Machine learning for neuroimaging using a very large scale clinical datawarehouse

Simona Bottani

Sorbonne Université

Doctoral Thesis

Machine learning for neuroimaging using a very large scale clinical datawarehouse

---

Author:
Simona Bottani

Supervisors:
Olivier Colliot
Ninon Burgos

Referees:
Emmanuel Barbier
Renaud Lopes
Xavier Tannier
Betty Tijms

A thesis submitted in fulfillment of the requirements for the degree of PhD in Computer Science

in the

ARAMIS Lab
Sorbonne Université, Inria Paris Center, Institut du Cerveau - Paris Brain Institute (ICM), Inserm U 1127, CNRS UMR 7225, AP-HP Hôpital de la Pitié Salpêtrière

April 1, 2022

Abstract

Machine learning (ML) and deep learning (DL) have been widely used for the computer-aided diagnosis (CAD) of neurodegenerative diseases The main limitation of these tools is that they have been mostly validated using research data sets that are very different from clinical routine ones: strict image acquisition protocols ensure good quality and homogeneous data, and well-defined diagnostic criteria guarantee unambiguous classification tasks. The validation on large clinical data sets is necessary to understand the performance of these tools in a real setting. Clinical data warehouses (CDW), gathering data of hundred of thousands of patients from different hospitals, allow access to such clinical data.

This PhD work consisted in applying ML/DL algorithms to data originating from the CDW of the Greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]) to validate CAD of neurodegenerative diseases. In particular, we aimed to address some of the challenges posed by the use of this type of data.

In the first work we developed, thanks to the manual annotation of 5500 images, an automatic approach for the quality control (QC) of T1-weighted (T1w) brain magnetic resonance images (MRI) from a clinical data set. QC is fundamental as insufficient image quality can prevent CAD systems from working properly. The automatic QC was able to identify images that are not proper T1w brain MRIs, to identify acquisitions for which gadolinium was injected and to rate the overall image quality.

In the second work, we focused on the homogenization of T1w brain MRIs from a CDW: heterogeneity must be reduced to avoid potential biases in downstream tasks. We proposed to homogenize such large clinical data set by converting images acquired after the injection of gadolinium into non-contrast-enhanced images using 3D U-Net models and conditional generative adversarial networks.

Lastly, we assessed whether ML/DL algorithms could detect dementia in a CDW using T1w brain MRI. We identified the population of interest using ICD-10 codes assigned during hospitalization. We compared the ability of ML/DL algorithms to detect dementia patients in a research data set and in the AP-HP CDW set. We then studied how the imbalance of the training sets, in terms of contrast injection and image quality, may bias the results and we proposed strategies to attenuate these biases.

CDW offer fantastic opportunities for the translation of CAD systems from research to clinical practice, but they still pose considerable challenges.

Résumé

L'apprentissage automatique et l'apprentissage profond ont été largement utilisés pour le diagnostic assisté par ordinateur des maladies neurodégénératives. La principale limite de ces outils est qu'ils ont été validés en utilisant des données de recherche qui sont très différents des données de routine clinique: les protocoles stricts d'acquisition d'images garantissent des données de bonne qualité et homogènes, et les critères de diagnostic bien définis garantissent des tâches de classification sans ambiguïté. La validation sur de grands ensembles de données cliniques est nécessaire pour comprendre la performance de ces outils dans un contexte réel. Les entrepôts de données de santé (EDS), qui rassemblent les données de centaines de milliers de patients de différents hôpitaux, permettent d'accéder à de telles données. Ce travail de thèse a consisté à appliquer des algorithmes d'apprentissage automatique à des données provenant de l'EDS de l'Assistance Publique-Hôpitaux de Paris (AP-HP) pour valider les outils pour le diagnostic assisté par ordinateur de maladies neurodégénératives. En particulier, nous avons cherché à relever certains des défis posés par l'utilisation de ce type de données.

Dans le premier travail, nous avons développé, grâce à l'annotation manuelle de 5500 images, une approche automatique pour le contrôle qualité des images par résonance magnétique (IRM) cérébrales pondérées en T1 provenant d'un EDS. Le contrôle qualité est fondamental car une qualité d'image insuffisante peut empêcher les systèmes de fonctionner correctement. Le contrôle qualité automatique a permis d'identifier les images qui ne sont pas de véritables IRM cérébrales pondérées en T1, d'identifier les acquisitions pour les quelles du gadolinium a été injecté et d'évaluer la qualité globale de l'image.

