCJC (2022) |

PARTICIPATION TO RECRUITMENT JURIES

- | | | |
|------|---|----------------------------------|
| 2020 | Jury member , Permanent researcher competitive recruitment procedure of the Inria Paris centre (concours CRCN) | France (virtual) |
|------|---|----------------------------------|

PARTICIPATION TO PHD COMMITTEES AND JURIES

- | | | |
|------------|--|----------------------------------|
| 2022 | PhD jury member , Gauthier Dot, supervised by Thomas Schouman, Laurent Gajny and Philippe Rouch | Paris, France |
| 2022 | Mid-thesis committee member , Francesco Galati, supervised by Maria A. Zuluaga | Paris, France |
| 2022 | Mid-thesis committee member , Camille Ruppli, supervised by Isabelle Bloch | Paris, France |
| 2021, 2022 | Mid-thesis committee member , Charlotte Godard, supervised by Jean-Baptiste Masson | France (virtual) |

SCHOOL ORGANISATION

- | | | |
|------|--|----------------------------------|
| 2022 | Scientific & Organisation Committees , AI4Health Winter School (ai4healthschool.org) | France (virtual) |
| 2021 | Scientific & Organisation Committees , AI4Health Winter School (ai4healthschool.org) | France (virtual) |

WORKSHOP ORGANISATION

2021	Programme & Organisation committees , Simulation and Synthesis in Medical Imaging (SASHIMI) 2021, a satellite workshop of MICCAI 2021 (www.sashimi.aramislab.fr)	Strasbourg, France (virtual)
2020	Programme Chair & Organisation Committee , SASHIMI 2020	Lima, Peru (virtual)
2020	Organisation Committee , CompAge 2020: Computational approaches for ageing and age-related diseases (compage2020.com)	Paris, France (virtual)
2020	Organisation Committee , Hands-on Workshop on Machine Learning Applied to Medical Imaging (laclauc.github.io/workshop)	Paris, France
2019	Programme Chair & Organisation Committee , SASHIMI 2019	Shenzhen, China
2018	Programme & Organisation Committees , SASHIMI 2018	Granada, Spain

TEACHING

Since 2021	CENIR courses , Deep Learning for Medical Imaging (1h30)	Paris Brain Institute
Since 2021	AI4Health Winter School , Practical session on Deep Learning for Medical Imaging (2x8h)	Virtual
Since 2020	DU Intelligence artificielle IA appliquée en santé , Deep Learning for Medical Imaging (1h)	Université de Paris
2020, 2022	DIU Neuroradiologie diagnostique et thérapeutique , Deep Learning for Neuro Imaging (1h)	Sorbonne Université
2020	Educational Courses of the OHBM 2020 conference , Machine Learning for NeuroImaging (30 min)	Virtual
2020	Hands-on Workshop on Machine Learning Applied to Medical Imaging , Introduction to Deep Learning & Deep Learning for Neuro Imaging (3h)	Paris Brain Institute
2018	Educational Courses of the OHBM 2018 conference , Pattern Recognition for NeuroImaging (45 min)	Singapore

TRAINING COURSES FOLLOWED

02/2022	London Mathematical Society (LMS) Invited Lecturers Series , Mathematics of Deep Learning	University of Cambridge (online)
12/2018	Formation continue des encadrants , Management d'un projet doctoral	Sorbonne Université
12/2018	FUN MOOC , Intégrité scientifique dans les métiers de la recherche par l'Université de Bordeaux	Online

SCIENTIFIC ANIMATION

Since 2019	Member of the scientific animation committee at the Paris Brain Institute , Participating to the organisation of weekly plenary talks from prestigious high-profile international speakers (e.g., Yann LeCun, Nick Fox, Katrin Amunts)	
------------	---	--

DISSEMINATION OF SCIENTIFIC KNOWLEDGE

2022	MIT-France Symposium on AI , Presentation on AI-based computer-aided diagnosis of dementia	Collège de France, Paris, France
Since 2019	Rendez-vous des Jeunes Mathématiciennes et Informaticiennes , Presentation and discussion with high school girls	Inria Paris, France
2021	Paris Brain Institute Donors' Conference , Presentation on the computer-aided diagnosis of Alzheimer's disease	Paris Brain Institute
2021	MIT Symposium on AI & Medicine: Promises and Limits , Panel discussion on image-guided clinical practice	Virtual
2020	France is AI , Panel discussion on AI in decision support systems with medical images	Virtual
2017	Fête de la science , Science fair showcasing research done within the ARAMIS Lab	Paris Brain Institute, France
2015	University College London Hospitals Research Open Day , Focus on clinical research	London, UK

MEDIA COVERAGE

- 2022 **Podcast ‘Braincast - La voix des neurones’ by Cerveau & Psycho magazine on the use of AI for the diagnosis of Alzheimer’s disease**, <https://www.cerveauetpsycho.fr/sr/braincast>
- 2021 **Interview for an article published in the magazine “Femme Actuelle Senior” on the use of AI for computer-aided diagnosis**, N°42
- 2019 **Interview published on the Inria website following the ERCIM Cor Baayen Young Researcher Award**, <https://www.inria.fr/en/ninon-burgos-wins-2019-ercim-cor-baayen-young-researcher-award-her-work-computational-imaging>
- 2019 **Interview published on the CNRS INS2I website following the ERCIM Cor Baayen Young Researcher Award**, <https://ins2i.cnrs.fr/fr/cnrsinfo/ninon-burgos-des-outils-informatiques-pour-detecter-des-maladies-comme-alzheimer> (in French)
- 2017 **Interview for the Nuclear Medicine and Molecular Medicine Podcast following an invited presentation at the Annual Congress of the European Association of Nuclear Medicine**, https://nucomedpodcast.blogspot.fr/2017/12/episode-74-n-burgos-and-attenuation_20.html
- 2017 **Interview published in the MICCAI Daily magazine, section “Women in Science”**, <http://www.rsipvision.com/MICCAI2017-Wednesday>

Ninon Burgos

LIST OF PUBLICATIONS

Contents

A International journal publications	1
B Book	3
C Book chapters	3
D Conference proceedings	4
E Conferences with full-length peer-reviewed proceedings	4
F Conference abstracts	5
G Thesis	7
H Submitted publications and preprints	8

Note that articles preceded by a ★ are the product of doctoral projects that I (co-)supervised.

A International journal publications

- A.1 ★ Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Routier, A., Didier, D., Colliot, O., **Burgos, N.**, ‘ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing’, *Computer Methods and Programs in Biomedicine*, 220: 106818, 2022. [doi:10.1016/j.cmpb.2022.106818](https://doi.org/10.1016/j.cmpb.2022.106818) • [hal-03351976](https://hal.archives-ouvertes.fr/hal-03351976)
- A.2 ★ Bottani, S., **Burgos, N.**, Maire, A., Wild, A., Ströer, S., Dormont, D., Colliot, O.: ‘Automatic Quality Control of Brain T1-Weighted Magnetic Resonance Images for a Clinical Data Warehouse’, *Medical Image Analysis*, 75: 102219, 2022. [doi:10.1016/j.media.2021.102219](https://doi.org/10.1016/j.media.2021.102219) • [hal-03154792](https://hal.archives-ouvertes.fr/hal-03154792)
- A.3 ★ Chadebec, C., Thibeau-Sutre, E., **Burgos, N.**, Allasonnière, S., ‘Data augmentation on neuroimaging data with variational autoencoders’. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. [doi:10.1109/TPAMI.2022.3185773](https://doi.org/10.1109/TPAMI.2022.3185773) • [arXiv: 2105.00026](https://arxiv.org/abs/2105.00026)
- A.4 Epelbaum, S., **Burgos, N.**, Canney, M., Matthews, D., Houot, M., Santin, M. D., Desseaux, C., Bouchoux, G., Ströer, S., Martin, C., Habert, M.-O., Levy, M., Bah, A., Martin, K., Delatour, B., Riche, M., Dubois, B., Belin, L., Carpentier, A., ‘Pilot study of repeated blood-brain barrier disruption in patients with mild Alzheimer’s disease with an implantable ultrasound device’. *Alzheimer’s Research & Therapy*, 14(1): 40, 2022. [doi:10.1186/s13195-022-00981-1](https://doi.org/10.1186/s13195-022-00981-1) • [hal-03484130](https://hal.archives-ouvertes.fr/hal-03484130)
- A.5 **Burgos, N.**, Bottani, S., Faouzi, J., Thibeau-Sutre, E., Colliot, O.: ‘Deep learning in brain disorders: from data processing to disease treatment’. *Briefings in Bioinformatics*, 22(2): 1560–1576, 2021. [doi:10.1093/bib/bbaa310](https://doi.org/10.1093/bib/bbaa310) • [hal-03070554](https://hal.archives-ouvertes.fr/hal-03070554) — **INVITED REVIEW**
- A.6 **Burgos, N.**, Cardoso, M.J., Samper-González, J., Habert, M.-O., Durrleman, S., Ourselin, S., Colliot, O.: ‘Anomaly Detection for the Individual Analysis of Brain PET Images’. *Journal of Medical Imaging*, 8(2): 024003, 2021. [doi:10.1117/1.JMI.8.2.024003](https://doi.org/10.1117/1.JMI.8.2.024003) • [hal-03193306](https://hal.archives-ouvertes.fr/hal-03193306)
- A.7 Routier, A., **Burgos, N.**, Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau-Sutre, E., Vailant, G., Wen, J., Wild, A., Habert, M.-O., Durrleman, S., Colliot, O.: ‘Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies’. *Frontiers in Neuroinformatics*, 15: 39, 2021. [doi:10.3389/fninf.2021.689675](https://doi.org/10.3389/fninf.2021.689675) • [hal-02308126](https://hal.archives-ouvertes.fr/hal-02308126) — **47 CITATIONS ACCORDING TO GOOGLE SCHOLAR**

- A.8 Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-González, J., **Burgos, N.**, Charlier, B., Bertrand, A., Epelbaum, S., Colliot, O., Allassonnière, S., Durrleman, S.: 'AD Course Map charts Alzheimer's disease progression', *Scientific Reports*, 11(1): 8020, 2021. doi:10.1038/s41598-021-87434-1 • hal-01964821
- A.9 Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronne, R., Faouzi, J., Koval, I., Louis, M., Thibeau-Sutre, E., Wen, J., Wild, A., **Burgos, N.**, Dormont, D., Colliot, O., Durrleman, S.: 'Predicting the Progression of Mild Cognitive Impairment Using Machine Learning: A Systematic Quantitative Review', *Medical Image Analysis*, 67: 101848, 2021. doi:10.1016/j.media.2020.101848 • hal-02337815 — **40 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.10 **Burgos, N.**, Colliot, O.: 'Machine Learning for Classification Prediction of Brain Diseases: Recent Advances Upcoming Challenges'. *Current Opinion in Neurology*, 33(4): 439–450, 2020. doi:10.1097/WCO.0000000000000838 • hal-02902586 — **INVITED REVIEW, 19 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.11 ★ Wen, J., Thibeau-Sutre, E., Samper-González, J., Routier, A., Bottani, S., Durrleman, S., **Burgos, N.**, Colliot, O.: 'Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview Reproducible Evaluation', *Medical Image Analysis*, 63: 101694, 2020. doi:10.1016/j.media.2020.101694 • hal-02562504 — **264 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.12 Couvy-Duchesne, B., Faouzi, J., Martin, B., Thibeau-Sutre, E., Wild, A., Ansart, M., Durrleman, S., Dormont, D., **Burgos, N.**, Colliot, O.: 'Ensemble Learning of Convolutional Neural Network, Support Vector Machine, Best Linear Unbiased Predictor for Brain Age Prediction: ARAMIS Contribution to the Predictive Analytics Competition 2019 Challenge'. *Frontiers in Psychiatry*, 11, Frontiers, 2020. doi:10.3389/fpsy.2020.593336 • hal-03136463
- A.13 Wen, J., Samper-González, J., Bottani, S., Routier, A., **Burgos, N.**, Jacquemont, T., Fontanella, S., Durrleman, S., Epelbaum, S., Bertrand, A., Colliot, O.: 'Reproducible Evaluation of Diffusion MRI Features for Automatic Classification of Patients with Alzheimer's Disease', *Neuroinformatics*, 2020. doi:10.1007/s12021-020-09469-5 • hal-02566361
- A.14 ★ Samper-González, J., **Burgos, N.**, Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., Colliot, O.: 'Reproducible Evaluation of Classification Methods in Alzheimer's Disease: Framework Application to MRI PET Data'. *NeuroImage*, 183: 504–521, 2018. doi:10.1016/j.neuroimage.2018.08.042 • hal-01858384 — **128 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.15 Marcoux, A., **Burgos, N.**, Bertrand, A., Teichmann, M., Routier, A., Wen, J., Samper-Gonzalez, J., Bottani, S., Durrleman, S., Habert, M.-O., Colliot, O.: 'An Automated Pipeline for the Analysis of PET Data on the Cortical Surface'. *Frontiers in Neuroinformatics*, 12, 2018. doi:10.3389/fninf.2018.00094
- A.16 Arabi, H., Dowling, J. A., **Burgos, N.**, Han, X., Greer, P. B., Koutsouvelis, N. Zaidi, H.: 'Comparative Study of Algorithms for Synthetic CT Generation from MRI: Consequences for MRI-Guided Radiation Planning in the Pelvic Region'. *Medical Physics*, 45(11): 5218–5233, 2018. doi:10.1002/mp.13187 • hal-01890646 — **95 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.17 Kieselmann, J. P., Kamerling, C. P., **Burgos, N.**, Menten, M. J., Ding, Y., Fuller, C. D., Jomaa, M. K., Petkar, I., McCormick, G., Hunt, A., Nill, S., Cardoso, M. J., Oelfke, U.: 'Geometric Dosimetric Evaluations of Atlas-Based Segmentation Methods of MR Images in the Head Neck Region'. *Physics in Medicine Biology*, 63(14): 145007, 2018. doi:10.1088/1361-6560/aac665 — **27 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.18 Scott, C.J., Jiao, J., Cardoso, M.J., Melbourne, A., Thomas, D.L., De Vita, E., **Burgos, N.**, Markiewicz, P., Schott, J.M., Hutton, B.F., Ourselin, S.: 'Reduced Acquisition Time PET Quantification Using Simultaneously Acquired Arterial Spin Labelled MRI'. *Journal of Cerebral Blood Flow Metabolism*, 2018. doi:10.1177/0271678X18797343
- A.19 **Burgos, N.**, Guerreiro, F., McClelland, J., Presles, B., Modat, M., Nill, S., Dearnaley, D., deSouza, N., Oelfke, U., Knopf, A.-C., Ourselin, S., Cardoso, M.J.: 'Iterative Framework for the Joint Segmentation CT Synthesis of MR Images: Application to MRI-Only Radiotherapy Treatment Planning'. *Physics in Medicine Biology*, 62(11): 4237, 2017. doi:10.1088/1361-6560/aa66bf — **AN INVITED PAPER IN THE SPECIAL ISSUES ON RECENT PROGRESS IN APPLICATIONS OF COMPUTING TO RADIOTHERAPY, 40 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.20 Guerreiro, F., **Burgos*, N.**, Dunlop, A., Wong, K., Petkar, I., Nutting, C., Harrington, K., Bhide, S., Newbold, K., Dearnaley, D., deSouza, N.M., Morgan, V.A., McClelland, J., Nill, S., Cardoso, M.J., Ourselin, S., Oelfke, U., Knopf, A.C.: 'Evaluation of a Multi-Atlas CT Synthesis Approach for MRI-Only Radiotherapy Treatment Planning'. *Physica Medica*, 35: 7–17, 2017 (*: joint first authorship). doi:10.1016/j.ejmp.2017.02.017 — **GALILEO GALILEI AWARD 2017 • BEST PUBLICATION IN THE EUROPEAN JOURNAL OF MEDICAL PHYSICS - PHYSICA MEDICA IN 2017, 61 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.21 Ladefoged, C.N., Law, I., Anazodo, U., St. Lawrence, K., Izquierdo-Garcia, D., Catana, C., **Burgos, N.**, Cardoso, M.J., Ourselin, S., Hutton, B., Mérida, I., Costes, N., Hammers, A., Benoit, D., Holm, S., Juttukonda, M., An, H., Cabello, J., Lukas, M., Nekolla, S., Ziegler, S., Fenchel, M., Jakoby, B., Casey, M.E., Benzinger, T., Højgaard, L., Hansen, A.E., Andersen, F.L.: 'A Multi-Centre Evaluation of Eleven Clinically Feasible Brain PET/MRI Attenuation Correction Techniques Using a Large Cohort of Patients'. *NeuroImage*, 147: 346–359, 2017. doi:10.1016/j.neuroimage.2016.12.010 — **194 CITATIONS ACCORDING TO GOOGLE SCHOLAR**