Dans le second travail, nous nous sommes concentrés sur l'homogénéisation des IRM cérébrales pondérées en T1 provenant d'un EDS : l'hétérogénéité doit être réduite pour éviter les biais potentiels dans les tâches en aval. Nous avons proposé d'homogénéiser ce grand ensemble de données cliniques en convertissant les images acquises après l'injection de gadolinium en images sans contraste à l'aide de modèles U-Net 3D et de réseaux antagonistes génératifs conditionnels.

Enfin, nous avons évalué si les algorithmes d'apprentissage automatique pouvaient détecter la démence dans un EDS en utilisant l'IRM cérébrale pondérées en T1. Nous avons identifié la population d'intérêt grace aux codes CIM-10 attribués pendant l'hospitalisation. Nous avons comparé la capacité des algorithmes à détecter les patients atteints de démence dans un ensemble de données de recherche et dans l'ensemble de l'EDS de l'AP-HP. Nous avons ensuite étudié comment le déséquilibre des ensembles d'entraînement, en termes d'injection de produit de contraste et de qualité d'image, peuvent biaiser les résultats et nous avons proposé des stratégies pour atténuer ces biais.

Les EDS offrent des possibilités fantastiques pour faire passer les systèmes d'aide au diagnostic de la recherche à la pratique clinique, mais ils posent encore des défis considérables.

Scientific production

First author journal papers

1. **Bottani, S.**, Burgos, N., Maire, A., Wild, A., Ströer, S., Dormont, D., Colliot, O. and the APPRIMAGE Study Group, 2022. "Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse". *Medical Image Analysis*, 75, p.102219.

2. Burgos*, N., **Bottani*, S.**, Faouzi*, J., Thibeau-Sutre*, E. and Colliot, O., 2021. "Deep learning for brain disorders: from data processing to disease treatment". *Briefings in Bioinformatics*, 22(2), pp.1560-1576. (*: joint first authorship)

Submitted first author journal papers

1. **Bottani, S.**, Thibeau-Sutre, E., Maire, A., Stroër, S., Dormont, D., Colliot, O., Burgos, N. and the APPRIMAGE Study Group, "Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation". Submitted to the SPIE Journal of Medical Imaging

2. **Bottani, S.**, Burgos, N., Maire, A., Stroër, S., Saracino, D. Dormont, D., Colliot, O., and the APPRIMAGE Study Group, "Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse". Submitted to Medical Image Analysis

Journal papers as co-author

1. Routier, A., Burgos, N., Díaz, M., Bacci, M., **Bottani, S.**, El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Jacquemont T., Lu P., Marcoux A., Moreau T., Samper-González J., Teichmann M., Thibeau-Sutre E., Vallant G., Wen J., Wild A., Habert M-O., Durrleman S., and Colliot O., 2021. "Clinica: an open-source software platform for reproducible clinical neuroscience studies". *Frontiers in Neuroinformatics*, 15.

2. Saracino, D., Géraudie, A., Remes, A.M., Ferrieux, S., Noguès-Lassiaille, M., **Bottani, S.**, Cipriano, L., Houot, M., Funkiewiez, A., Camuzat, A., Rinaldi, D.,Teichmann, M., Parientede, J., Couratier, P., Boutoleau-Bretonnière, C., Auriacombe, S., Etcharry-Bouyx, F., Levy, R., Migliaccio R., Solje E., Le Ber E and The French research network on FTD/FTD-ALS and PREV-DEMALS study groups, 2021, "Primary progressive aphasias associated with C9orf72 expansions: Another side of the story". *Cortex*, 145, pp.145-159.

3. Saracino, D., Ferrieux, S., Noguès-Lassiaille, M., Houot, M., Funkiewiez, A., Sellami, L., Deramecourt, V., Pasquier, F., Couratier, P., Pariente, J., Géraudie, A., Epelbaum S., Wallon D., Hannequin D., Martinaud O., Clot F., Camuzat A., **Bottani S.**, Rinaldi, D., Auriacombe, S., Sarazin M., Didic, M., Boutoleau-Bretonnière, C., Thauvin-Robinet, C., Lagarde J., Roué-Jagot, C., Sellal, F., Gabelle, A., Etcharry-Bouyx F., Morin A., Coppola, C., Levy, R., Dubois, B., Brice, A., Colliot, O., Gorno-Tempini, M.L., Teichmann, M., Migliaccio, R., Le Ber, I, on behalf of the French Research Network on FTD/FTD-ALS, 2021. "Primary Progressive Aphasia Associated With GRN Mutations: New Insights Into the Non-amyloid Logopenic Variant". *Neurology*, 97(1), pp.e88-e102.