- A.22 Lane, C.A., Parker, T.D., Cash, D.M., Macpherson, K., Donnachie, E., Murray-Smith, H., Barnes, A., Barker, S., Beasley, D.G., Bras, J., Brown, D., **Burgos, N.**, Byford, M., Jorge Cardoso, M., Carvalho, A., Collins, J., De Vita, E., Dickson, J.C., Epie, N., Espak, M., Henley, S.M.D., Hoskote, C., Hutel, M., Klimova, J., Malone, I.B., Markiewicz, P., Melbourne, A., Modat, M., Schrag, A., Shah, S., Sharma, N., Sudre, C.H., Thomas, D.L., Wong, A., Zhang, H., Hardy, J., Zetterberg, H., Ourselin, S., Crutch, S.J., Kuh, D., Richards, M., Fox, N.C., Schott, J.M.: ‘Study Protocol: Insight 46 • a Neuroscience Sub-Study of the MRC National Survey of Health Development’. *BMC Neurology*, 17: 75, 2017. doi:10.1186/s12883-017-0846-x — **72 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.23 Jiao, J., Bousse, A., Thielemans, K., **Burgos, N.**, Weston, P.S.J., Schott, J.M., Atkinson, D., Arridge, S.R., Hutton, B.F., Markiewicz, P., Ourselin, S.: ‘Direct Parametric Reconstruction with Joint Motion Estimation/Correction for Dynamic Brain PET Data’. *IEEE Transactions on Medical Imaging*, 36(1): 203–213, 2017. doi:10.1109/TMI.2016.2594150
- A.24 Sekine, T., **Burgos, N.**, Warnock, G., Huellner, M., Buck, A., Voert, E.E.G.W. ter, Cardoso, M.J., Hutton, B.F., Ourselin, S., Veit-Haibach, P., Delso, G.: ‘Multi Atlas-Based Attenuation Correction for Brain FDG- PET Imaging Using a TOF-PET/MR Scanner: Comparison with Clinical Single Atlas- CT-Based Attenuation Correction’. *Journal of Nuclear Medicine*, 57(8): 1258–1264, 2016. doi:10.2967/jnumed.115.169045 — **32 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.25 **Burgos, N.**, Cardoso, M.J., Thielemans, K., Modat, M., Dickson, J., Schott, J.M., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Multi-Contrast Attenuation Map Synthesis for PET/MR Scanners: Assessment on FDG Florbetapir PET Tracers’. *European Journal of Nuclear Medicine Molecular Imaging*, 42(9): 1447–1458, 2015. doi:10.1007/s00259-015-3082-x — **48 CITATIONS ACCORDING TO GOOGLE SCHOLAR**
- A.26 Zuluaga*, M.A., **Burgos***, N., Mendelson, A.F., Taylor, A.M., Ourselin, S.: ‘Voxelwise Atlas Rating for Computer Assisted Diagnosis: Application to Congenital Heart Diseases of the Great Arteries’. *Medical Image Analysis*, 26(1): 185–194, 2015 (*: joint first authorship). doi:10.1016/j.media.2015.09.001
- A.27 Kochan, M., Daga, P., **Burgos, N.**, White, M., Cardoso, M.J., Mancini, L., Winston, G.P., McEvoy, A.W., Thornton, J., Yousry, T., Duncan, J.S., Stoyanov, D., Ourselin, S.: ‘Simulated Field Maps for Susceptibility Artefact Correction in Interventional MRI’. *International Journal of Computer Assisted Radiology Surgery*, 10(9): 1405–1416, 2015. doi:10.1007/s11548-015-1253-7
- A.28 Weston, P.S.J., Paterson, R.W., Modat, M., Burgos, N., Cardoso, M.J., Magdalinou, N., Lehmann, M., Dickson, J.C., Barnes, A., Bomanji, J.B., Kayani, I., Cash, D.M., Ourselin, S., Toombs, J., Lunn, M.P., Mummery, C.J., Warren, J.D., Rossor, M.N., Fox, N.C., Zetterberg, H., Schott, J.M.: ‘Using Florbetapir Positron Emission Tomography to Explore Cerebrospinal Fluid Cut Points Gray Zones in Small Sample Sizes’. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(4): 440–446, 2015. doi:10.1016/j.dadm.2015.10.001
- A.29 **Burgos, N.**, Cardoso, M.J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C.J., Schott, J.M., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Application to Brain Studies’. *IEEE Transactions on Medical Imaging*, 33(12): 2332–2341, 2014. doi:10.1109/TMI.2014.2340135 — **330 CITATIONS ACCORDING TO GOOGLE SCHOLAR**

B Book

- B.1 **Burgos, N.**, Svoboda, D., eds.: *Biomedical Image Synthesis Simulation: Methods and Applications*, MICCAI Book series, Elsevier, 2022. doi:10.1016/C2020-0-01250-8

C Book chapters

- C.1 Svoboda, D., **Burgos, N.**: ‘Introduction to Medical Biomedical Image Synthesis’. In *Biomedical Image Synthesis Simulation: Methods and Applications*, edited by **Burgos, N.** and Svoboda, D., MICCAI Book series, Elsevier, 2022. doi:10.1016/B978-0-12-824349-7.00008-6 • hal-03721967
- C.2 **Burgos, N.**: ‘Medical Image Synthesis Using Segmentation Registration’. In *Biomedical Image Synthesis Simulation: Methods and Applications*, edited by **Burgos, N.** and Svoboda, D., MICCAI Book series, Elsevier, 2022. doi:10.1016/B978-0-12-824349-7.00011-6 • hal-03721697
- C.3 Nečasová, T., **Burgos, N.**, Svoboda, D.: ‘Validation Evaluation Metrics for Medical Biomedical Image Synthesis’. In *Biomedical Image Synthesis Simulation: Methods and Applications*, edited by **Burgos, N.** and Svoboda, D., MICCAI Book series, Elsevier, 2022. doi:10.1016/B978-0-12-824349-7.00032-3 • hal-03721947
- C.4 **Burgos, N.**, Tsiftaris, S., Svoboda, D.: ‘Future Trends in Medical Image Synthesis’. In *Biomedical Image Synthesis Simulation: Methods and Applications*, edited by **Burgos, N.** and Svoboda, D., MICCAI Book series, Elsevier, 2022. doi:10.1016/B978-0-12-824349-7.00034-7 • hal-03721950

D Conference proceedings

- D.1 Svoboda, D., **Burgos, N.**, Wolterink, J.M., Zhao, C., eds.: Simulation Synthesis in Medical Imaging: 6th International Workshop, SASHIMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 2021, Proceedings. Vol. 12965. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021. doi:10.1007/978-3-030-87592-3
- D.2 **Burgos, N.**, Svoboda, D., Wolterink, J.M., Zhao, C., eds.: Simulation Synthesis in Medical Imaging: 5th International Workshop, SASHIMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 2020, Proceedings. Vol. 12417. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020. doi:10.1007/978-3-030-59520-3
- D.3 **Burgos, N.**, Gooya, A., Svoboda, D., eds.: Simulation Synthesis in Medical Imaging: 4th International Workshop, SASHIMI 2019, Held in conjunction with MICCAI 2019, Shenzhen, China, October 2019, Proceedings. Vol. 11827. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019. doi:10.1007/978-3-030-32778-1
- D.4 Gooya, A., Goksel, O., Oguz, I., **Burgos, N.**, eds.: Simulation Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 2018, Proceedings. Vol. 11037. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018. doi:10.1007/978-3-030-00536-8

E Conferences with full-length peer-reviewed proceedings

- E.1 ★ Hassanaly, R., Bottani, S., Sauty, B., Colliot, O., **Burgos, N.**: ‘Simulation-based evaluation framework for deep learning unsupervised anomaly detection on brain FDG PET’. In *SPIE Medical Imaging 2023*, 2023 (accepted). hal-03835015
- E.2 ★ Loizillon, S., Bottani, S., Maire, A., Ströer, S., Dormont, D., Colliot, O., **Burgos, N.**: ‘Transfer learning from synthetic to routine clinical data for motion artefact detection in brain T1-weighted MRI’. In *SPIE Medical Imaging 2023*, 2023 (accepted). hal-03831746
- E.3 ★ Thibeau-Sutre, E., Wolterink, J. M., Colliot, O., **Burgos, N.**: ‘How can data augmentation improve attribution maps for disease subtype explainability?’. In *SPIE Medical Imaging 2023*, 2023 (accepted).
- E.4 ★ Thibeau-Sutre, E., Couvy-Duchesne, B., Dormont, D., Colliot, O., **Burgos, N.**: ‘MRI field strength predicts Alzheimer’s disease: A case example of bias in the ADNI data set’. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–4, 2022. doi:10.1109/ISBI52829.2022.9761504 • hal-03542213
- E.5 ★ Bottani, S., Thibeau-Sutre, E., Maire, A., Ströer, S., Dormont, D., Colliot, O., **Burgos, N.**: ‘Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-Net derived models’. In *SPIE Medical Imaging 2022*, 12032:576–582, 2022. doi:10.1117/12.2608565 • hal-03478798
- E.6 ★ Thibeau-Sutre, E., Colliot, O., Dormont, D., **Burgos, N.**: ‘Visualization Approach to Assess the Robustness of Neural Networks for Medical Image Classification’. In *SPIE Medical Imaging 2020*, 11313: 113131J, 2020. doi:10.1117/12.2548952 • hal-02370532 — **ORAL PRESENTATION**
- E.7 ★ Samper-González, J., **Burgos, N.**, Bottani, S., Habert, M.-O., Evgeniou, T., Epelbaum, S., Colliot, O.: ‘Reproducible Evaluation of Methods for Predicting Progression to Alzheimer’s Disease from Clinical Neuroimaging Data.’ In *SPIE Medical Imaging 2019*, 10949:109490V, 2019. doi:10.1117/12.2512430 • hal-02025880 — **ORAL PRESENTATION**
- E.8 **Burgos, N.**, Samper-González, J., Bertrand, A., Habert, M.-O., Ourselin, S., Durrleman, S., Cardoso, M.J., Colliot, O.: ‘Individual Analysis of Molecular Brain Imaging Data through Automatic Identification of Abnormality Patterns’. In *Molecular Imaging, Reconstruction Analysis of Moving Body Organs, Stroke Imaging Treatment*, LNCS, 10555: 13–22, Springer, 2017. doi:10.1007/978-3-319-67564-0_2 • hal-01567343 — **ORAL PRESENTATION**
- E.9 ★ Samper-González, J., **Burgos, N.**, Fontanella, S., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., Colliot, O.: ‘Yet Another ADNI Machine Learning Paper? Paving the Way towards Fully-Reproducible Research on Classification of Alzheimer’s Disease’. In *Machine Learning in Medical Imaging*, LNCS, 10541: 53–60, Springer, 2017. doi:10.1007/978-3-319-67389-9_7 • hal-01578479
- E.10 Scott, C.J., Jiao, J., Cardoso, M.J., Melbourne, A., De Vita, E., Thomas, D.L., **Burgos, N.**, Markiewicz, P., Schott, J.M., Hutton, B.F., Ourselin, S.: ‘Short Acquisition Time PET Quantification Using MRI-Based Pharmacokinetic Parameter Synthesis’. In *Medical Image Computing Computer-Assisted Intervention • MICCAI 2017*, LNCS, 10434: 737–744, Springer, 2017. doi:10.1007/978-3-319-66185-8_83
- E.11 **Burgos, N.**, Guerreiro, F., McClelland, J., Nill, S., Dearnaley, D., deSouza, N., Oelfke, U., Knopf, A.-C., Ourselin, S., Cardoso, M.J.: ‘Joint Segmentation CT Synthesis for MRI-Only Radiotherapy Treatment Planning’. In *Medical Image Computing Computer-Assisted Intervention • MICCAI 2016*, LNCS, 9901: 547–555, Springer, 2016. doi:10.1007/978-3-319-46723-8_63 — **ACCEPTANCE RATE BELOW 35%, STUDENT TRAVEL AWARD**
- E.12 **Burgos, N.**, Cardoso, M.J., Guerreiro, F., Veiga, C., Modat, M., McClelland, J., Knopf, A.-C., Punwani, S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Robust CT Synthesis for Radiotherapy Planning: Application to the Head & Neck Region’. In *Medical Image Computing Computer-Assisted Intervention • MICCAI 2015*, LNCS, 9350: 476–484, Springer, 2015. doi:10.1007/978-3-319-24571-3_57 — **ACCEPTANCE RATE BELOW 35%, STUDENT TRAVEL AWARD, 54 CITATIONS ACCORDING TO GOOGLE SCHOLAR**

- E.13 **Burgos, N.**, Cardoso, M.J., Mendelson, A.F., Schott, J.M., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Subject-Specific Models for the Analysis of Pathological FDG PET Data’. In *Medical Image Computing Computer-Assisted Intervention • MICCAI 2015*, LNCS, 9350: 651–658, Springer, 2015. doi:10.1007/978-3-319-24571-3_78 — **ACCEPTANCE RATE BELOW 35%, STUDENT TRAVEL AWARD**
- E.14 Jiao, J., Markiewicz, P., **Burgos, N.**, Atkinson, D., Hutton, B., Arridge, S., Ourselin, S.: ‘Detail-Preserving PET Reconstruction with Sparse Image Representation Anatomical Priors’. In *Information Processing in Medical Imaging*, LNCS, 9123: 540–551, Springer, 2015. doi:10.1007/978-3-319-19992-4_42
- E.15 Zuluaga*, M.A., **Burgos***, N., Taylor, A.M., Ourselin, S.: ‘Multi-Atlas Synthesis for Computer Assisted Diagnosis: Application to Cardiovascular Diseases’. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 290–293, 2015 (*: joint first authorship). doi:10.1109/ISBI.2015.7163870 — **ORAL PRESENTATION**
- E.16 **Burgos, N.**, Thielemans, K., Cardoso, M.J., Markiewicz, P., Jiao, J., Dickson, J., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Effect of Scatter Correction When Comparing Attenuation Maps: Application to Brain PET/MR’. In *2014 IEEE Nuclear Science Symposium Medical Imaging Conference (NSS/MIC)*, 1–5, 2014. doi:10.1109/NSS-MIC.2014.7430775
- E.17 Jiao, J., Bousse, A., Thielemans, K., Markiewicz, P., **Burgos, N.**, Atkinson, D., Arridge, S., Hutton, B.F., Ourselin, S.: ‘Joint Parametric Reconstruction Motion Correction Framework for Dynamic PET Data’. In *Medical Image Computing Computer-Assisted Intervention • MICCAI 2014*, LNCS, 8673: 114–121, Springer, 2014. doi:10.1007/978-3-319-10404-1_15
- E.18 Kochan, M., Daga, P., **Burgos, N.**, White, M., Cardoso, M.J., Mancini, L., Winston, G.P., McEvoy, A.W., Thornton, J., Yousry, T., Duncan, J.S., Stoyanov, D., Ourselin, S.: ‘Simulated Field Maps: Toward Improved Susceptibility Artefact Correction in Interventional MRI’. In *Information Processing in Computer-Assisted Interventions*, LNCS, 8498: 226–235, Springer, 2014. doi:10.1007/978-3-319-07521-1_24
- E.19 **Burgos, N.**, Cardoso, M.J., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: ‘Attenuation Correction Synthesis for Hybrid PET-MR Scanners’. In *Medical Image Computing Computer-Assisted Intervention • MICCAI 2013*, LNCS, 8149: 147–154, Springer, 2013. doi:10.1007/978-3-642-40811-3_19 — **ACCEPTANCE RATE BELOW 35%, STUDENT TRAVEL AWARD, 62 CITATIONS ACCORDING TO GOOGLE SCHOLAR**

F Conference abstracts

- F.1 ★ Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Colliot, O., and **Burgos, N.**: ‘A Glimpse of ClinicaDL, an Open-Source Software for Reproducible Deep Learning in Neuroimaging’. In *Medical Imaging with Deep Learning - MIDL 2022* (short paper), 2022. [Open Review gsqiNMdPSYK](#)
- F.2 ★ Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Colliot, O., and **Burgos, N.**: ‘ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing’. In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2022*, 2022.
- F.3 El Rifai, O., Díaz, M., Hassanaly, R., Joulot, M., Routier, A.M., Thibeau-Sutre, E., Vaillant, G., Durrleman, S., **Burgos, N.**, and Colliot, O.: ‘Advances in the Clinica software platform for clinical neuroimaging studies’. In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2022*, 2022. [hal-03728243](#)
- F.4 Canney, M., Epelbaum, S., **Burgos, N.**, Matthews, D., Houot, M., Santin, M. D., Desseaux, C., Bouchoux, G., Ströer, S., Martin, C., Habert, M.-O., Levy, M., Martin, K., Delatour, B., Riche, M., Dubois, B., Belin, L., Carpentier, A., ‘Pilot study of blood-brain barrier disruption in Alzheimer’s disease’. In *21st Annual International Symposium on Therapeutic Ultrasound - ISTU 2022*, 2022.
- F.5 Maire, A., Bottani, S., Jacob, Y., Ströer, S., **Burgos, N.**, Colliot, O., Dormont, D., Hilka, M.: ‘Apports de la Plateforme Données Massive AP-HP pour la recherche en IA: le projet APPRIMAGE’. In *Journées Francophones de Radiologie*, 2021.
- F.6 Routier, A., Marcoux, A., Melo, M.D., Samper-González, J., Wild, A., Guyot, A., Wen, J., Thibeau-Sutre, E., Bottani, S., Durrleman, S., **Burgos, N.**, Colliot, O.: ‘New Longitudinal Deep Learning Pipelines in the Clinica Software Platform’. In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2020*, 2020. [hal-02549242](#)
- F.7 Cash, D.M., Markiewicz, P.J., Jiao, J., Coath, W., Modat, M., Lane, C.A., Parker, T.D., Keuss, S.E., Buchanan, S.M., **Burgos, N.**, Dickson, J., Barnes, A., Cardoso, J., Alves, I.L., Barkhof, F., Thomas, D.L., Beasley, D., Wong, A., Schöll, M., Richards, M., Ourselin, S., Fox, N.C., and Schott, J.M.: ‘Comparison of Static and Dynamic Analysis Techniques for Longitudinal Analysis of Amyloid PET’. In *Alzheimer’s Association International Conference - AAIC 2020*, 2020.
- F.8 ★ Wen, J., Thibeau-Sutre, E., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Colliot, O., **Burgos, N.**: ‘How Serious Is Data Leakage in Deep Learning Studies on Alzheimer’s Disease Classification?’ In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2019*, 2019. [hal-02105133](#)
- F.9 Routier, A., Marcoux, A., Díaz Melo, M., Guillon, J., Samper-González, J., Wen, J., Bottani, S., Guyot, A., Thibeau-Sutre, E., Teichmann, M., Habert, M.-O., Durrleman, S., **Burgos, N.**, Colliot, O.: ‘New Advances in the Clinica Software Platform for Clinical Neuroimaging Studies’. In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2019*, 2019. [hal-02132147](#)

- F.10 ★ Samper-González, J., **Burgos, N.**, Bottani, S., Habert, M.-O., Evgeniou, T., Epelbaum, S., Colliot, O.: 'Predicting Progression to Alzheimer's Disease from Clinical Imaging Data: A Reproducible Study.' In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2019*, 2019. [hal-02142315](#)
- F.11 Ansart, M., **Burgos, N.**, Colliot, O., Dormont, D., Durrleman, S.: 'Prediction of Future Cognitive Scores Dementia Onset in Mild Cognitive Impairment Patients.' In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2019*, 2019. [hal-02098427](#)
- F.12 Koval, I., Marcoux, A., **Burgos, N.**, Allasonnière, S., Colliot, O., Durrleman, S.: 'Deciphering the Progression of PET Alterations Using Surface-Based Spatiotemporal Modeling.' In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2019*, 2019. [hal-02134909](#)
- F.13 Wen, J., Samper-González, J., Routier, A., Bottani, S., Durrleman, S., **Burgos, N.**, Colliot, O.: 'Beware of Feature Selection Bias! Example on Alzheimer's Disease Classification from Diffusion MRI.' In *Annual Meeting of the Organization for Human Brain Mapping - OHBM 2019*, 2019. [hal-02105134](#)
- F.14 Cash, D.M., Modat, M., Coath, W., Cardoso, M.J., Markiewicz, P., Lane, C.A., Parker, T., Keuss, S., Buchanan, S., **Burgos, N.**, Dickson, J., Barnes, A., Thomas, D.L., Beasley, D., Malone, I.B., Erlandsson, K., Thomas, B.A., Ourselin, S., Fox, N.C., Schott, J.M., Richards, M.: 'Longitudinal Rates of Amyloid Accumulation in a 70-Year Old British Birth Cohort'. In *Alzheimer's Association International Conference - Aaic 2019*, 2019.
- F.15 Coath, W., Modat, M., Cardoso, M.J., Markiewicz, P., Lane, C.A., Parker, T., Keuss, S., Buchanan, S., **Burgos, N.**, Dickson, J., Barnes, A., Thomas, D.L., Beasley, D., Malone, I.B., Wong, A., Thomas, B.A., Ourselin, S., Richards, M., Fox, N.C., Schott, J.M., Cash, D.M.: 'Centiloid Scale Transformation of Florbetapir Data Acquired on a PET/MR Scanner'. In *Alzheimer's Association International Conference - Aaic 2019*, 2019.
- F.16 Marcoux, A., **Burgos, N.**, Bertrand, A., Routier, A., Wen, J., Samper-González, J., Bottani, S., Durrleman, S., Habert, M.-O., Colliot, O.: 'A pipeline for the analysis of 18F-FDG PET data on the cortical surface its evaluation on ADNI'. *Annual Meeting of the Organization for Human Brain Mapping • OHBM 2018*, 2018. [hal-01757646](#)
- F.17 Routier, A., Guillon, J., **Burgos, N.**, Samper-González, Wen, J., Fontanella, S., Bottani, S., Jacquemont, T., Marcoux, A., Gori, P., Lu, P., Moreau, T., Bacci, M., Durrleman, S., Colliot, O.: 'Clinica: an open source software platform for reproducible clinical neuroscience studies'. *Annual Meeting of the Organization for Human Brain Mapping • OHBM 2018*, 2018. [hal-01760658](#)
- F.18 ★ Samper-González, J., Bottani, S., **Burgos, N.**, Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., Colliot, O.: 'Reproducible evaluation of Alzheimer's disease classification from MRI PET data'. *Annual Meeting of the Organization for Human Brain Mapping • OHBM 2018*, 2018. [hal-01761666](#)
- F.19 Wen, J., Samper-González, J., Bottani, S., Routier, A., **Burgos, N.**, Jacquemont, T., Fontanella, S., Durrleman, S., Bertrand, A., Colliot, O.: 'Comparison of DTI features for the classification of Alzheimer's disease: A reproducible study'. *Annual Meeting of the Organization for Human Brain Mapping • OHBM 2018*, 2018. [hal-01758206](#)
- F.20 ★ Samper-González, J., **Burgos, N.**, Bottani, S., Habert, M.-O., Evgeniou, T., Epelbaum, S., Colliot, O.: 'Three Simple Ideas for Predicting Progression to Alzheimer's Disease.' In *International Workshop on Pattern Recognition in Neuroimaging - PRNI 2018*, 2018. [hal-01891996](#)
- F.21 Wen, J., Samper-González, J., Bottani, S., Routier, A., **Burgos, N.**, Jacquemont, T., Fontanella, S., Durrleman, S., Bertrand, A., Colliot, O.: 'Using diffusion MRI for classification prediction of Alzheimer's Disease: a reproducible study'. *Alzheimer's Association International Conference - Aaic 2018*, 2018. [hal-01758167](#)
- F.22 **Burgos, N.**, Samper-González, J., Bertrand, A., Habert, M.-O., Ourselin, S., Durrleman, S., Cardoso, M.J., Colliot, O.: 'Diagnosis of Alzheimer's Disease through Identification of Abnormality Patterns in FDG PET Data'. In *Proceedings of the 30th Annual Congress of the European Association of Nuclear Medicine (EANM)*, S253–S254, Springer, 2017. [doi:10.1007/s00259-017-3822-1](#) • [hal-01632509](#) — **ORAL PRESENTATION**
- F.23 **Burgos, N.**, Samper-González, J., Cardoso, M.J., Durrleman, S., Ourselin, S., Colliot, O.: 'Early Diagnosis of Alzheimer's Disease Using Subject-Specific Models of FDG-PET Data'. *Alzheimer's & Dementia*, 13(7): P1117, 2017. [doi:10.1016/j.jalz.2017.06.1618](#) • [hal-01621383](#)
- F.24 Cash, D.M., **Burgos, N.**, Modat, M., Dickson, J., Beasley, D., Markiewicz, P., Lane, C.A., Parker, T., Barnes, A., Thomas, D.L., Cardoso, M.J., Malone, I.B., Veale, T., Wallon, D., Klimova, J., Erlandsson, K., Wong, A., Richards, M., Fox, N.C., Ourselin, S., Schott, J.M.: 'A Comparison of Techniques for Quantifying Amyloid Burden on a Combined PET/MR Scanner'. *Alzheimer's & Dementia*, 13(7): P12–P13, 2017. [doi:10.1016/j.jalz.2017.06.2276](#)
- F.25 Schott, J.M., Cash, D.M., Lane, C.A., Parker, T., **Burgos, N.**, Modat, M., Beasley, D., Dickson, J., Barnes, A., Thomas, D.L., Murray-Smith, H., Wong, A., Macpherson, K., James, S.-N., Cardoso, M.J., Malone, I.B., Klimova, J., Markiewicz, P., Crutch, S.J., Kuh, D., Ourselin, S., Richards, M., Fox, N.C.: 'Exploring the Population Prevalence of β -Amyloid Burden: An Analysis of 250 Individuals Born in Main Britain in the Same Week in 1946'. *Alzheimer's & Dementia*, 13(7): P1088–P1089, 2017. [doi:10.1016/j.jalz.2017.06.1563](#)

- F.26 James, S.-N., Parker, T., Lane, C.A., Cash, D.M., Wong, A., Barnes, A., Beasley, D., **Burgos, N.**, Cardoso, M.J., Dickson, J., Klimova, J., Malone, I.B., Modat, M., Thomas, D.L., Kuh, D., Ourselin, S., Fox, N.C., Schott, J.M., Richards, M.: 'Midlife Affective Symptoms Are Associated with Lower Brain Volumes in Later Life: Evidence from a Prospective UK Birth Cohort'. *Alzheimer's & Dementia*, 13(7): P212, 2017. doi:10.1016/j.jalz.2017.07.086
- F.27 Parker, T., Cash, D.M., Lane, C.A., Murray-Smith, H., Wong, A., Malone, I.B., **Burgos, N.**, Modat, M., Beasley, D., Dickson, J., Barnes, A., Thomas, D.L., Cardoso, M.J., Klimova, J., Ourselin, S., Frost, C., Kuh, D., Richards, M., Fox, N.C., Schott, J.M.: 'Brain volume, cerebral β -amyloid deposition, ageing: A study of over 200 individuals born in the same week in 1946'. *Alzheimer's & Dementia*, 13(7): P1464–P1465, 2017. doi:10.1016/j.jalz.2017.07.534
- F.28 Kieselmann, J. P., Kamerling, C. P., **Burgos, N.**, Menten, M. J., Nill, S., Cardoso, M. J., Oelfke, U.: 'Geometric Dosimetric Evaluation of Three Atlas-based Segmentation Methods for Head Neck Cancer Patients on MR Images'. *MR in RT Symposium*, 2017
- F.29 **Burgos, N.**, Cardoso, M.J., Guerreiro, F., McClelland, J., Knopf, A.-C., Ourselin, S.: 'Simultaneous Organ-at-Risk Segmentation CT Synthesis in the Pelvic Region for MRI-Only Radiotherapy Treatment Planning'. In *International Conference on the Use of Computers in Radiation Therapy (ICCR)*, 2016 — **HIGHLIGHTED ORAL PRESENTATION**
- F.30 **Burgos, N.**, Cardoso, M.J., Guerreiro, F., McClelland, J., Knopf, A.-C., Punwani, Ourselin, S.: 'CT Synthesis in the Head & Neck Pelvic Regions for Radiotherapy Treatment Planning'. In *IPEM Workshop on MRI Guided Radiotherapy*, 2016 — **ORAL PRESENTATION**
- F.31 Ladefoged, C.N., Law, I., Anazodo, U., Izquierdo-Garcia, D., **Burgos, N.**, Mérida, I., Benoit, D., Juttukonda, M., Cabello, J., Fenchel, M., Jakoby, B., Højgaard, L., Hansen, A.E., Andersen, F.L.: 'A Multi-Method, Multi-Center Study of PET/MRI Brain Attenuation Correction on a Large Cohort of [18F]- FDG Patients: Ready for Clinical Implementation'. In *Annual Meeting of the Radiological Society of North America (RSNA)*, 2016
- F.32 Ladefoged, C.N., Law, I., Anazodo, U., St. Lawrence, K., Izquierdo-Garcia, D., Catana, C., **Burgos, N.**, Cardoso, M.J., Hutton, B., Ourselin, S., Mérida, I., Costes, N., Hammers, A., Benoit, D., Holm, S., Juttukonda, M., An, H., Cabello, J., Lukas, M., Nekolla, S., Ziegler, S., Fenchel, M., Jakoby, B., Casey, M.E., Benzinger, T., Højgaard, L., Hansen, A.E., Andersen, F.L.: 'A Multi-Centre Evaluation of Eleven Clinically Feasible Brain PET/MRI Attenuation Correction Techniques Using a Large Cohort of Patients'. In *2016 IEEE Nuclear Science Symposium Medical Imaging Conference (NSS/MIC)*, 2016
- F.33 Prados Carrasco, F., Cardoso, M.J., **Burgos, N.**, Wheeler-Kingshott, C.A.M., Ourselin, S.: 'NiftyWeb: Web Based Platform for Image Processing on the Cloud'. In *Proceedings of the 24th Scientific Meeting Exhibition of the International Society for Magnetic Resonance in Medicine (ISMRM)*, 2016
- F.34 Sekine, T., **Burgos, N.**, Warnock, G., Huellner, M., Buck, A., Voert, E.E.G.W. ter, Cardoso, M.J., Hutton, B.F., Ourselin, S., Veit-Haibach, P., Delso, G.: 'Multi Atlas-Based Attenuation Correction for Brain FDG- PET Imaging Using a TOF-PET/MR Scanner: Comparison with Clinical Single Atlas- CT-Based Attenuation Correction'. In *Proceedings of the 24th Scientific Meeting Exhibition of the International Society for Magnetic Resonance in Medicine (ISMRM)*, 2016
- F.35 **Burgos, N.**, Cardoso, M.J., Modat, M., Punwani, S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: 'CT Synthesis in the Head & Neck Region for PET/MR Attenuation Correction: An Iterative Multi-Atlas Approach'. *EJNMMI Physics*, 2(1): A31, 2015. doi:10.1186/2197-7364-2-S1-A31 — **RUNNER-UP AWARD FOR BEST ORAL PRESENTATION**
- F.36 Dickson, J.C., Erlandsson, K., Lehmann, M., Modat, M., **Burgos, N.**, Groves, A., Schott, J.: 'Partial Volume Correction of Amyvid FDG PET Data Using the Discrete Iterative Yang Technique'. In *Proceedings of the 28th Annual Congress of the European Association of Nuclear Medicine (EANM)*, S69, Springer, 2015. doi:10.1007/s00259-015-3198-z
- F.37 Guerreiro, F., McClelland, J., **Burgos, N.**, Cardoso, M.J., Dunlop, A., Wong, K., Nill, S., Oelfke, U., Knopf, A.C.: 'Evaluation of Different Approaches to Obtain Synthetic CT Images for a MRI-Only Radiotherapy Workflow'. In *MR in RT Symposium*, 2015
- F.38 Mota, A., Cuplov, V., Schott, J., Hutton, B., Thielemans, K., Drobnyak, I., Dickson, J., Bert, J., **Burgos, N.**, Cardoso, J., Modat, M., Ourselin, S., Erlandsson, K.: 'Establishment of an Open Database of Realistic Simulated Data for Evaluation of Partial Volume Correction Techniques in Brain PET/MR'. *EJNMMI Physics*, 2(1): A44, 2015. doi:10.1186/2197-7364-2-S1-A44
- F.39 **Burgos, N.**, Cardoso, M.J., Thielemans, K., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: 'Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Validation for Brain Study Applications'. *EJNMMI Physics*, 1(1): A52, 2014. doi:10.1186/2197-7364-1-S1-A52
- F.40 Markiewicz, P., Thielemans, K., **Burgos, N.**, Manber, R., Jiao, J., Barnes, A., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S.: 'Image Reconstruction of mMR PET Data Using the Open Source Software STIR'. *EJNMMI Physics*, 1(1): A44, 2014. doi:10.1186/2197-7364-1-S1-A44

G Thesis

- G.1 Burgos, N., Image synthesis for the attenuation correction analysis of PET/MR data. Doctoral thesis, UCL (University College London), 2016, <http://discovery.ucl.ac.uk/1517860>

H Submitted publications and preprints

- H.1 ★ Bottani, S., Thibeau-Sutre, E., Maire, A., Ströer, S., Dormont, D., Colliot, O., **Burgos, N.**: ‘Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation’. Submitted to *MELBA*. [hal-03497645](#)
- H.2 ★ Bottani, S., **Burgos, N.**, Maire, A., Saracino, D., Ströer, S., Dormont, D., and Colliot, O.: ‘Evaluation of MRI-based Machine Learning Approaches for Computer-Aided Diagnosis of Dementia in a Clinical Data Warehouse’. Submitted to *Medical Image Analysis*. [hal-03656136](#)
- H.3 ★ Thibeau-Sutre, E., Collin, S., **Burgos, N.**, and Colliot, O.: ‘Interpretability of Machine Learning Methods Applied to Neuroimaging’. In *Machine Learning for Brain Disorders*, edited by Colliot O., Springer. To be published in 2022. [hal-03615163](#)
- H.4 **Burgos, N.**: ‘Neuroimaging in Machine Learning for Brain Disorders’. In *Machine Learning for Brain Disorders*, edited by Colliot O., Springer. To be published in 2022. [hal-03814787](#)
- H.5 Berenbaum, A., **Burgos, N.**, Thibeau-Sutre, E., Bottani, S., Habert, M.-O., Colliot, O., Kas, A., ‘Classification automatisée des TEP-TDM cérébrales au 18F-FDG par intelligence artificielle : preuve de concept’. Submitted to *Médecine Nucléaire*.
- H.6 Fu, G., Jimenez, G., Loizillon, S., Jurdi, R.E., Chougar, L., Dormont, D., Valabregue, R., **Burgos, N.**, Lehericy, S., Racoceanu, D., Colliot, O.: ‘Fourier Disentangled Multimodal Prior Knowledge Fusion for Red Nucleus Segmentation in Brain MRI’, arXiv, 2022. [doi:10.48550/arXiv.2211.01353](#)