4. Koval, I., Bône, A., Louis, M., Lartigue, T., **Bottani, S.**, Marcoux, A., Samper-Gonzalez, J., Burgos, N., Charlier, B., Bertrand, A. and Epelbaum, S., Colliot O., Allassonnière S. and Durreleman S., 2021. "AD Course Map charts Alzheimer's disease progression". *Scientific Reports*, 11(1), pp.1-16.

5. Wen, J., Samper-González, J., **Bottani, S.**, Routier, A., Burgos, N., Jacquemont, T., Fontanella, S., Durrleman, S., Epelbaum, S., Bertrand, A. and Colliot, O., 2021." Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease". *Neuroinformatics*, 19(1), pp.57-78.

6. Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., **Bottani, S.**, Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M. and Thibeau-Sutre, E., Wen J., Wild A., Burgos, N., Dormont, D., Colliot, O. and Durrleman, S., 2021. "Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review". *Medical Image Analysis*, 67, p.101848.

7. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., **Bottani, S.**, Dormont, D., Durrleman, S., Burgos, N., Colliot, O. and Alzheimer's Disease Neuroimaging Initiative, 2020. "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation". *Medical Image Analysis*, 63, p.101694.

8. Lartigue, T., **Bottani, S.**, Baron, S., Colliot, O., Durrleman, S. and Allassonnière, S., 2020. "Gaussian Graphical Model exploration and selection in high dimension low sample size setting". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9), pp.3196-3213.

9. Marcoux, A., Burgos, N., Bertrand, A., Teichmann, M., Routier, A., Wen, J., Samper-González, J., **Bottani, S.**, Durrleman, S., Habert, M.O. and Colliot, O., 2018. "An automated pipeline for the analysis of PET data on the cortical surfac"e. *Frontiers in Neuroinformatics*, 12, p.94.

10. Samper-González, J., Burgos, N., **Bottani, S.**, Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J. and Bertrand, A., Bertin, H., Habert, M-O., Durrleman, S., Evgeniou, T., Colliot, O., 2018. "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data". *NeuroImage*, 183, pp.504-521.

Submitted journal papers

1. Couvy-Duchesne, B., **Bottani, S.**, Camenen, E., Fang, F., Fikere, M., Gonzalez-Astudillo, J., Harvey, J., Hassanaly, R., Kassam, I., Lind, P., Liu, Q., Lu, Y., Nabais, M., Rolland, T., Sidorenko, J., Strike, J., Wright M. "Main existing datasets for open data research on humans".

2. 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.

Conference papers

1. **Bottani, S.**, Thibeau-Sutre, E., Maire A., Ströer, S., Dormont, D., Colliot, O., Burgos, N. and the APPRIMAGE Study Group, 2022, "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*, San Diego, United States

2. Samper-Gonzalez, J., Burgos, N., **Bottani, S.**, Habert, M.O., Evgeniou, T., Epelbaum, S. and Colliot, O., 2019. "Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data". In *SPIE - Medical Imaging 2019* (Vol. 10949, p. 109490V).

3. Samper-Gonzalez, J., Burgos, N., **Bottani, S.**, Habert, M.O., Evgeniou, T., Epelbaum, S. and Colliot, O., 2018. "Three simple ideas for predicting progression to Alzheimer's disease". In *8^{th} International Workshop on Pattern Recognition in Neuroimaging.*

Conference abstracts

1. Routier, A., Marcoux, A., Melo, M.D., Samper-González, J., Wild, A., Guyot, A., Wen, J., Thibeau-Sutre, E., **Bottani, S.**, Durrleman, S. and Burgos, N., Colliot, O., 2020. "New longitudinal and deep learning pipelines in the Clinica software platform." In *OHBM 2020 - Organization for Human Brain Mapping Annual Meeting 2020*.

2. Samper-Gonzalez, J., Burgos, N., **Bottani, S.**, Habert, M.O., Evgeniou, T., Epelbaum, S. and Colliot, O., 2019. "Predicting progression to Alzheimer's disease from clinical and imaging data: a reproducible stud"y. In *OHBM 2019 - Organization for Human Brain Mapping Annual Meeting 2019*.

3. Wen, J., Samper-González, J., Routier, A., **Bottani, S.**, Durrleman, S., Burgos, N. and Colliot, O., 2019. "Beware of feature selection bias! Example on Alzheimer's disease classification from diffusion MRI". In *OHBM 2019 - Organization for Human Brain Mapping Annual Meeting 2019*.