Appendix A

Scopus and PubMed database queries

Scopus query

Scopus query corresponding to Figure 2 displaying the distribution by imaging modality and brain disorder of 1327 articles presenting a study using machine learning:

(TITLE (dementia OR "alzheimer*" OR "AD" OR "Mild Cognitive Impairment" OR "MCI" OR "Posterior cortical atrophy" OR "frontotemporal dementia" OR FTD OR "Frontotemporal lobar degeneration" OR FTLD OR Pick OR "Primary Progressive Aphasia\$" OR PPA OR "semantic dementia") **OR**

TITLE ("Parkinson" OR PD OR "Progressive supranuclear palsy" OR "Corticobasal * Degeneration" OR "Lewy Bod*" OR LBD OR "Multiple System Atrophy" OR MSA) **OR**

TITLE (epilepsy) **OR**

TITLE ("Multiple sclerosis" OR MS) **OR**

TITLE ("vascular dementia" OR stroke OR "Cerebrovascular accident\$" OR "ischemic attack\$" OR "Brain aneurysm\$" OR "Subdural hematoma\$" OR "Epidural hematoma\$" OR "Traumatic brain injur*" OR TBI OR "Intracerebral hemorrhage\$" OR "Concussion\$") **OR**

TITLE ("Brain tumor\$" OR "Brain tumour\$" OR "Glioma\$" OR "Glioblastoma\$" OR "Pseudotumor cerebr*" OR "meningioma\$" OR "astrocytoma\$" OR "medulloblastoma\$" OR "Pituitary adenoma\$" OR "Nerve sheath" OR "Pilocytic Astrocytoma\$" OR "Ependymom\$" OR "Oligodendroglioma\$" OR "Medulloblastoma\$") **OR**

TITLE ("attention deficit hyperactivity disorder" or "ADHD" or "autism" or "ASD" or "asperger") **OR**

TITLE ("psychiatric disorder\$" or "mental disorder\$" or "behavioural disorder\$" or "anxiety" or "depression" or "depressive disorder\$" or "MDD" or "depressive episode" or "paranoi*" or "mania" or "hypomania" or "bipolar" or "personality disorder\$" or "impulse disorder\$" or "identity disorder\$" or "mood disorder\$" or "*phobia\$" or "panic disorder\$" or "adjustment disorder\$" or "stress disorder\$" or "stress reaction" or "eating disorder\$" or "anorexia nervosa" or "bulimia nervosa" or "sleep disorder\$" or "dissociative disorder\$" or "conversion disorder\$" or "affective disorder\$" or "psychosis" or "psychotic" or "schizophrenia" or "delusion*" or "schizoaffective" or "schizophreniform" or "schizotypal" or "neurotic" or "somatoform" or "somatization" or "addiction" or "substance" or "post-traumatic stress disorder" or "PTSD" or "obsessive compulsive disorder" or "OCD" or "dyslexia" or "language disorder\$" or "conduct disorder\$"))

AND

(TITLE (magnetic OR MRI OR MRS OR perfusion OR DCE OR DSC OR ASL OR MRA OR SWI OR QSM OR "voxel based morphometry") **OR**

TITLE (fMRI OR "functional MRI" OR BOLD OR "resting state") **OR**

TITLE ("diffusion weighted" OR DWI OR "diffusion tensor" OR DTI OR tractography) **OR**

TITLE (positron OR PET OR FDG OR fluorodeoxuglucose OR pittsburgh OR PiB OR AV45 OR florbetapir OR florbetaben OR flutemetamol OR AV1451 OR flortaucipir OR DOPA OR TSPO OR PK11195 OR PBR28 OR DAA1106 OR FEPPA OR DPA-714 OR CLINME OR FET OR MET OR FLT) **OR**

TITLE ("Single-photon emission computed tomography" OR SPECT OR HMPAO OR ECD OR "FP-CIT" OR "99mTC" OR Ioflupane) **OR**

TITLE ("computed tomography" OR CT) **OR**

TITLE ("X*ray") **OR**

TITLE ("angiography" OR "arteriography"))

AND

(TITLE ("Deep learning" OR "neural network\$" OR "convolutional network\$" OR "CNN\$" OR "RNN\$" OR "LSTM\$" OR "bayesian network\$" OR "adversarial" OR "GAN\$" OR "cGAN\$" OR "cycle\$GAN\$" OR "U*net" OR "auto*encoder" OR "perceptron\$") **OR**

TITLE ("Matrix completion" OR "Support vector machine\$" OR "linear mixed\$effect\$" OR "logistic regression\$" OR "Random Forest\$" OR "kernel classifier\$" OR "kernel" OR "decision tree\$" OR "least-squares" OR "Naive Bayes" OR "Linear discriminant" OR "K\$nearest neighbor\$") **OR**

TITLE (" Machine learning" OR "pattern recognition" OR "pattern classification" OR "classifier" OR "algorithm" OR " classification") **OR**

TITLE ("radiomic\$"))

AND (LIMIT-TO (SRCTYPE , "j"))

AND (LIMIT-TO (LANGUAGE , "English"))

PubMed query

PubMed query corresponding to Figure 3 displaying the number of articles presenting machine learning and deep learning approaches for the computer-aided diagnosis of Alzheimer's disease published over the years according.

Machine learning query

(alzheimer [Title] OR "Cognitive Impairment" [Title])

AND ("classif*" [Title] OR "diagnos*" [Title] OR "identif*" [Title] OR "detect*" [Title] OR "recogni*" [Title] OR "prognos*" [Title] OR "predict*" [Title])

AND (mri OR "Magnetic Resonance Imaging" OR neuroimaging OR (brain AND imaging) OR positron OR PET)

AND ("Matrix completion" [Title/Abstract] OR "Support vector machine\$" [Title/Abstract] OR "linear mixed-effect\$" [Title/Abstract] OR "Machine Learning" [Title/Abstract] OR "logistic regression" [Title/Abstract] OR "Random Forest" [Title/Abstract] OR "kernel\$" [Title/Abstract] OR "decision tree\$" [Title/Abstract] OR "least-squares" [Title/Abstract])

NOT ("cnn\$" [Title] OR "Convolutional Network\$" [Title] OR "Convolutional neural Network\$" [Title] OR "Deep Learning" [Title] OR "Neural Network\$" [Title] OR "autoencoder\$" [Title] OR gan [Title] OR adversarial [Title] OR "deep belief network\$" [Title])

Deep learning query

(alzheimer [Title] OR "Cognitive Impairment" [Title])

AND ("classif*" [Title] OR "diagnos*" [Title] OR "identif*" [Title] OR "detect*" [Title] OR "recogni*" [Title] OR "prognos*" [Title] OR "predict*" [Title])

AND (mri OR "Magnetic Resonance Imaging" OR neuroimaging OR (brain AND imaging) OR positron OR PET)

AND ("cnn\$" [Title/Abstract] OR "Convolutional Network\$" [Title/Abstract] OR "Convolutional neural Network\$" [Title/Abstract] OR "Deep Learning" [Title/Abstract] OR "Neural Network\$" [Title/Abstract] OR "autoencoder\$" [Title/Abstract] OR gan [Title/Abstract] OR adversarial [Title/Abstract] OR "deep belief network\$" [Title/Abstract])

NOT ("Matrix completion" [Title] OR "Support vector machine" [Title] OR "linear mixed-effect" [Title] OR "Machine Learning" [Title] OR "logistic regression" [Title] OR "Random Forest" [Title] OR "kernel" [Title] OR "decision tree" [Title] OR "decision trees" [Title] OR "least-squares" [Title])

Appendix B

Data access

ADNI

Data collection and sharing for this work was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

NIFD

Data used in preparation of this work was obtained from the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI) database (<http://4rtni-ftldni.ini.usc.edu/>). The investigators at NIFD/FTLDNI contributed to the design and implementation of FTLDNI and/or provided data, but did not participate in analysis or writing of this report (unless otherwise listed).

AIBL

Data used in the preparation of this work was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

OASIS

The OASIS Cross-Sectional project (Principal Investigators: D. Marcus, R. Buckner, J. Csernansky J. Morris) was supported by the following grants: P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, and U24 RR021382.

Appendix C

APPRIMAGE Study Group

Olivier Colliot, Ninon Burgos, Simona Bottani, Sophie Loizillon ¹

Didier Dormont ^{1,2}, Samia Si Smail Belkacem, Sebastian Ströer ²

Nathalie Boddaert ³

Farida Benoudiba, Ghaida Nasser, Claire Ancelet, Laurent Spelle ⁴

Hubert Ducou-Le-Pointe⁵

Catherine Adamsbaum⁶

Marianne Alison⁷

Emmanuel Houdart⁸

Robert Carlier ^{9,17}

Myriam Edjlali⁹

Betty Marro^{10,11}

Lionel Arrive¹⁰

Alain Luciani¹²

Antoine Khalil¹³

Elisabeth Dion¹⁴

Laurence Rocher¹⁵

Pierre-Yves Brillet¹⁶

Paul Legmann, Jean-Luc Drape ¹⁸

Aurélien Maire, Stéphane Bréant, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel, Cyrina Saussol, Rafael Gozlan ¹⁹

Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret ²⁰

¹ Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, Inria, Aramis project-team, F-75013, Paris, France

² AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

³ AP-HP, Hôpital Necker, Department of Radiology, F-75015, Paris, France

⁴ AP-HP, Hôpital Bicêtre, Department of Radiology, F-94270, Le Kremlin-Bicêtre, France

⁵ AP-HP, Hôpital Armand-Trousseau, Department of Radiology, F-75012, Paris, France

⁶ AP-HP, Hôpital Bicêtre, Department of Pediatric Radiology, F-94270, Le Kremlin-Bicêtre, France

⁷ AP-HP, Hôpital Robert-Debré, Department of Radiology, F-75019, Paris, France

⁸ AP-HP, Hôpital Lariboisière, Department of Neuroradiology, F-75010, Paris, France

⁹ AP-HP, Hôpital Raymond-Poincaré, Department of Radiology, F-92380, Garches, France

¹⁰ AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75012, Paris, France

- ¹¹ AP-HP, Hôpital Tenon, Department of Radiology, F-75020, Paris, France
- ¹² AP-HP, Hôpital Henri-Mondor, Department of Radiology, F-94000, Créteil, France
- ¹³ AP-HP, Hôpital Bichat, Department of Radiology, F-75018, Paris, France
- ¹⁴ AP-HP, Hôpital Hôtel-Dieu, Department of Radiology, F-75004, Paris, France
- ¹⁵ AP-HP, Hôpital Antoine-Béclère, Department of Radiology, F-92140, Clamart, France
- ¹⁶ AP-HP, Hôpital Avicenne, Department of Radiology, F-93000, Bobigny, France
- ¹⁷ AP-HP, Hôpital Ambroise Paré, Department of Radiology, F-92100 104, Boulogne-Billancourt, France
- ¹⁸ AP-HP, Hôpital Cochin, Department of Radiology, F-75014, Paris, France
- ¹⁹ AP-HP, WIND department, F-75012, Paris, France
- ²⁰ AP-HP, Unité de Recherche Clinique, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

Bibliography

- Abdulkadir, A., B. Mortamet, P. Vemuri, C. R. Jack, G. Krueger, and S. Klöppel (2011). “Effects of hardware heterogeneity on the performance of SVM Alzheimer’s disease classifier”. In: *NeuroImage* 58.3, pp. 785–792.
- Abraham, A., F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux (2014). “Machine learning for neuroimaging with scikit-learn”. In: *Frontiers in neuroinformatics* 8, p. 14.
- Adebayo, J., J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim (2018). “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems*, pp. 9505–9515.
- Aderghal, K., A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline (2018). “Classification of Alzheimer Disease on Imaging Modalities with Deep CNNs Using Cross-Modal Transfer Learning”. In: *IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 345–350.
- Aderghal, K., J. Benois-Pineau, K. Afdel, and C. Gwenaëlle (2017a). “FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+ ϵ projections”. In: *15th International Workshop on Content-Based Multimedia Indexing*. ACM, p. 34.
- Aderghal, K., M. Boissenin, J. Benois-Pineau, G. Catheline, and K. Afdel (2017b). “Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+ ϵ Study on ADNI”. In: *International Conference on Multimedia Modeling*, pp. 690–701.
- Aguilar, C., E. Westman, J.-S. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Kloszewska, H. Soininen, S. Lovestone, C. Spenger, A. Simmons, and L.-O. Wahlund (2013). “Different multivariate techniques for automated classification of MRI data in Alzheimer’s disease and mild cognitive impairment”. In: *Psychiatry Research* 212.2, pp. 89–98.
- Alfaro-Almagro, F., M. Jenkinson, N. K. Bangerter, J. L. R. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, D. Vidauvre, M. Webster, P. McCarthy, C. Rorden, A. Daducci, D. C. Alexander, H. Zhang, I. Dragonu, P. M. Matthews, K. L. Miller, and S. M. Smith (2018). “Image Processing and Quality Control for the First 10,000 Brain Imaging Datasets from UK Biobank”. In: *Neuroimage* 166, pp. 400–424.
- Allen, G. I. et al. (2016). “Crowdsourced estimation of cognitive decline and resilience in Alzheimer’s disease”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 12.6, pp. 645–653.
- Alzheimer Europe (2013). *Dementia in Europe Yearbook 2013*. Tech. rep.
- Alzheimer’s Disease International (2015). *World Alzheimer Report 2015*. Tech. rep.
- Ambrose, J. (1973). “Computerized Transverse Axial Scanning (Tomography): Part 2. Clinical Application”. In: *The British Journal of Radiology* 46.552, pp. 1023–1047.

- Andermatt, S., A. Horváth, S. Pezold, and P. Cattin (2019). “Pathology Segmentation Using Distributional Differences to Images of Healthy Origin”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 228–238.
- Arbabshirani, M. R., S. Plis, J. Sui, and V. D. Calhoun (2017). “Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls”. In: *NeuroImage* 145.Pt B, pp. 137–165.
- Ashburner, J. (2007). “A fast diffeomorphic image registration algorithm”. In: *NeuroImage* 38.1, pp. 95–113.
- (2012). “SPM: A History”. In: *NeuroImage. 20 years of fMRI* 62.2, pp. 791–800.
- Ashburner, J. and K. J. Friston (2005). “Unified Segmentation”. In: *NeuroImage* 26.3, pp. 839–851.
- Avants, B. B., C. L. Epstein, M. Grossman, and J. C. Gee (2008). “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain”. In: *Medical Image Analysis* 12.1, pp. 26–41.
- Avants, B. B., N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee (2014). “The Insight ToolKit image registration framework”. In: *Frontiers in Neuroinformatics* 8, p. 44.
- Bäckström, K., M. Nazari, I. Y.-H. Gu, and A. S. Jakola (2018). “An Efficient 3D Deep Convolutional Network for Alzheimer’s Disease Diagnosis Using MR Images”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 149–153.
- Barbano, R., S. Arridge, B. Jin, and R. Tanno (2022). “Uncertainty quantification in medical image synthesis”. In: *Biomedical Image Synthesis and Simulation*. Elsevier, pp. 601–641.
- Basaia, S., F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, and Alzheimer’s Disease Neuroimaging Initiative (2019). “Automated classification of Alzheimer’s disease and mild cognitive impairment using a single MRI and deep neural networks”. In: *Neuroimage Clin* 21, p. 101645.
- Baur, C., S. Denner, B. Wiestler, N. Navab, and S. Albarqouni (2021). “Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study”. In: *Medical Image Analysis*, p. 101952.
- Baur, C., B. Wiestler, S. Albarqouni, and N. Navab (2019). “Deep autoencoding models for unsupervised anomaly segmentation in brain MR images”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 161–169.
- Baydargil, H. B., J.-S. Park, and D.-Y. Kang (2021). “Anomaly analysis of Alzheimer’s disease in PET images using an unsupervised adversarial deep learning model”. In: *Applied Sciences* 11.5, p. 2187.
- Becquerel, H. (1903). *Recherches Sur Une Propriété Nouvelle de La Matière: Activité Radiante Spontanée Ou Radioactivité de La Matière*. Mémoires de l’Académie Des Sciences de l’Institut de France. L’Institut de France.
- Bengio, Y. and Y. Grandvalet (2004). “No Unbiased Estimator of the Variance of K-Fold Cross-Validation”. In: *J. Mach. Learn. Res.* 5, pp. 1089–1105.
- Benou, A., R. Veksler, A. Friedman, and T. R. Raviv (2017). “Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences”. In: *Medical Image Analysis* 42, pp. 145–159.

- Bensaïdane, M. R., J.-M. Beauregard, S. Poulin, F.-A. Buteau, J. Guimond, D. Bergeron, L. Verret, M.-P. Fortin, M. Houde, and R. W. Bouchard (2016). “Clinical Utility of Amyloid PET Imaging in the Differential Diagnosis of Atypical Dementias and Its Impact on Caregivers”. In: *Journal of Alzheimer’s Disease* 52.4, pp. 1251–1262.
- Bergstra, J. and Y. Bengio (2012). “Random Search for Hyper-Parameter Optimization”. In: *Journal of Machine Learning Research* 13.Feb, pp. 281–305.
- Bi, L., J. Kim, A. Kumar, D. Feng, and M. Fulham (2017). “Synthesis of Positron Emission Tomography (PET) Images via Multi-channel Generative Adversarial Networks (GANs)”. In: *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pp. 43–51.
- Bloch, F. (1946). “Nuclear Induction”. In: *Physical Review* 70.7-8, pp. 460–474.
- Bône, A., S. Ammari, J.-P. Lamarque, M. Elhaik, É. Chouzenoux, F. Nicolas, P. Robert, C. Balleyguier, N. Lassau, and M.-M. Rohé (2021). “Contrast-enhanced brain MRI synthesis with deep learning: key input modalities and asymptotic performance”. In: *2021 IEEE ISBI*.
- Bottani, S., N. Burgos, A. Maire, D. Saracino, S. Ströer, D. Dormont, and O. Colliot (2022a). “Evaluation of MRI-based Machine Learning Approaches for Computer-Aided Diagnosis of Dementia in a Clinical Data Warehouse”. In: *Preprint*.
- Bottani, S., N. Burgos, A. Maire, A. Wild, S. Ströer, D. Dormont, and O. Colliot (2022b). “Automatic Quality Control of Brain T1-weighted Magnetic Resonance Images for a Clinical Data Warehouse”. In: *Medical Image Analysis* 75, p. 102219.
- Bottani, S., E. Thibeau-Sutre, A. Maire, S. Ströer, D. Dormont, O. Colliot, and N. Burgos (2021). “Homogenization of Brain MRI from a Clinical Data Warehouse Using Contrast-Enhanced to Non-Contrast-Enhanced Image Translation”. In: *Preprint*.
- Bottani, S., E. Thibeau-Sutre, A. Maire, S. Ströer, D. Dormont, O. Colliot, and N. Burgos (2022c). “Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-Net derived models”. In: *SPIE Medical Imaging 2022*.
- Bottani, S., E. Thibeau-Sutre, A. Maire, S. Ströer, D. Dormont, O. Colliot, and N. Burgos (2022d). “Homogenization of Brain MRI from a Clinical Data Warehouse Using Contrast-Enhanced to Non-Contrast-Enhanced Image Translation with U-Net Derived Models”. In: *SPIE Medical Imaging*. Vol. 12032. SPIE, pp. 576–582.
- Bouts, M. J., J. van der Grond, M. W. Vernooij, M. Koini, T. M. Schouten, F. de Vos, R. A. Feis, L. G. Cremers, A. Lechner, R. Schmidt, et al. (2019). “Detection of mild cognitive impairment in a community-dwelling population using quantitative, multiparametric MRI-based classification”. In: *Human Brain Mapping* 40.9, pp. 2711–2722.
- Bouwman, F., S. Orini, F. Gandolfo, D. Altomare, C. Festari, F. Agosta, J. Arbizu, A. Drzezga, P. Nestor, F. Nobili, Z. Walker, S. Morbelli, and M. Boccardi (2018). “Diagnostic utility of FDG-PET in the differential diagnosis between different forms of primary progressive aphasia”. In: *Eur. J. Nucl. Med. Mol. Imaging* 45.9, pp. 1526–1533.
- Bowles, C., C. Qin, R. Guerrero, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert (2017). “Brain lesion segmentation through image synthesis and outlier detection”. In: *NeuroImage: Clinical* 16, pp. 643–658.

- Brett, M., M. Hanke, C. Markiewicz, Marc-Alexandre Côté, P. McCarthy, C. Cheng, Y. Halchenko, S. Ghosh, D. Wassermann, S. Gerhard, E. Larson, G. R. Lee, E. Kastman, C. M. A. Rokem, F. C. Morency, Moloney, M. Cottaar, J. Millman, R. Markello, Jaeillepp, A. Gramfort, R. D. Vincent, J. J. V. D. Bosch, K. Subramaniam, P. R. Raamana, M. Goncalves, N. Nichols, Embaker, and Basile (2019). *Nipy/Nibabel: 2.3.3*. Zenodo.
- Bron, E. E., S. Klein, J. M. Papma, L. C. Jiskoot, V. Venkatraghavan, J. Linders, P. Aalten, P. P. De Deyn, G. J. Biessels, J. A. Claassen, et al. (2021). “Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer’s disease”. In: *NeuroImage: Clinical* 31, p. 102712.
- Bron, E. E., M. Smits, W. M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. E. Steketee, C. Méndez Orellana, R. Meijboom, M. Pinto, J. R. Meireles, C. Garrett, A. J. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cárdenas-Peña, A. M. Álvarez Meza, C. V. Dolph, K. M. Iftekharuddin, S. F. Eskildsen, P. Coupé, V. S. Fonov, K. Franke, C. Gaser, C. Ledig, R. Guerrero, T. Tong, K. R. Gray, E. Moradi, J. Tohka, A. Routier, S. Durrleman, A. Sarica, G. Di Fatta, F. Sensi, A. Chincarini, G. M. Smith, Z. V. Stoyanov, L. Sørensen, M. Nielsen, S. Tangaro, P. Inglese, C. Wachinger, M. Reuter, J. C. van Swieten, W. J. Niessen, and S. Klein (2015). “Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge”. In: *NeuroImage* 111, pp. 562–579.
- Burgos, N., M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott, J. S. Duncan, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin (2014). “Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Application to Brain Studies”. In: *IEEE Transactions on Medical Imaging* 33.12, pp. 2332–2341.
- Burgos, N., S. Bottani, J. Faouzi, E. Thibeau-Sutre, and O. Colliot (2021a). “Deep Learning for Brain Disorders: From Data Processing to Disease Treatment”. In: *Briefings in Bioinformatics* 22.2, pp. 1560–1576.
- Burgos, N., M. J. Cardoso, A. F. Mendelson, J. M. Schott, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin (2015). “Subject-Specific Models for the Analysis of Pathological FDG PET Data”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, pp. 651–658.
- Burgos, N., M. J. Cardoso, J. Samper-González, M.-O. Habert, S. Durrleman, S. Ourselin, and O. Colliot (2021b). “Anomaly Detection for the Individual Analysis of Brain PET Images”. In: *Journal of Medical Imaging* 8.2, p. 024003.
- Burgos, N. and O. Colliot (2020). “Machine Learning for Classification and Prediction of Brain Diseases: Recent Advances and Upcoming Challenges”. In: *Current Opinion in Neurology* 33.4, pp. 439–450.
- Burgos, N., F. Guerreiro, J. McClelland, B. Presles, M. Modat, S. Nill, D. Dearnaley, N. deSouza, U. Oelfke, A.-C. Knopf, S. Ourselin, and M. J. Cardoso (2017a). “Iterative Framework for the Joint Segmentation and CT Synthesis of MR Images: Application to MRI-only Radiotherapy Treatment Planning”. In: *Physics in Medicine and Biology* 62.11, p. 4237.

- Burgos, N., J. Samper-González, A. Bertrand, M.-O. Habert, S. Ourselin, S. Durrleman, M. J. Cardoso, and O. Colliot (2017b). “Individual Analysis of Molecular Brain Imaging Data through Automatic Identification of Abnormality Patterns”. In: *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. Lecture Notes in Computer Science. Springer, pp. 13–22.
- Cabral, C., P. M. Morgado, D. Campos Costa, and M. Silveira (2015). “Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages”. In: *Computers in Biology and Medicine* 58, pp. 101–109.
- Cachier, P., E. Bardinet, D. Dormont, X. Pennec, and N. Ayache (2003). “Iconic feature based nonrigid registration: the PASHA algorithm”. In: *Comput. Vision Image Understanding* 89.2-3, pp. 272–298.
- Cackowski, S., E. L. Barbier, M. Dojat, and T. Christen (2021). “ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization”. In: *arXiv preprint arXiv:2109.06756*.
- Cardoso, M., M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin (2015). “Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion”. In: *IEEE Trans. Med. Imaging* 34.9, pp. 1976–1988.
- Cerami, C., A. Dodich, L. Greco, S. Iannaccone, G. Magnani, A. Marcone, E. Pelagallo, R. Santangelo, S. F. Cappa, and D. Perani (2016a). “The Role of Single-Subject Brain Metabolic Patterns in the Early Differential Diagnosis of Primary Progressive Aphasia and in Prediction of Progression to Dementia”. In: *J. Alzheimers Dis.* 55.1, pp. 183–197.
- Cerami, C., A. Dodich, G. Lettieri, S. Iannaccone, G. Magnani, A. Marcone, L. Gianolli, S. F. Cappa, and D. Perani (2016b). “Different FDG-PET metabolic patterns at single-subject level in the behavioral variant of fronto-temporal dementia”. In: *Cortex* 83, pp. 101–112.
- Chagué, P., B. Marro, S. Fadili, M. Houot, A. Morin, J. Samper-González, P. Beunon, L. Arrivé, D. Dormont, B. Dubois, et al. (2021). “Radiological classification of dementia from anatomical MRI assisted by machine learning-derived maps”. In: *Journal of Neuroradiology* 48.6, pp. 412–418.
- Chen, X. and E. Konukoglu (2022). “Unsupervised abnormality detection in medical images with deep generative methods”. In: *Biomedical Image Synthesis and Simulation*. Elsevier, pp. 303–324.
- Chen, X., S. You, K. C. Tezcan, and E. Konukoglu (2020). “Unsupervised Lesion Detection via Image Restoration with a Normative Prior”. In: *Medical Image Analysis* 64, p. 101713.
- Chen, Y., F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, and D. Li (2018). “Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 91–99.
- Cheng, D and M Liu (2017). “CNNs based multi-modality classification for AD diagnosis”. In: *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5.
- Choi, H., S. Ha, H. Kang, H. Lee, and D. S. Lee (2019). “Deep Learning Only by Normal Brain PET Identify Unheralded Brain Anomalies”. In: *EBioMedicine* 43, pp. 447–453.

- Choi, H. and D. S. Lee (2018). “Generation of structural MR images from amyloid PET: application to MR-less quantification”. In: *Journal of Nuclear Medicine* 59.7, pp. 1111–1117.
- Chu, C., A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin (2012). “Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images”. In: *NeuroImage* 60.1, pp. 59–70.
- Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger (2016). “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer, pp. 424–432.
- Cosgrove, K. P., C. M. Mazure, and J. K. Staley (2007). “Evolving Knowledge of Sex Differences in Brain Structure, Function, and Chemistry”. In: *Biol. Psychiatry* 62.8, pp. 847–855.
- Couvy-Duchesne, B., J. Faouzi, B. Martin, E. Thibeau-Sutre, A. Wild, M. Ansart, S. Durleman, D. Dormont, N. Burgos, and O. Colliot (2020). “Ensemble Learning of Convolutional Neural Network, Support Vector Machine, and Best Linear Unbiased Predictor for Brain Age Prediction: ARAMIS Contribution to the Predictive Analytics Competition 2019 Challenge”. In: *Frontiers in Psychiatry* 11.
- Crane, M. (2018). “Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results”. In: *Transactions of the Association for Computational Linguistics* 6, pp. 241–252.
- Cuingnet, R., E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot (2011). “Automatic classification of patients with Alzheimer’s disease from structural MRI: a comparison of ten methods using the ADNI database”. In: *NeuroImage* 56.2, pp. 766–781.
- Curiati, P. K., J. H. Tamashiro-Duran, F. L. S. Duran, C. A. Buchpiguel, P. Squarzoni, D. C. Romano, H. Vallada, P. R. Menezes, M. Scazufca, G. F. Busatto, and T. C. T. F. Alves (2011). “Age-Related Metabolic Profiles in Cognitively Healthy Elders: Results from a Voxel-Based [18F]Fluorodeoxyglucose-Positron-Emission Tomography Study with Partial Volume Effects Correction”. In: *Am. J. Neuroradiology* 32.3, pp. 560–565.
- Daniel, C. and E. Salamanca (2020). “Hospital Databases”. In: *Healthcare and Artificial Intelligence*. Springer, pp. 57–67.
- De Tiege, X., S. Goldman, S. Laureys, D. Verheulpen, C. Chiron, C. Wetzburger, P. Paquier, D. Chaigne, N. Poznanski, I. Jambaque, E. Hirsch, O. Dulac, and P. Van Bogaert (2004). “Regional cerebral glucose metabolism in epilepsies with continuous spikes and waves during sleep”. In: *Neurology* 63.5, pp. 853–857.
- Dewey, B. E., C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, et al. (2019). “DeepHarmony: a deep learning approach to contrast harmonization across scanner changes”. In: *Magnetic Resonance Imaging* 64, pp. 160–170.

- Diehl-Schmid, J., T. Grimmer, A. Drzezga, S. Bornschein, M. Riemenschneider, H. Förstl, M. Schwaiger, and A. Kurz (2007). “Decline of cerebral glucose metabolism in frontotemporal dementia: a longitudinal 18F-FDG-PET-study”. In: *Neurobiol. Aging* 28.1, pp. 42–50.
- Ding, Z., G. Fleishman, X. Yang, P. Thompson, R. Kwitt, and M. Niethammer (2019). “Fast Predictive Simple Geodesic Regression”. In: *Med. Image Anal.* 56, pp. 193–209.
- Drzezga, A., H. Barthel, S. Minoshima, and O. Sabri (2014). “Potential Clinical Applications of PET/MR Imaging in Neurodegenerative Diseases”. In: *Journal of Nuclear Medicine* 55.Supplement_2, 47S–55S.
- Drzezga, A., T. Grimmer, M. Riemenschneider, N. Lautenschlager, H. Siebner, P. Alexopoulos, S. Minoshima, M. Schwaiger, and A. Kurz (2005). “Prediction of Individual Clinical Outcome in MCI by Means of Genetic Assessment and ¹⁸F-FDG PET”. In: *J. Nucl. Med.* 46.10, pp. 1625–1632.
- Du, J., L. Wang, Y. Liu, Z. Zhou, Z. He, and Y. Jia (2020). “Brain mri super-resolution using 3d dilated convolutional encoder–decoder network”. In: *IEEE Access* 8, pp. 18938–18950.
- Dukart, J., K. Mueller, H. Barthel, A. Villringer, O. Sabri, and M. L. Schroeter (2013). “Meta-analysis based SVM classification enables accurate detection of Alzheimer’s disease across different clinical centers using FDG-PET and MRI”. In: *Psychiatry Research: Neuroimaging* 212.3, pp. 230–236.
- Ebrahimighahnavieh, M. A., S. Luo, and R. Chiong (2020). “Deep learning to detect Alzheimer’s disease from neuroimaging: A systematic literature review”. In: *Computer Methods and Programs in Biomedicine* 187, p. 105242.
- Ellis, K. A., A. I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N. T. Lautenschlager, N. Lenzo, R. N. Martins, P. Maruff, C. Masters, A. Milner, K. Pike, C. Rowe, G. Savage, C. Szoeki, K. Taddei, V. Villemagne, M. Woodward, D. Ames, and AIBL Research Group (2009). “The Australian Imaging, Biomarkers and Lifestyle (AIBL) Study of Aging: Methodology and Baseline Characteristics of 1112 Individuals Recruited for a Longitudinal Study of Alzheimer’s Disease”. In: *International Psychogeriatrics* 21.4, pp. 672–687.
- Ellis, K. A., C. C. Rowe, V. L. Villemagne, R. N. Martins, C. L. Masters, O. Salvado, C. Szoeki, D. Ames, and AIBL research group (2010). “Addressing Population Aging and Alzheimer’s Disease through the Australian Imaging Biomarkers and Lifestyle Study: Collaboration with the Alzheimer’s Disease Neuroimaging Initiative”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 6.3, pp. 291–296.
- Eskildsen, S. F., P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins (2013). “Prediction of Alzheimer’s disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning”. In: *NeuroImage* 65, pp. 511–521.
- Esteban, O., D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski (2017). “MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites”. In: *PLOS One* 12.9, e0184661.