4. Routier, A., Marcoux, A., Melo, M.D., Guillon, J., Samper-González, J., Wen, J., **Bottani, S.**, Guyot, A., Thibeau-Sutre, E., Teichmann, M. and Habert, M.O., Durrleman, S., Burgos, N., Colliot, O., 2019. "New advances in the Clinica software platform for clinical neuroimaging studies". In *OHBM 2019 - Organization for Human Brain Mapping Annual Meeting 2019*.

5. Wen, J., Thibeau, E., Samper-González, J., Routier, A., **Bottani, S.**, Dormont, D., Durrleman, S., Colliot, O. and Burgos, N., 2019. "How serious is data leakage in deep learning studies on Alzheimer's disease classification?". In *OHBM 2019 - Organization for Human Brain Mapping Annual Meeting 2019*.

6. Samper-Gonzalez, J., **Bottani, S.**, Burgos, N., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J. and Bertrand, A., Burtin, H., Habert, M-0., Durrleman S., Evgeniou, T., Colliot, O., 2018. "Reproducible evaluation of Alzheimer's Disease classification from MRI and PET data". In *OHBM 2018 - Organization for Human Brain Mapping Annual Meeting 2018*.

7. Wen, J., Samper-Gonzalez, J., **Bottani, S.**, Routier, A., Burgos, N., Jacquemont, T., Fontanella, S., Durrleman, S., Bertrand, A. and Colliot, O., 2018. "Comparison of DTI Features for the Classification of Alzheimer's Disease: A Reproducible Study". In *OHBM 2018 - Organization for Human Brain Mapping Annual Meeting 2018*.

8. Wen, J., Samper-Gonzalez, J.,**Bottani, S.**, Routier, A., Burgos, N., Jacquemont, T., Fontanella, S., Durrleman, S., Bertrand, A. and Colliot, O., 2018, July. "Using diffusion MRI for classification and prediction of Alzheimer's Disease: a reproducible study". In *AAIC 2018 - Alzheimer's Association International Conference*.

Talks

1. Maire* A., **Bottani* S.**, Jacob Y., Ströer S., Burgos N., Colliot O., Dormont D., Hilka M., 2021. "Apports de la Plateforme Données Massive AP-HP pour la recherche

en IA: le projet APPRIMAGE". In *JFR 2021 - Journées Francophones de Radiologie 2021.* (*: joint first authorship)

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AP-HP	Assistance Publique-Hôpitaux de Paris
BA	Balanced accuracy
BIDS	Brain Imaging Data Structure
BDP	Big Data Platform
CAD	Computer-aided diagnosis
CDW	Clinical Data Warehouse
CN	Cognitively normal
CNN	Convolutional Neural Network
CSF	Cerebrospinal fluid
CV	Cross-validation
D	Dementia
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
EDS	Entrepôt de Données de Santé - Clinical data warehouse in French
EHR	Electronic Health Records
FLAIR	Fluid attenuated inversion recovery
GAN	Generative Adversarial Networks
GM	Gray matter
HDFS	Hadoop Distributed File System
ICD	International Classification of Disease
IQM	Image Quality Metrics
I&D	Innovation and data division
MAE	Mean absolute error
ML	Machine learning
MNI	Standard space of the Montreal Neurological Institute
MRI	Magnetic resonance imaging
NDL	Not dementia with lesions
NDNL	Not dementia no lesions
PACS	Picture Archiving and Communication Systems
PSNR	Peak signal-to-noise ratio
QC	Quality Control
SR	Straight Reject
SSIM	Structural similarity metrix
SVM	Support Vector Machines
T1w	T1-weighted magnetic resonance imaging
TIV	Tissue intracranial volume
WM	White matter

Introduction

Machine learning (ML) is a field of artificial intelligence that allows computers to perform tasks learning by themselves the relevant decision rules by analysing training data sets. Classical ML models, such as support vector machines or random forests, are based on features pre-extracted from the training data thanks to expert's knowledge, while more recent deep learning (DL) models are able to extract suitable features by themselves. ML and DL models have been applied to medical images to perform many tasks, including computer-aided diagnosis (CAD). Such tools can assist doctors in the study of various diseases as they can extract patterns of the diseases for their detection.

This thesis focuses on the CAD of neurological diseases, and more particularly neurodegenerative dementias, using ML and DL models. In this introduction, we will first give an overview of recent works that have been published on the CAD of brain disorders using DL and we will then focus on the CAD of neurodegenerative diseases and its limitations. A major limitation is that most of the existing studies were performed on research data, which can largely differ from data acquired in clinical routine. We will introduce in the third section of the introduction how clinical data warehouses (CDW) could help translate CAD tools to clinical practice. The introduction will end with a brief summary of the thesis contributions and with the outline of the manuscript.