- Ewers, M., G. B. Frisoni, S. J. Teipel, L. T. Grinberg, E. Amaro, H. Heinsen, P. M. Thompson, and H. Hampel (2011a). “Staging Alzheimer’s Disease Progression with Multimodality Neuroimaging”. In: *Progress in Neurobiology*. Biological Markers for Neurodegenerative Diseases 95.4, pp. 535–546.
- Ewers, M., R. A. Sperling, W. E. Klunk, M. W. Weiner, and H. Hampel (2011b). “Neuroimaging markers for the prediction and early diagnosis of Alzheimer’s disease dementia”. In: *Trends in Neurosciences* 34.8, pp. 430–442.
- Falahati, F., E. Westman, and A. Simmons (2014). “Multivariate data analysis and machine learning in Alzheimer’s disease with a focus on structural magnetic resonance imaging”. In: *Journal of Alzheimer’s disease* 41.3, pp. 685–708.
- Fan, J., X. Cao, P.-T. Yap, and D. Shen (2019). “BIRNet: Brain image registration using dual-supervised fully convolutional networks”. In: *Med. Image Anal.* 54, pp. 193–206.
- Fan, Y., N. Batmanghelich, C. M. Clark, C. Davatzikos, A. D. N. Initiative, et al. (2008). “Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline”. In: *Neuroimage* 39.4, pp. 1731–1743.
- Fathi, S., M. Ahmadi, and A. Dehnad (2022). “Early diagnosis of Alzheimer’s disease based on deep learning: A systematic review”. In: *Computers in Biology and Medicine*, p. 105634.
- Fischl, B. (2012). “FreeSurfer”. In: *NeuroImage*. 20 Years of fMRI 62.2, pp. 774–781.
- Fong, R. C. and A. Vedaldi (2017). “Interpretable Explanations of Black Boxes by Meaningful Perturbation”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457.
- Fonov, V. S., A. C. Evans, R. C. McKinstry, C. R. Almlı, and D. L. Collins (2009). “Unbiased nonlinear average age-appropriate brain templates from birth to adulthood”. In: *Neuroimage* Supplement 1.47, S102.
- Fonov, V., A. C. Evans, K. Botteron, C. R. Almlı, R. C. McKinstry, D. L. Collins, and Brain Development Cooperative Group (2011). “Unbiased average age-appropriate atlases for pediatric studies”. In: *Neuroimage* 54.1, pp. 313–327.
- Forsberg, A., H. Engler, O. Almkvist, G. Blomquist, G. Hagman, A. Wall, A. Ringheim, B. Långström, and A. Nordberg (2008). “PET imaging of amyloid deposition in patients with mild cognitive impairment”. In: *Neurobiol. Aging* 29.10, pp. 1456–1465.
- Franke, K., G. Ziegler, S. Klöppel, and C. Gaser (2010). “Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters”. In: *NeuroImage* 50.3, pp. 883–892.
- Friston, K. J., A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak (1994). “Statistical parametric maps in functional imaging: A general linear approach”. In: *Hum. Brain Mapp.* 2.4, pp. 189–210.
- Frizzell, T. O., M. Glashutter, C. C. Liu, A. Zeng, D. Pan, S. G. Hajra, R. C. D’Arcy, and X. Song (2022). “Artificial intelligence in brain MRI analysis of Alzheimer’s disease over the past 12 years: A systematic review”. In: *Ageing Research Reviews*, p. 101614.

- Gal, Y. and Z. Ghahramani (2016). “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *International Conference on Machine Learning*.
- Gerardin, E., G. Chételat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, M. Niethammer, B. Dubois, S. Lehéricy, L. Garnero, et al. (2009). “Multidimensional classification of hippocampal shape features discriminates Alzheimer’s disease and mild cognitive impairment from normal aging”. In: *Neuroimage* 47.4, pp. 1476–1486.
- Gilmore, A., N. Buser, and J. L. Hanson (2019). “Variations in structural MRI quality impact measures of brain anatomy: Relations with age and other sociodemographic variables”. In: *Biorxiv*, p. 581876.
- Gong, K., J. Yang, K. Kim, G. El Fakhri, Y. Seo, and Q. Li (2018). “Attenuation correction for brain PET imaging using deep neural network based on Dixon and ZTE MR images”. In: *Physics in Medicine & Biology* 63.12, p. 125011.
- Goodman, R. A., K. A. Lochner, M. Thambisetty, T. S. Wingo, S. F. Posner, and S. M. Ling (2017). “Prevalence of Dementia Subtypes in United States Medicare Fee-for-Service Beneficiaries, 2011–2013”. In: *Alzheimer’s & Dementia* 13.1, pp. 28–37.
- Gorgolewski, K., C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh (2011). “Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python”. In: *Frontiers in Neuroinformatics* 5.
- Gorgolewski, K. J., T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, and R. A. Pollock (2016). “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments”. In: *Scientific Data* 3, p. 160044.
- Gousias, I. S., D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers (2008). “Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest”. In: *NeuroImage* 40.2, pp. 672–684.
- Grimmer, T., J. Diehl, A. Drzezga, H. Förstl, and A. Kurz (2004). “Region-Specific Decline of Cerebral Glucose Metabolism in Patients with Frontotemporal Dementia: A Prospective 18F-FDG-PET Study”. In: *Dement. Geriatr. Cogn. Disord.* 18.1, pp. 32–36.
- Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger (2017). “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, pp. 1321–1330.
- Gómez-Sancho, M., J. Tohka, and V. Gómez-Verdejo (2018). “Comparison of feature representations in MRI-based MCI-to-AD conversion prediction”. In: *Magnetic Resonance Imaging* 50, pp. 84–95.
- Haller, S., K. O. Lovblad, and P. Giannakopoulos (2011). “Principles of classification analyses in mild cognitive impairment (MCI) and Alzheimer disease”. In: *Journal of Alzheimer’s disease* 26 Suppl 3, pp. 389–394.
- Hammers, A., R. Allom, M. J. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks, and J. S. Duncan (2003). “Three-dimensional maximum probability atlas

- of the human brain, with particular reference to the temporal lobe”. In: *Human Brain Mapping* 19.4, pp. 224–247.
- Han, X. (2017). “MR-Based Synthetic CT Generation Using a Deep Convolutional Neural Network Method”. In: *Medical Physics* 44.4, pp. 1408–1419.
- Hashimoto, F., H. Ohba, K. Ote, A. Teramoto, and H. Tsukada (2019). “Dynamic PET image denoising using deep convolutional neural networks without prior training datasets”. In: *IEEE Access* 7, pp. 96594–96603.
- He, K., X. Zhang, S. Ren, and J. Sun (2015). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, pp. 1026–1034.
- (2016a). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- (2016b). “Identity mappings in deep residual networks”. In: *European conference on computer vision*. Springer, pp. 630–645.
- Heckemann, R. A., A. Hammers, D. Rueckert, R. I. Aviv, C. J. Harvey, and J. V. Hajnal (2008). “Automatic volumetry on MR brain images can support diagnostic decision making”. In: *BMC Medical Imaging* 8.1, pp. 1–6.
- Heiss, W.-D. and S. Zimmermann-Meinzingen (2012). “PET imaging in the differential diagnosis of vascular dementia”. In: *J. Neurol. Sci.* 322.1-2, pp. 268–273.
- Herholz, K. (1995). “FDG PET and differential diagnosis of dementia”. In: *Alzheimer Dis. Assoc. Disord.* 9.1, pp. 6–16.
- Herholz, K. (2014). “The Role of PET Quantification in Neurological Imaging: FDG and Amyloid Imaging in Dementia”. In: *Clinical and Translational Imaging* 2.4, pp. 321–330.
- Hosseini-Asl, E., R. Keynton, and A. El-Baz (2016). “Alzheimer’s disease diagnostics by adaptation of 3D convolutional network”. In: *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 126–130.
- Hosseini Asl, E., M. Ghazal, A. Mahmoud, A. Aslantas, A. Shalaby, M. Casanova, G. Barnes, G. Gimel’farb, R. Keynton, and A. El Baz (2018). “Alzheimer’s disease diagnostics by a 3D deeply supervised adaptable convolutional network”. In: *Front. Biosci.* 23.2, pp. 584–596.
- Hounsfield, G. N. (1973). “Computerized Transverse Axial Scanning (Tomography): Part 1. Description of System”. In: *The British Journal of Radiology* 46.552, pp. 1016–1022.
- Hutton, B. F., M. Occhipinti, A. Kuehne, D. Máthé, N. Kovács, H. Waiczies, K. Erlandsson, D. Salvado, M. Carminati, G. L. Montagnani, et al. (2018). “Development of clinical simultaneous SPECT/MRI”. In: *The British journal of radiology* 91.1081, p. 20160690.
- Hutton, C., J. Declerck, M. A. Mintun, M. J. Pontecorvo, M. D. Devous, A. D. Joshi, A. D. N. Initiative, et al. (2015). “Quantification of ^{18}F -florbetapir PET: comparison of two analysis methods”. In: *Eur. J. Nucl. Med. Mol. Imaging* 42.5, pp. 725–732.
- Isensee, F., M. Schell, I. Pfleger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, and P. Kikingereder (2019). “Automated brain extraction of multisequence MRI using artificial neural networks”. In: *Human Brain Mapping* 40.17, pp. 4952–4964.

- Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros (2017). “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.
- Jack, C. R., M. A. Bernstein, B. J. Borowski, J. L. Gunter, N. C. Fox, P. M. Thompson, N. Schuff, G. Krueger, R. J. Killiany, C. S. DeCarli, A. M. Dale, O. W. Carmichael, D. Tosun, and M. W. Weiner (2010). “Update on the magnetic resonance imaging core of the Alzheimer’s disease neuroimaging initiative”. In: *Alzheimer’s & Dementia* 6.3, pp. 212–220.
- Jack, C. R., V. J. Lowe, M. L. Senjem, S. D. Weigand, B. J. Kemp, M. M. Shiung, D. S. Knopman, B. F. Boeve, W. E. Klunk, C. A. Mathis, and R. C. Petersen (2008). “11C PiB and Structural MRI Provide Complementary Information in Imaging of Alzheimer’s Disease and Amnestic Mild Cognitive Impairment”. In: *Brain* 131.3, pp. 665–680.
- Jagust, W. (2006). “Positron emission tomography and magnetic resonance imaging in the diagnosis and prediction of dementia”. In: *Alzheimer’s & Dementia* 2.1, pp. 36–42.
- Jagust, W. J., S. M. Landau, R. A. Koeppe, E. M. Reiman, K. Chen, C. A. Mathis, J. C. Price, N. L. Foster, and A. Y. Wang (2015). “The Alzheimer’s disease neuroimaging initiative 2 PET core: 2015”. In: *Alzheimer’s & Dementia* 11.7, pp. 757–771.
- Jaszczak, R. J., P. H. Murphy, D. Huard, and J. A. Burdine (1977). “Radionuclide Emission Computed Tomography of the Head with 99mTc and a Scintillation Camera”. In: *Journal of Nuclear Medicine* 18.4, pp. 373–380.
- Jenkinson, M., C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith (2012). “FSL”. In: *NeuroImage*. 20 YEARS OF fMRI 62.2, pp. 782–790.
- Jeong, Y., S. S. Cho, J. M. Park, S. J. Kang, J. S. Lee, E. Kang, D. L. Na, and S. E. Kim (2005). “18F-FDG PET Findings in Frontotemporal Dementia: An SPM Analysis of 29 Patients”. In: *J. Nucl. Med.* 46.2, pp. 233–239.
- Jiang, D., W. Dou, L. Vosters, X. Xu, Y. Sun, and T. Tan (2018). “Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network”. In: *Japanese journal of radiology* 36.9, pp. 566–574.
- Jo, T., K. Nho, and A. J. Saykin (2019). “Deep learning in Alzheimer’s disease: diagnostic classification and prognostic prediction using neuroimaging data”. In: *Frontiers in Aging Neuroscience* 11, p. 220.
- Joliot, M., G. Jobard, M. Naveau, N. Delcroix, L. Petit, L. Zago, F. Crivello, E. Mellet, B. Mazoyer, and N. Tzourio-Mazoyer (2015). “AICHA: An atlas of intrinsic connectivity of homotopic areas”. In: *Journal of Neuroscience Methods* 254, pp. 46–59.
- Jones, E., T. Oliphant, P. Peterson, et al. (2001). *SciPy: Open Source Scientific Tools for Python*.
- Jónsson, B. A., G. Bjornsdottir, T. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. Gudbjartsson, H. Stefansson, K. Stefansson, and M. Ulfarsson (2019). “Brain age prediction using deep learning uncovers associated sequence variants”. In: *Nature Communications* 10.1, pp. 1–10.
- Jungo, A. and M. Reyes (2019). “Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib,

- C. Essert, S. Zhou, P.-T. Yap, and A. Khan. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 48–56.
- Kalpouzos, G., G. Chételat, J.-C. Baron, B. Landeau, K. Mevel, C. Godeau, L. Barré, J.-M. Constans, F. Viader, F. Eustache, and B. Desgranges (2009). “Voxel-based mapping of brain gray matter volume and glucose metabolism profiles in normal aging”. In: *Neurobiol. Aging* 30.1, pp. 112–124.
- Kanda, T., K. Ishii, T. Uemura, N. Miyamoto, T. Yoshikawa, A. K. Kono, and E. Mori (2008). “Comparison of grey matter and metabolic reductions in frontotemporal dementia using FDG-PET and voxel-based morphometric MR studies”. In: *Eur. J. Nucl. Med. Mol. Imaging* 35.12, pp. 2227–2234.
- “Reproducibility in Machine Learning Research” (2017). In: *Workshop of the International Conference on Machine Learning, Sydney, Australia*. Ed. by N. R. Ke, A. Goyal, A. Lamb, J. Pineau, S. Bengio, and Y. Bengio.
- Kendall, A. and Y. Gal (2017). “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: *Conference on Neural Information Processing Systems*.
- Keshavan, A., E. Datta, I. M. McDonough, C. R. Madan, K. Jordan, and R. G. Henry (2018). “Mindcontrol: A web application for brain segmentation quality control”. In: *NeuroImage* 170, pp. 365–372.
- Keyes, J. W., N. Orlandea, W. J. Heetderks, P. F. Leonard, and W. L. Rogers (1977). “The Humongotron—A Scintillation-Camera Transaxial Tomograph”. In: *Journal of Nuclear Medicine* 18.4, pp. 381–387.
- Kim, H., A. Irimia, S. M. Hobel, M. Pogosyan, H. Tang, P. Petrosyan, R. E. C. Blanco, B. A. Duffy, L. Zhao, K. L. Crawford, S.-L. Liew, K. Clark, M. Law, P. Mukherjee, G. T. Manley, J. D. Van Horn, and A. W. Toga (2019). “LONI QC system: a semi-automated, web-based and freely-available environment for the comprehensive quality control of neuroimaging data”. In: *Frontiers in Neuroinformatics* 13, p. 60.
- Kim, K. H., W.-J. Do, and S.-H. Park (2018). “Improving resolution of MR images with an adversarial network incorporating images with different contrast”. In: *Medical Physics* 45.7, pp. 3120–3131.
- Kläser, K., P. Markiewicz, M. Ranzini, W. Li, M. Modat, B. F. Hutton, D. Atkinson, K. Thielemans, M. J. Cardoso, and S. Ourselin (2018). “Deep Boosted Regression for MR to CT Synthesis”. In: *Simulation and Synthesis in Medical Imaging*, pp. 61–70.
- Kleesiek, J., J. N. Morshuis, F. Isensee, K. Deike-Hofmann, D. Paech, P. Kickingeder, U. Köthe, C. Rother, M. Forsting, W. Wick, et al. (2019). “Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study”. In: *Investigative Radiology* 54.10, pp. 653–660.
- Klöppel, S., J. Peter, A. Ludl, A. Pilatus, S. Maier, I. Mader, B. Heimbach, L. Frings, K. Egger, J. Dukart, et al. (2015). “Applying automated MR-based diagnostic methods to the memory clinic: a prospective study”. In: *Journal of Alzheimer’s disease* 47.4, pp. 939–954.
- Klöppel, S., C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. Frackowiak (2008). “Automatic classification of MR scans in Alzheimer’s disease”. In: *Brain* 131.3, pp. 681–689.

- Knopman, D. S., C. R. Jack, H. J. Wiste, E. S. Lundt, S. D. Weigand, P. Vemuri, V. J. Lowe, K. Kantarci, J. L. Gunter, M. L. Senjem, M. M. Mielke, R. O. Roberts, B. F. Boeve, and R. C. Petersen (2014). “¹⁸F-fluorodeoxyglucose positron emission tomography, aging, and apolipoprotein E genotype in cognitively normal persons”. In: *Neurobiol. Aging* 35.9, pp. 2096–2106.
- Knopman, D. S., J. H. Kramer, B. F. Boeve, R. J. Caselli, N. R. Graff-Radford, M. F. Mendez, B. L. Miller, and N. Mercaldo (2008). “Development of methodology for conducting clinical trials in frontotemporal lobar degeneration”. In: *Brain* 131.11, pp. 2957–2968.
- Koikkalainen, J., H. Rhodius-Meester, A. Tolonen, F. Barkhof, B. Tijms, A. W. Lemstra, T. Tong, R. Guerrero, A. Schuh, C. Ledig, et al. (2016). “Differential diagnosis of neurodegenerative diseases using structural MRI data”. In: *NeuroImage: Clinical* 11, pp. 435–449.
- Kriegeskorte, N., W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker (2009). “Circular analysis in systems neuroscience: the dangers of double dipping”. In: *Nat. Neurosci.* 12.5, pp. 535–540.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). “Imagenet Classification with Deep Convolutional Neural Networks”. In: *Advances in neural information processing systems* 25, pp. 1097–1105.
- Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. Springer, New York, NY.
- Ladefoged, C. N., L. Marner, A. Hindsholm, I. Law, L. Højgaard, and F. L. Andersen (2019). “Deep learning based attenuation correction of PET/MRI in pediatric brain tumor patients: evaluation in a clinical setting”. In: *Frontiers in neuroscience* 12, p. 1005.
- Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles”. In: *Conference on Neural Information Processing Systems*.
- Landau, S. M., M. A. Mintun, A. D. Joshi, R. A. Koeppe, R. C. Petersen, P. S. Aisen, M. W. Weiner, and W. J. Jagust (2012). “Amyloid deposition, hypometabolism, and longitudinal cognitive decline”. In: *Ann. Neurol.* 72.4, pp. 578–586.
- Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller (2019). “Unmasking Clever Hans predictors and assessing what machines really learn”. In: *Nature communications* 10.1, pp. 1–8.
- LeCun, Y., Y. Bengio, and G. Hinton (2015). “Deep learning”. In: *Nature* 521.7553, pp. 436–444.
- Lee, J. Y., S. W. Oh, M. S. Chung, J. E. Park, Y. Moon, H. J. Jeon, and W.-J. Moon (2021). “Clinically available software for automatic brain volumetry: comparisons of volume measurements and validation of intermethod reliability”. In: *Korean journal of radiology* 22.3, p. 405.
- Li, F., M. Liu, and Alzheimer’s Disease Neuroimaging Initiative (2018). “Alzheimer’s disease diagnosis based on multiple cluster dense convolutional networks”. In: *Comput. Med. Imaging Graph.* 70, pp. 101–110.