Computer-aided diagnosis of brain disorders

Machine learning has been used for many years for the CAD of brain disorders (Rathore et al., 2017; Pellegrini et al., 2018; De Filippis et al., 2019; Moon et al., 2019; Burgos and Colliot, 2020). The use of DL is more recent and was analyzed in a review paper published in Briefings in Bioinformatics (Burgos et al., 2021). The following section was extracted from this review, to which I contributed as joint first author.

Most of the studies on disease detection have dealt with Alzheimer's and Parkinson's diseases (Noor et al., 2019; Gautam and Sharma, 2020). This is partly due to the public availability of large data sets such as from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Parkinson's Progression Markers Initiative (PPMI) cohorts. The aim of disease detection and diagnosis can be to differentiate healthy controls from subjects with a disease or to distinguish between different diseases (disease recognition), but also, once a disease has been singled out, to quantify its severity or to differentiate between subtypes.

Many studies focusing on Alzheimer's disease aim to differentiate healthy controls from subjects with dementia, a relatively easy task, that is useful for assessing and benchmarking classification methods but with little clinical relevance (Wen et al., 2020). The most commonly used approach is an end-to-end CNN for classification (Gautam and Sharma,

2020; Noor et al., 2019; Wen et al., 2020). Wen et al., 2020 provided a review on the use of CNN for Alzheimer's disease classification showing that results obtained with CNN are comparable to those obtained with traditional machine learning techniques. Nevertheless, other approaches exist. Silva et al., 2019 used a CNN for feature extraction only and not classification. A variational autoencoder was used by Choi et al., 2019 to detect anomalies in positron emission tomography (PET) images, thereby providing a score of abnormality used to identify Alzheimer's disease patients. The main imaging modalities used for Alzheimer's disease classification are T1w MRI and ^{18}F-fluorodeoxyglucose PET (Gautam and Sharma, 2020), but others, for example amyloid PET, have been used (Punjabi et al., 2019). Other types of data, such as speech data (Chien et al., 2019), also bring meaningful information.

Several studies have dealt with classification of Parkinson's disease patients vs controls. This can be achieved using single photon computed tomography (Choi et al., 2017), neuromelanin sensitive MRI (Shinde et al., 2019), connectivity graphs computed from diffusion MRI (Zhang et al., 2019b) or handwriting images (Afonso et al., 2019; Naseer et al., 2020).

Differentiating healthy controls from subjects with a psychiatric disease is also a research question widely addressed. Depression was studied using electroencephalograms as input (Acharya et al., 2018; Yang et al., 2018; Yang et al., 2020). To overcome the lack of patient data, transfer learning was used in the work of Banerjee et al., 2019, where they classified patients with post-traumatic stress disorder using a deep belief network model. Classification of schizophrenia versus healthy control is performed in several studies with a sparse multilayer perceptron (Zeng et al., 2018), or a CNN combined with a pre-trained convolutional autoencoder (Oh et al., 2019), and classification of bipolar disorders versus healthy controls is studied in (Campese et al., 2019) with a 3D CNN. The public availability of the Autism Brain Imaging Data Exchange data set has propelled research on autism spectrum disorder. For example, functional MRI data were used in (Eslami et al., 2019; Xiao et al., 2018) to distinguish patients with an autism spectrum disorder from controls using a CNN. Other works used genomic data with a neural network (Ghafouri-Fard et al., 2019) or eye tracking data with a long short-term memory (LSTM) (Li et al., 2019b). A multimodal approach for the integration of functional and structural MRI was proposed by Zou et al., 2017 for the classification of attention deficit hyperactivity disorder versus healthy children using a 3D CNN. Finally, Zhang et al., 2019a identified patients with conduct disorder using T1w MRI and a 3D variation of AlexNet.

The control vs disease task is not limited to neurodegenerative and psychiatric disorders. Classification of epileptic subjects vs HC has been addressed by Aoe et al., 2019 who built a CNN called M-Net from magnetoencephalography signals. Fu et al., 2019 performed natural language processing by using a CNN to detect individuals with silent brain infarction using radiological reports, as early detection can be useful for stroke prevention. Using different features extracted from functional MRI data, Yang et al., 2018 proposed to distinguish between migraine patients and healthy controls (but also between two subtypes of migraine) using an Inception CNN. Finally, MR angiography was used to detect cerebral aneurysms (Nakao et al., 2018; Ueda et al., 2019) using a custom CNN (Nakao et al., 2018) or a ResNet-18 (Ueda et al., 2019).

Few studies have explored differential diagnosis with DL. Wada et al., 2019 classified

Alzheimer's disease versus Lewy body dementia using a 2D CNN while Huang, Wu, and Su, 2019 classified bipolar disorder and unipolar depression using a CNN followed by an LSTM, both with attention mechanisms.

Neurological disorders can be complex and several works aim to identify known disease subtypes or quantify their severity. This is particularly the case in oncology. In the brain cancer domain, most of the studies (Akkus et al., 2017; Ge et al., 2018; Li et al., 2017) focused on low-grade gliomas. Low-grade gliomas are less aggressive tumors with better prognosis compared to high-grade gliomas. In low-grade gliomas, the genetics of the tumor can provide prognostic information, but this analysis requires biopsy, which is an invasive procedure. To rely only on non-invasive examinations, these studies proposed DL methods to distinguish between different genetic classes based on different structural MRI modalities. Ge et al., 2018 performed two classification tasks: low-grade vs high-grade gliomas (tumor grading) and low-grade gliomas with or without 1p19q codeletion, a biomarker predictive of chances of survival (tumor subtyping). They used a 2D CNN on T1w, T2w and FLAIR MRI slices. Akkus et al., 2017 also performed the same tumor subtyping task using 2D CNN on T2w and post-contrast T1w MRI and Li et al., 2017 predicted the mutation status of isocitrate dehydrogenase 1 in low-grade gliomas, using a 2D CNN associated with a support vector machine applied to post-contrast T1w and FLAIR MRI. Additionally, Hollon et al., 2020 classified images of biopsies (stimulated Raman histology) between 13 common subtypes covering 90% of the diversity of brain tumors with a 2D CNN. They compared their workflow with pathologists interpreting conventional histologic images and achieved a similar diagnostic accuracy for a large gain in diagnostic time (less than 2.5 minutes versus 30 minutes). They also successfully identified rare phenotypes as they did not belong to any of their predefined classes, though they could not distinguish between them.

The distinction of disease subtypes has also been explored for other neurological disorders. Choi et al., 2020 aimed to differentiate Parkinson's disease patients with dementia from those without dementia following a transfer learning strategy using a CNN initially trained to distinguish controls vs Alzheimer's disease patients. In (Zhang et al., 2019b) the objective was to identify subtypes of Parkinson's disease progression using a LSTM with clinical and imaging data. Kiryu et al., 2019 aimed to differentiate Parkinsonian syndromes. Different clinical profiles of patients with multiple sclerosis were identified in (Marzullo et al., 2019) using graphs extracted from diffusion MRI and graph CNN. Non-contrast head CT scanners were used by Ye et al., 2019 for the detection of intracranial hemorrhage and its five subtypes. They used a CNN to identify the presence or absence of intracranial hemorrhage and a recurrent neural network for the classification of intracranial hemorrhage subtypes. In the context of epilepsy, Acharya et al., 2018 used a CNN to distinguish three classes of electroencephalography signal: normal, preictal and seizure. San-Segundo et al., 2019 studied two different tasks using a CNN applied to electroencephalography: classification of epileptic vs non-epileptic brain areas and detection of epileptic seizures.