- Li, R., W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji (2014). “Deep learning based imaging data completion for improved brain disease diagnosis”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 305–312.
- Li, X., P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden (2016). “The first step for neuroimaging data analysis: DICOM to NIFTI conversion”. In: *Journal of Neuroscience Methods* 264, pp. 47–56.
- Lian, C., M. Liu, J. Zhang, and D. Shen (2018). “Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer’s Disease Diagnosis using Structural MRF”. In: *IEEE Trans. Pattern Anal. Mach. Intell.*
- Lin, W., T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, and Alzheimer’s Disease Neuroimaging Initiative (2018). “Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer’s Disease Prediction From Mild Cognitive Impairment”. In: *Front. Neurosci.* 12, p. 777.
- Lipton, Z. C. (2018). “The Mythos of Model Interpretability”. In: *Communications of the ACM* 61.10, pp. 36–43.
- Litjens, G., T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. Sánchez (2017). “A Survey on Deep Learning in Medical Image Analysis”. In: *Medical Image Analysis* 42, pp. 60–88.
- Liu, M., D. Cheng, K. Wang, Y. Wang, and Alzheimer’s Disease Neuroimaging Initiative (2018a). “Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer’s Disease Diagnosis”. In: *Neuroinformatics* 16.3-4, pp. 295–308.
- Liu, M., J. Zhang, E. Adeli, and D. Shen (2018b). “Landmark-based deep multi-instance learning for brain disease diagnosis”. In: *Med. Image Anal.* 43, pp. 157–168.
- Liu, M., J. Zhang, D. Nie, P.-T. Yap, and D. Shen (2018c). “Anatomical Landmark Based Deep Feature Representation for MR Images in Brain Disease Diagnosis”. In: *IEEE J Biomed Health Inform* 22.5, pp. 1476–1485.
- Ma, D., D. Lu, K. Popuri, L. Wang, M. F. Beg, A. D. N. Initiative, et al. (2020). “Differential Diagnosis of Frontotemporal Dementia, Alzheimer’s Disease, and Normal Aging Using a Multi-Scale Multi-Type Feature Generative Adversarial Deep Neural Network on Structural Magnetic Resonance Images”. In: *Frontiers in Neuroscience* 14, p. 853.
- Maggipinto, T., R. Bellotti, N. Amoroso, D. Diacono, G. Donvito, E. Lella, A. Monaco, M. A. Scelsi, and S. Tangaro (2017). “DTI measurements for Alzheimer’s classification”. In: *Physics in Medicine & Biology* 62.6, p. 2361.
- Manera, A. L., M. Dadar, J. C. Van Swieten, B. Borroni, R. Sanchez-Valle, F. Moreno, R. Laforce Jr, C. Graff, M. Synofzik, D. Galimberti, et al. (2021). “MRI data-driven algorithm for the diagnosis of behavioural variant frontotemporal dementia”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 92.6, pp. 608–616.
- Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley (2017). “Least squares generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802.
- Marcus, D. S., T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner (2007). “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data

- in Young, Middle Aged, Nondemented, and Demented Older Adults”. In: *Journal of Cognitive Neuroscience* 19.9, pp. 1498–1507.
- Marinescu, R. V., N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, S. Klein, D. C. Alexander, et al. (2018). “Tadpole challenge: Prediction of longitudinal evolution in Alzheimer’s disease”. In: *arXiv preprint arXiv:1805.03909*.
- McDermott, M. B., S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi (2021). “Reproducibility in machine learning for health research: Still a ways to go”. In: *Science Translational Medicine* 13.586, eabb1655.
- McKinney, W. (2010). “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman, pp. 51–56.
- Milletari, F., N. Navab, and S.-A. Ahmadi (2016). “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE, pp. 565–571.
- Minoshima, S., K. A. Frey, N. L. Foster, and D. E. Kuhl (1995a). “Preserved pontine glucose metabolism in Alzheimer disease: a reference region for functional brain image (PET) analysis”. In: *J. Comp. Assist. Tomo.* 19.4, pp. 541–547.
- Minoshima, S., K. A. Frey, R. A. Koeppe, N. L. Foster, and D. E. Kuhl (1995b). “A Diagnostic Approach in Alzheimer’s Disease Using Three-Dimensional Stereotactic Surface Projections of Fluorine-18-FDG PET”. In: *J. Nucl. Med.*, p. 12.
- Modat, M., D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, and S. Ourselin (2014). “A symmetric block-matching framework for global registration”. In: *Proc. SPIE Medical Imaging* 9034.
- Modat, M., G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin (2010). “Fast free-form deformation using graphics processing units”. In: *Comput. Methods Programs Biomed.* 98.3, pp. 278–84.
- Moradi, E., A. Pepe, C. Gaser, H. Huttunen, and J. Tohka (2015). “Machine learning framework for early MRI-based Alzheimer’s conversion prediction in MCI subjects”. In: *NeuroImage* 104, pp. 398–412.
- Morin, A., J. Samper-González, A. Bertrand, S. Ströer, D. Dormont, A. Mendes, P. Coupé, J. Ahdidan, M. Lévy, D. Samri, et al. (2020). “Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort”. In: *Journal of Alzheimer’s Disease* 74.4, pp. 1157–1166.
- Mosconi, L., W. H. Tsui, K. Herholz, A. Pupi, A. Drzezga, G. Lucignani, E. M. Reiman, V. Holthoff, E. Kalbe, S. Sorbi, J. Diehl-Schmid, R. Perneczky, F. Clerici, R. Caselli, B. Beuthien-Baumann, A. Kurz, S. Minoshima, and M. J. de Leon (2008). “Multicenter Standardized 18F-FDG PET Diagnosis of Mild Cognitive Impairment, Alzheimer’s Disease, and Other Dementias”. In: *J. Nucl. Med.* 49.3, pp. 390–398.
- Murphy, D. G. M., C. DeCarli, A. R. McIntosh, E. Daly, M. J. Mentis, P. Pietrini, J. Szczepanik, M. B. Schapiro, C. L. Grady, B. Horwitz, and S. I. Rapoport (1996). “Sex Differences in Human Brain Morphometry and Metabolism: An In Vivo Quantitative Magnetic Resonance Imaging and Positron Emission Tomography Study on the Effect of Aging”. In: *Arch. Gen. Psychiatry* 53.7, pp. 585–594.

- Nadeau, C. and Y. Bengio (2003). “Inference for the Generalization Error”. In: *Machine Learning* 52.3, pp. 239–281.
- Neppl, S., G. Landry, C. Kurz, D. C. Hansen, B. Hoyle, S. Stöcklein, M. Seidensticker, J. Weller, C. Belka, K. Parodi, et al. (2019). “Evaluation of proton and photon dose distributions recalculated on 2D and 3D Unet-generated pseudoCTs from T1-weighted MR head scans”. In: *Acta Oncologica* 58.10, pp. 1429–1434.
- Nie, D., X. Cao, Y. Gao, L. Wang, and D. Shen (2016). “Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks”. In: *Deep Learning and Data Labeling for Medical Applications*, pp. 170–178.
- Oktay, O., J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. (2018). “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999*.
- Ossenkoppele, R., N. D. Prins, Y. A. Pijnenburg, A. W. Lemstra, W. M. van der Flier, S. F. Adriaanse, A. D. Windhorst, R. L. Handels, C. A. Wolfs, P. Aalten, F. R. Verhey, M. M. Verbeek, M. A. van Buchem, O. S. Hoekstra, A. A. Lammertsma, P. Scheltens, and B. N. van Berckel (2013). “Impact of molecular imaging on the diagnostic process in a memory clinic”. In: *Alzheimer’s & Dementia* 9.4, pp. 414–421.
- Ota, K., N. Oishi, K. Ito, and H. Fukuyama (2014). “A comparison of three brain atlases for MCI prediction”. In: *Journal of Neuroscience Methods* 221, pp. 139–150.
- (2015). “Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer’s disease”. In: *Journal of Neuroscience Methods* 256, pp. 168–183.
- Panegyres, P. K., J. M. Rogers, M. McCarthy, A. Campbell, and J. S. Wu (2009). “Fluorodeoxyglucose-Positron Emission Tomography in the differential diagnosis of early-onset dementia: a prospective, community-based study”. In: *BMC Neurology* 9.1.
- Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimsheine, L. Antiga, et al. (2019). “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32.
- Patterson, J. C., D. L. Lilien, A. Takalkar, and J. B. Pinkston (2011). “Early Detection of Brain Pathology Suggestive of Early AD Using Objective Evaluation of FDG-PET Scans”. In: *Int. J. Alzheimer’s Disease*.
- Patton, J. A., D. Delbeke, and M. P. Sandler (2000). “Image fusion using an integrated, dual-head coincidence camera with X-ray tube-based attenuation maps”. In: *Journal of Nuclear Medicine* 41.8, pp. 1364–1368.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). “Scikit-learn: Machine learning in Python”. In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830.
- Penny, W. D., K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.
- Perani, D., P. A. Della Rosa, C. Cerami, F. Gallivanone, F. Fallanca, E. G. Vanoli, A. Panzacchi, F. Nobili, S. Pappatà, A. Marcone, V. Garibotto, I. Castiglioni, G. Maggiani, S. F. Cappa, and L. Gianolli (2014). “Validation of an optimized SPM procedure

- for FDG-PET in dementia diagnosis in a clinical setting”. In: *NeuroImage: Clinical* 6, pp. 445–454.
- Petersen, R. C., P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner (2010). “Alzheimer’s Disease Neuroimaging Initiative (ADNI): clinical characterization”. In: *Neurology* 74.3, pp. 201–209.
- Pham, C.-H., A. Ducournau, R. Fablet, and F. Rousseau (2017). “Brain MRI super-resolution using deep 3D convolutional networks”. In: *2017 IEEE ISBI*, pp. 197–200.
- Pinaya, W. H., P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso (2022). “Unsupervised brain imaging 3D anomaly detection and segmentation with transformers”. In: *Medical Image Analysis* 79, p. 102475.
- Platero, C., M. E. López, M. d. Carmen Tobar, M. Yus, and F. Maestu (2019). “Discriminating Alzheimer’s disease progression using a new hippocampal marker from T1-weighted MRI: The local surface roughness”. In: *Human brain mapping* 40.5, pp. 1666–1676.
- Poldrack, R. A., C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews, M. R. Munafò, T. E. Nichols, J.-B. Poline, E. Vul, and T. Yarkoni (2017). “Scanning the horizon: towards transparent and reproducible neuroimaging research”. In: *Nature Reviews. Neuroscience* 18.2, pp. 115–126.
- Querbes, O., F. Aubry, J. Pariente, J.-A. Lotterie, J.-F. Démonet, V. Duret, M. Puel, I. Berry, J.-C. Fort, and P. Celsis (2009). “Early diagnosis of Alzheimer’s disease using cortical thickness: impact of cognitive reserve”. In: *Brain: A Journal of Neurology* 132.Pt 8, pp. 2036–2047.
- Raamana, P. R. (2017). *Neuropredict: Easy Machine Learning And Standardized Predictive Analysis Of Biomarkers*.
- Raamana, P. R. and S. C. Strother (2017). “Impact of spatial scale and edge weight on predictive power of cortical thickness networks”. In: *bioRxiv*.
- Rabinovici, G. D., W. J. Jagust, A. J. Furst, J. M. Ogar, C. A. Racine, E. C. Mormino, J. P. O’Neil, R. A. Lal, N. F. Dronkers, B. L. Miller, and M. L. Gorno-Tempini (2008). “A β amyloid and glucose metabolism in three variants of primary progressive aphasia”. In: *Ann. Neurol.* 64.4, pp. 388–401.
- Raschka, S. (2015). *Python Machine Learning*. Packt Publishing Ltd.
- Rathore, S., M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos (2017). “A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer’s disease and its prodromal stages”. In: *NeuroImage* 155, pp. 530–548.
- Renard, D., R. Vandenberghe, L. Collombier, P.-O. Kotzki, J.-P. Pouget, and V. Boudousq (2013). “Glucose metabolism in nine patients with probable sporadic Creutzfeldt-Jakob disease: FDG-PET study using SPM and individual patient analysis”. In: *J. Neurol.* 260.12, pp. 3055–3064.
- Reuter, M., M. D. Tisdall, A. Qureshi, R. L. Buckner, A. J. van der Kouwe, and B. Fischl (2015). “Head motion during MRI acquisition reduces gray matter volume and thickness estimates”. In: *NeuroImage* 107, pp. 107–115.
- Rice, L. and S. Bisdas (2017). “The diagnostic value of FDG and amyloid PET in Alzheimer’s disease—A systematic review”. In: *Eur. J. Radiol.* 94, pp. 16–24.

- Ripley, B. D. (1996). *Pattern Recognition and Neural Networks by Brian D. Ripley*. Cambridge University Press.
- Rogers, T. T., J. Hocking, U. Noppeney, A. Mechelli, M. L. Gorno-Tempini, K. Patterson, and C. J. Price (2006). “Anterior temporal cortex and semantic memory: Reconciling findings from neuropsychology and functional imaging”. In: *Cogn. Affect. Behav. Neurosci.* 6.3, pp. 201–213.
- Rohrer, J. D. and H. J. Rosen (2013). “Neuroimaging in frontotemporal dementia”. In: *Int. Rev. Psychiatry* 25.2, pp. 221–229.
- Rolls, E. T., M. Joliot, and N. Tzourio-Mazoyer (2015). “Implementation of a New Parcelation of the Orbitofrontal Cortex in the Automated Anatomical Labeling Atlas”. In: *NeuroImage* 122, pp. 1–5.
- Ronneberger, O., P. Fischer, and T. Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.
- Röntgen, W. C. (1896). “On a New Kind of Rays”. In: *Science* 3.59, pp. 227–231.
- Routier, A., N. Burgos, M. Díaz, M. Bacci, S. Bottani, O. El-Rifai, S. Fontanella, P. Gori, J. Guillon, A. Guyot, R. Hassanaly, T. Jacquemont, P. Lu, A. Marcoux, T. Moreau, J. Samper-González, M. Teichmann, E. Thibeau-Sutre, G. Vaillant, J. Wen, A. Wild, M.-O. Habert, S. Durrleman, and O. Colliot (2021). “Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies”. In: *Frontiers in Neuroinformatics* 15, p. 39.
- Sabuncu, M. R. and E. Konukoglu (2014). “Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study”. In: *Neuroinformatics*.
- Samper-González, J., N. Burgos, S. Bottani, S. Fontanella, P. Lu, A. Marcoux, A. Routier, J. Guillon, M. Bacci, J. Wen, A. Bertrand, H. Bertin, M.-O. Habert, S. Durrleman, T. Evgeniou, and O. Colliot (2018). “Reproducible Evaluation of Classification Methods in Alzheimer’s Disease: Framework and Application to MRI and PET Data”. In: *NeuroImage* 183, pp. 504–521.
- Sarle, W. S. (1997). “Neural Network FAQ, part 1 of 7”. In: *Introduction, periodic posting to the Usenet newsgroup comp. ai. neural-nets URL: ftp://ftp.sas.com/pub/neural/FAQ.html*.
- Schlegl, T., P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth (2019). “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks”. In: *Med. Image Anal.* 54, pp. 30–44.
- Schlemmer, H.-P., B. Pichler, K. Wienhard, M. Schmand, C. Nahmias, D. Townsend, W.-D. Heiss, and C. Claussen (2007). “Simultaneous MR/PET for Brain Imaging: First Patient Scans”. In: *Journal of Nuclear Medicine* 48.supplement 2, 45P–45P.
- Schmand, M., Z. Burbar, J. Corbeil, N. Zhang, C. Michael, L. Byars, L. Eriksson, R. Grazioso, M. Martin, A. Moor, J. Camp, V. Matschl, R. Ladebeck, W. Renz, H. Fischer, K. Jattke, G. Schnur, N. Rietsch, B. Bendriem, and W.-D. Heiss (2007). “BrainPET: First Human Tomograph for Simultaneous (Functional) PET and MR Imaging”. In: *Journal of Nuclear Medicine* 48.supplement 2, 45P–45P.