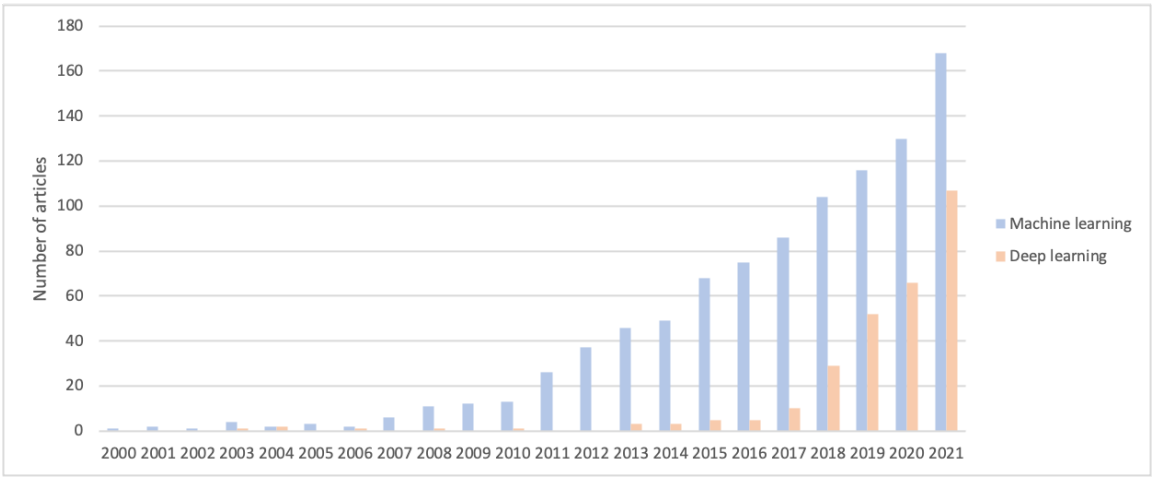

FIGURE 1: Number of articles presenting computer-aided diagnosis approaches based on machine learning and deep learning applied to neurodegenerative diseases published over the years according to PubMed (queries are available in Appendix A).

Computer-aided diagnosis of neurodegenerative diseases: current challenges

Among the brain disorders, this thesis focuses on neurodegenerative diseases. Neurodegenerative diseases are characterized by a progressive degeneration of brain tissues, mainly of gray matter, leading to cognitive deficits. They are mainly diagnosed following a neurological examination including neuropsychological tests, but imaging can also play an important role, in particular T1-weighted (T1w) brain magnetic resonance imaging (MRI) that enables the assessment of atrophy.

As highlighted in Figure 1, the number of papers presenting CAD systems for neurodegenerative diseases has been rising for the past fifteen years. This increase can be associated with the appearance of public data sets, as already mentioned in the previous section, and in advances in ML/DL research. Among the public data sets the most important are ADNI (Alzheimer's Disease Neuroimaging Initiative) [1], AIBL (Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing) [2] including participants with Alzheimer's disease and mild cognitive impairment, OASIS (Open Access Series Of Imaging Studies) [3] with Alzheimer's disease patients, PPMI (Parkinson's Progression Markers Initiative) [4] including Parkinson's disease patients and NIFD (Frontotemporal lobar degeneration neuroimaging initiative) [5] with fronto-temporal dementia subjects. Public research data sets are easy to access (data can be downloaded directly from the study websites) and use (data are homogenized and of good quality).

Most of the works in the literature share the same limitations: they are developed using only research data sets (e.g. (Wen et al., 2020; Samper-González et al., 2018; Gautam and Sharma, 2020; Koikkalainen et al., 2016)) and they are rarely based on clinical data.

[1] http://adni.loni.usc.edu/
[2] https://aibl.csiro.au/
[3] https://www.oasis-brains.org/
[4] https://www.ppmi-info.org/
[5] https://ida.loni.usc.edu/home/projectPage.jsp?project=NIFD

Furthermore the few works based on clinical data (Morin et al., 2020; Chagué et al., 2021) use data sets with a small sample size and which often come from one or a few highly specialized medical centers, thus not reflecting the reality of clinical routine in general. The generalization of the ML/DL models in a real clinical setting is not straightforward if they are trained using small clinical data sets or research data sets. In fact, the size of the data set must be large in order to ensure the generalization of the ML/DL models and research data are very different from clinical data. Indeed the quality of the images coming from research data sets is guaranteed by strict research protocols, where the acquisition parameters are harmonized among the different sites, the number of scanners employed is limited and the classes of the diseases are homogeneous. Images of clinical data sets are heterogeneous, they have different qualities, acquisition parameters are not harmonized, there is a larger number of scanners and they are acquired during a long time span. All these factors affect the images used as input for the ML/DL models. Accordingly, performance of the classifiers can greatly vary.

Research in the use of CAD for neurodegenerative diseases must undertake a step forward and focus on the validation of the ML/DL models using clinical routine data sets in order to prove their utility in a clinical setting. Nowadays challenge is the translation of the research work to the clinical environment. The principal limitation is the difficult access to clinical data. Due to privacy reasons and to confidentiality among the patients and the clinicians, they are very hard to access and they cannot be shared.

Use of clinical data warehouse for the development of CAD in a clinical setting

Thanks to the technology advancement in the domain of big data and the awareness of the need for more clinical data to validate the algorithms, clinical data warehouses (CDW) have been developed. CDW gather electronic health records, which can assemble demographics data, results from biological tests, prescribed medications and images acquired in clinical routine, sometimes for millions of patients from multiple sites. CDW allow for large-scale epidemiological studies and they offer unique data sets to validate ML and DL algorithms in a clinical context.

CDW offer a great opportunity for researchers to validate their algorithms but the use of this type of data is not straightforward. As mentioned above, the images can be very heterogeneous because of various acquisition setups. This heterogeneity may not be evenly represented among the diagnostic classes, which could bias the results of CAD systems. In addition, clinical data are not systematically checked by a neurologist.