- Schwarz, C. G., J. L. Gunter, H. J. Wiste, S. A. Przybelski, S. D. Weigand, C. P. Ward, M. L. Senjem, P. Vemuri, M. E. Murray, D. W. Dickson, J. E. Parisi, K. Kantarci, M. W. Weiner, R. C. Petersen, and C. R. Jack (2016). “A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer’s disease severity”. In: *NeuroImage. Clinical* 11, pp. 802–812.
- Shattuck, D. W., M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga (2008). “Construction of a 3D probabilistic atlas of human cortical structures”. In: *NeuroImage* 39.3, pp. 1064–1080.
- Shiri, I., P. Ghafarian, P. Geramifar, K. H.-Y. Leung, M. Ghelichoghli, M. Oveisi, A. Rahmim, and M. R. Ay (2019). “Direct attenuation correction of brain PET images using only emission data via a deep convolutional encoder-decoder (Deep-DAC)”. In: *European Radiology* 29.12, pp. 6867–6879.
- Shmulev, Y. and M. Belyaev (2018). “Predicting conversion of mild cognitive impairments to Alzheimer’s disease and exploring impact of neuroimaging”. In: *Graphs in biomedical image analysis and integrating medical imaging and non-imaging modalities*. Springer, pp. 83–91.
- Signorini, M., E. Paulesu, K. Friston, D. Perani, A. Colleluori, G. Lucignani, F. Grassi, V. Bettinardi, R. S. J. Frackowiak, and F. Fazio (1999). “Rapid Assessment of Regional Cerebral Metabolic Abnormalities in Single Subjects with Quantitative and Nonquantitative [18F]FDG PET: A Clinical Validation of Statistical Parametric Mapping”. In: *NeuroImage* 9.1, pp. 63–80.
- Sikka, A., S. V. Peri, and D. R. Bathula (2018). “MRI to FDG-PET: Cross-Modal Synthesis Using 3D U-Net for Multi-modal Alzheimer’s Classification”. In: *Simulation and Synthesis in Medical Imaging*, pp. 80–89.
- Simonyan, K., A. Vedaldi, and A. Zisserman (2014). “Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *In Workshop at International Conference on Learning Representations*.
- Smith-Bindman, R., M. L. Kwan, E. C. Marlow, M. K. Theis, W. Bolch, S. Y. Cheng, E. J. A. Bowles, J. R. Duncan, R. T. Greenlee, L. H. Kushi, J. D. Pole, A. K. Rahm, N. K. Stout, S. Weinmann, and D. L. Miglioretti (2019). “Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016”. In: *JAMA* 322.9, pp. 843–856.
- Sohn, B. K., D. Yi, E. H. Seo, Y. M. Choe, J. W. Kim, S. G. Kim, H. J. Choi, M. S. Byun, J. H. Jhoo, J. I. Woo, et al. (2015). “Comparison of regional gray matter atrophy, white matter alteration, and glucose metabolism as a predictor of the conversion to Alzheimer’s disease in mild cognitive impairment”. In: *Journal of Korean medical science* 30.6, pp. 779–787.
- Spuhler, K. D., J. Gardus, Y. Gao, C. DeLorenzo, R. Parsey, and C. Huang (2019). “Synthesis of patient-specific transmission data for PET attenuation correction for PET/MRI neuroimaging using a convolutional neural network”. In: *Journal of nuclear medicine* 60.4, pp. 555–560.
- Staffaroni, A. M., P. A. Ljubenkov, J. Kornak, Y. Cobigo, S. Datta, G. Marx, S. M. Walters, K. Chiang, N. Olney, F. M. Elahi, D. S. Knopman, B. C. Dickerson, B. F. Boeve, M. L.

- Gorno-Tempini, S. Spina, L. T. Grinberg, W. W. Seeley, B. L. Miller, J. H. Kramer, A. L. Boxer, and H. J. Rosen (2019). “Longitudinal multimodal imaging and clinical endpoints for frontotemporal dementia clinical trials”. In: *Brain* 142.2, pp. 443–459.
- Sujit, S. J., I. Coronado, A. Kamali, P. A. Narayana, and R. E. Gabr (2019). “Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks”. In: *Journal of Magnetic Resonance Imaging* 50.4, pp. 1260–1267.
- Sun, H., X. Liu, X. Feng, C. Liu, N. Zhu, S. J. Gjerswold-Selleck, H.-J. Wei, P. S. Upadhyayula, A. Mela, C.-C. Wu, P. D. Canoll, A. F. Laine, J. T. Vaughan, S. A. Small, and J. Guo (2020a). “Substituting Gadolinium in Brain MRI Using DeepContrast”. In: *2020 IEEE ISBI*, pp. 908–912.
- Sun, L., J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley (2020b). “An Adversarial Learning Approach to Medical Image Synthesis for Lesion Detection”. In: *IEEE Journal of Biomedical and Health Informatics* 24.8, pp. 2303–2314.
- Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.
- Sørensen, L. and M. Nielsen (2018). “Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination”. In: *Journal of Neuroscience Methods* 302, pp. 66–74.
- Ter-Pogossian, M. M., M. E. Phelps, E. J. Hoffman, and N. A. Mullani (1975). “A Positron-Emission Transaxial Tomograph for Nuclear Imaging (PETT)”. In: *Radiology* 114.1, pp. 89–98.
- Thibeau-Sutre, E. (2021). “Reproducible and Interpretable Deep Learning for the Diagnosis, Prognosis and Subtyping of Alzheimer’s Disease from Neuroimaging Data”. PhD thesis. Sorbonne université.
- Thibeau-Sutre, E., S. Collin, N. Burgos, and O. Colliot (2022a). “Interpretability of Machine Learning Methods Applied to Neuroimaging”. In.
- Thibeau-Sutre, E., O. Colliot, D. Dormont, and N. Burgos (2020). “Visualization Approach to Assess the Robustness of Neural Networks for Medical Image Classification”. In: *SPIE Medical Imaging 2020*. Vol. 11313. International Society for Optics and Photonics, 113131J.
- Thibeau-Sutre, E., M. Díaz, R. Hassanaly, A. Routier, D. Dormont, O. Colliot, and N. Burgos (2022b). “ClinicaDL: An Open-Source Deep Learning Software for Reproducible Neuroimaging Processing”. In: *Computer Methods and Programs in Biomedicine* 220, p. 106818.
- Thomas, B. A., V. Cuplov, A. Bousse, A. Mendes, K. Thielemans, B. F. Hutton, and K. Erlandsson (2016). “PETPVC: A Toolbox for Performing Partial Volume Correction Techniques in Positron Emission Tomography”. In: *Physics in Medicine and Biology* 61.22, pp. 7975–7993.
- Thompson, W. H., J. Wright, P. G. Bissett, and R. A. Poldrack (2020). “Dataset Decay and the Problem of Sequential Analyses on Open Datasets”. In: *eLife* 9, e53498.

- Tohka, J., E. Moradi, and H. Huttunen (2016). “Comparison of Feature Selection Techniques in Machine Learning for Anatomical Brain MRI in Dementia”. In: *Neuroinformatics* 14.3, pp. 279–296.
- Tosun, D., N. Schuff, G. D. Rabinovici, N. Ayakta, B. L. Miller, W. Jagust, J. Kramer, M. M. Weiner, and H. J. Rosen (2016). “Diagnostic utility of ASL-MRI and FDG-PET in the behavioral variant of FTD and AD”. In: *Ann. Clin. Transl. Neurol.* 3.10, pp. 740–751.
- Tournier, J.-D., R. Smith, D. Raffelt, R. Tabbara, T. Dhollander, M. Pietsch, D. Christiaens, B. Jeurissen, C.-H. Yeh, and A. Connelly (2019). “MRtrix3: A Fast, Flexible and Open Software Framework for Medical Image Processing and Visualisation”. In: *NeuroImage* 202, p. 116137.
- Townsend, D. W., T. Beyer, and T. M. Blodgett (2003). “PET/CT Scanners: A Hardware Approach to Image Fusion”. In: *Seminars in Nuclear Medicine* 33.3, pp. 193–204.
- Tsunoda, Y., M. Moribe, H. Orii, H. Kawano, and H. Maeda (2014). “Pseudo-normal Image Synthesis from Chest Radiograph Database for Lung Nodule Detection”. In: *Advanced Intelligent Systems*, pp. 147–155.
- Tustison, N. J., B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee (2010). “N4ITK: improved N3 bias correction”. In: *IEEE Trans. Med. Imaging* 29.6, pp. 1310–1320.
- Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (2002). “Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain”. In: *NeuroImage* 15.1, pp. 273–289.
- Uzunova, H., S. Schultz, H. Handels, and J. Ehrhardt (2019). “Unsupervised pathology detection in medical images using conditional variational autoencoders”. In: *Int. J. Comput. Assist. Radiol. Surg.* 14.3, pp. 451–461.
- van der Walt, S., S. C. Colbert, and G. Varoquaux (2011). “The NumPy Array: A Structure for Efficient Numerical Computation”. In: *Computing in Science & Engineering* 13.2, pp. 22–30.
- van der Walt, S., J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu (2014). “Scikit-Image: Image Processing in Python”. In: *PeerJ* 2, e453.
- Varoquaux, G., P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion (2017). “Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines”. In: *NeuroImage. Individual Subject Prediction* 145, pp. 166–179.
- Vemuri, P., J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack Jr (2008). “Alzheimer’s disease diagnosis in individual subjects using structural MR images: validation studies”. In: *Neuroimage* 39.3, pp. 1186–1197.
- Voevodskaya, O., A. Simmons, R. Nordenskjöld, J. Kullberg, H. Ahlström, L. Lind, L.-O. Wahlund, E.-M. Larsson, and E. Westman (2014). “The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer’s disease”. In: *Frontiers in Aging Neuroscience* 6, p. 264.

- Vos, B. D. de, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum (2019). “A Deep Learning Framework for Unsupervised Affine and Deformable Image Registration”. In: *Med. Image Anal.* 52, pp. 128–143.
- Vu, T.-D., N.-H. Ho, H.-J. Yang, J. Kim, and H.-C. Song (2018). “Non-white matter tissue extraction and deep convolutional neural network for Alzheimer’s disease detection”. In: *Soft Comput* 22.20, pp. 6825–6833.
- Vu, T. D., H.-J. Yang, V. Q. Nguyen, A.-R. Oh, and M.-S. Kim (2017). “Multimodal learning using Convolution Neural Network and Sparse Autoencoder”. In: *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 309–312.
- Wallis, D. and I. Buvat (2022). “Clever Hans effect found in a widely used brain tumour MRI dataset”. In: *Medical Image Analysis*, p. 102368.
- Wang, H., Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, and X. Zhao (2019). “Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer’s disease”. In: *Neurocomputing* 333, pp. 145–156.
- Wang, S.-H., P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng (2018). “Classification of Alzheimer’s Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling”. In: *J. Med. Syst.* 42.5, p. 85.
- Wang, W., C. Chen, M. Ding, H. Yu, S. Zha, and J. Li (2021). “Transbts: Multimodal brain tumor segmentation using transformer”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 109–119.
- Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli (2004). “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Trans. Image Proc.* 13.4, pp. 600–612.
- Watson, P. and A. Petrie (2010). “Method agreement analysis: a review of correct methodology”. In: *Theriogenology* 73.9, pp. 1167–1179.
- Wei, W., E. Poirion, B. Bodini, S. Durrleman, N. Ayache, B. Stankoff, and O. Colliot (2018). “Learning Myelin Content in Multiple Sclerosis from Multimodal MRI Through Adversarial Training”. In: *Medical Image Computing and Computer Assisted Intervention*, pp. 514–522.
- (2019). “Predicting PET-derived Demyelination from Multimodal MRI using Sketcher-Refiner Adversarial Training for Multiple Sclerosis”. In: *Medical Image Analysis*, p. 101546.
- Wen, J., E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot (2020). “Convolutional Neural Networks for Classification of Alzheimer’s Disease: Overview and Reproducible Evaluation”. In: *Medical Image Analysis* 63, p. 101694.
- Westman, E., C. Aguilar, J.-S. Muehlboeck, and A. Simmons (2013). “Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer’s disease and mild cognitive impairment”. In: *Brain Topography* 26.1, pp. 9–23.
- Whitwell, J. L., S. A. Przybelski, S. D. Weigand, R. J. Ivnik, P. Vemuri, J. L. Gunter, M. L. Senjem, M. M. Shiung, B. F. Boeve, D. S. Knopman, J. E. Parisi, D. W. Dickson, R. C. Petersen, J. Jack Clifford R., and K. A. Josephs (2009). “Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: a cluster analysis study”. In: *Brain* 132.11, pp. 2932–2946.

- Whitwell, J. L., S. A. Przybelski, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack (2007). “3D Maps from Multiple MRI Illustrate Changing Atrophy Patterns as Subjects Progress from Mild Cognitive Impairment to Alzheimer’s Disease”. In: *Brain* 130.7, pp. 1777–1786.
- Wolterink, J. M., A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum (2017). “Deep MR to CT synthesis using unpaired data”. In: *International workshop on simulation and synthesis in medical imaging*. Springer, pp. 14–23.
- Wolz, R., V. Julkunen, J. Koikkalainen, E. Niskanen, D. P. Zhang, D. Rueckert, H. Soininen, and J. Lötjönen (2011). “Multi-method analysis of MRI images in early diagnostics of Alzheimer’s disease”. In: *PloS One* 6.10, e25446.
- World Health Organization et al. (2007). “International classification of diseases and related health problems, 10th revision”. In: <http://www.who.int/classifications/apps/icd/icd10online>.
- Worsley, K., J. Taylor, F Carbonell, M. Chung, E Duerden, B Bernhardt, O Lyttelton, M Boucher, and A. Evans (2009). “SurfStat: A Matlab Toolbox for the Statistical Analysis of Univariate and Multivariate Surface and Volumetric Data Using Linear Mixed Effects Models and Random Field Theory”. In: *NeuroImage*. Organization for Human Brain Mapping 2009 Annual Meeting 47, S102.
- Xia, T., A. Chatsias, and S. A. Tsaftaris (2020). “Pseudo-Healthy Synthesis with Pathology Disentanglement and Adversarial Learning”. In: *Medical Image Analysis* 64, p. 101719.
- Xiang, L., Q. Wang, D. Nie, L. Zhang, X. Jin, Y. Qiao, and D. Shen (2018). “Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image”. In: *Med. Image Anal.* 47, pp. 31–44.
- Xie, N., G. Ras, M. van Gerven, and D. Doran (2020). “Explainable Deep Learning: A Field Guide for the Uninitiated”. In: *arXiv:2004.14545 [cs, stat]*.
- Yaakub, S. N., C. J. McGinnity, J. R. Clough, E. Kerfoot, N. Girard, E. Guedj, and A. Hammers (2019). “Pseudo-Normal PET Synthesis with Generative Adversarial Networks for Localising Hypometabolism in Epilepsies”. In: *Simulation and Synthesis in Medical Imaging*, pp. 42–51.
- Yagis, E., S. W. Atnafu, A. García Seco de Herrera, C. Marzi, R. Sceda, M. Giannelli, C. Tessa, L. Citi, and S. Diciotti (2021). “Effect of data leakage in brain MRI classification using 2D convolutional neural networks”. In: *Scientific reports* 11.1, pp. 1–13.
- Yang, J., D. Park, G. T. Gullberg, and Y. Seo (2019). “Joint correction of attenuation and scatter in image space using deep convolutional neural networks for dedicated brain 18F-FDG PET”. In: *Physics in medicine & biology* 64.7, p. 075019.
- Yang, X., X. Han, E. Park, S. Aylward, R. Kwitt, and M. Niethammer (2016). “Registration of Pathological Images”. In: *Simulation and Synthesis in Medical Imaging*, pp. 97–107.
- Yang, Z., X. Zhuang, K. Sreenivasan, V. Mishra, T. Curran, and D. Cordes (2020). “A robust deep neural network for denoising task-based fMRI data: An application to working memory and episodic memory”. In: *Medical Image Analysis* 60, p. 101622.
- Ye, D. H., D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu (2013). “Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 606–613.

- You, S., K. C. Tezcan, X. Chen, and E. Konukoglu (2019). “Unsupervised Lesion Detection via Image Restoration with a Normative Prior”. In: *International Conference on Medical Imaging with Deep Learning*, pp. 540–556.
- Young, I. R., D. R. Bailes, M. Burl, A. G. Collins, D. T. Smith, M. J. McDonnell, J. S. Orr, L. M. Banks, G. M. Bydder, R. H. Greenspan, and R. E. Steiner (1982). “Initial Clinical Evaluation of a Whole Body Nuclear Magnetic Resonance (NMR) Tomograph”. In: *Journal of Computer Assisted Tomography* 6.1, pp. 1–18.
- Yu, Q., Y. Mai, Y. Ruan, Y. Luo, L. Zhao, W. Fang, Z. Cao, Y. Li, W. Liao, S. Xiao, et al. (2021). “An MRI-based strategy for differentiation of frontotemporal dementia and Alzheimer’s disease”. In: *Alzheimer’s research & therapy* 13.1, pp. 1–12.
- Zaki, L. A., M. W. Vernooij, M. Smits, C. Tolman, J. M. Papma, J. J. Visser, and R. M. Steketee (2022). “Comparing two artificial intelligence software packages for normative brain volumetry in memory clinic imaging”. In: *Neuroradiology*, pp. 1–8.
- Zeng, K., H. Zheng, C. Cai, Y. Yang, K. Zhang, and Z. Chen (2018). “Simultaneous single- and multi-contrast super-resolution for brain MRI images based on a convolutional neural network”. In: *Computers in Biology and Medicine* 99, pp. 133–141.
- Zimmerer, D., F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein (2019). “Unsupervised Anomaly Localization Using Variational Auto-Encoders”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 289–297.