In this PhD work, we aimed to address some of the challenges posed by CDW for developing and validating ML/DL algorithms for neurological diseases from neuroimaging data. More specifically, we worked with T1w brain MRI data originating from the CDW of the greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]). Our contributions concern image quality control, image harmonization as well as validation of CAD systems.

Contributions

This thesis includes three main contributions.

The quality of images coming from clinical data warehouses can greatly vary, which can prevent classification algorithms from working properly. Quality control is thus a fundamental step before training and evaluating machine learning approaches on clinical routine data. We proposed a framework for the automatic quality control of brain T1w MRI. Thanks to the manual annotation of 5500 images, we trained and validated convolutional neural networks that are able to discard images that are of no interest, recognise the injection of a gadolinium-based contrast agent and rate the overall image quality. It is, to our knowledge, the first brain imaging paper from the AP-HP CDW.

The heterogeneity of the T1w brain MRI must be reduced to avoid potential biases. In the second contribution, we proposed to homogenize such large clinical data set by converting images acquired after the injection of gadolinium into non-contrast-enhanced images using 3D U-Net models and conditional generative adversarial networks.

The third contribution consists in an experimental study that aimed to assess whether machine learning algorithms could detect dementia in a clinical data warehouse using anatomical brain magnetic resonance imaging. At first we identified the population of interest by exploiting the diagnostic codes from the 10^{th} revision of the International Classification of Diseases that are assigned to each patient. Then we assessed the ability of machine and deep learning classification algorithms to detect neurodegenerative dementia in a research data set and in the CDW set. We studied how the imbalance of the training sets, in terms of contrast injection and image quality, may bias the results and we proposed strategies to attenuate these biases.

Outline of the manuscript

The manuscript has the following structure:

- In **Chapter 1** we describe the clinical data warehouse of the Paris Greater Area (AP-HP). After an overview of their structure and how data (imaging and clinical) are collected, we focus on the project called APPRIMAGE within which the present thesis work was carried out.

- In **Chapter 2** we present the automatic quality control framework developed to classify images which are not proper 3D T1w brain MRI, images injected with gadolinium and to rate the overall image quality.

- In **Chapter 3** we demonstrate how image translation models, such as U-Net or conditional generative adversarial networks, may be used to homogenize MRI sequences, in particular to transform images with gadolinium into images without gadolinium.

- In **Chapter 4** we evaluate computer-aided diagnosis tools to classify patients with dementia from the CDW: we compare the performance on the CDW with the performance obtained on a research data set and we show how the characteristics of a training data set could bias the analysis.

- Finally, in the Conclusion and Perspectives chapter, we discuss our results and provide potential future research directions.

Chapter 1

Clinical data warehouse of the Greater Paris university hospitals

In this PhD thesis, we relied on 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.

In this chapter, we first provide some general information about the AP-HP CDW (Section 1.1). We then describe the APPRIMAGE project, within which this PhD was carried out (Section 1.2). We finally describe the different data management procedures that were carried out as part of the present PhD thesis as well as the resulting dataset (Section 1.3).

1.1 Clinical data warehouse of the Greater Paris area

One of the first CDW in France was launched in 2017 by the AP-HP, which gathers 39 hospitals of the Greater Paris area (Daniel and Salamanca, 2020). AP-HP obtained the authorization of the CNIL in 2017 (*Commission Nationale de l'informatique et des Libertés*, the French regulatory body for data collection and management) to share data for research purposes in compliance with the MR004 reference methodology (Daniel and Salamanca, 2020). The MR004 reference controls data processing for the purpose of studying, evaluating and/or researching that does not involve human patients (in the sense of not involving an intervention or a prospective collection of research data in patients that would not be necessary for clinical evaluation, but which allows retrospective use of data previously acquired in patients). The goals of the CDW are the development of decision support algorithms, the support of clinical trials and the promotion of multi-centre studies.

According to French regulation, and as authorised by the CNIL, patients' consent to use their data in the projects of the CDW can be waived as these data were acquired as part of the clinical routine care of the patients. At the same time, AP-HP committed to keep patients updated about the different research projects of the CDW through a portal on the internet (`https://eds.aphp.fr/recherches-en-cours`) and individual information is systematically provided to all the patients admitted to the AP-HP. In addition, a retrospective information campaign was conducted by the AP-HP in 2017: it involved around

500,000 patients who were contacted by e-mail and by postal mail to be informed of the development of the CDW.

Accessing the data is possible with the following procedure. A detailed project must be submitted to the Scientific and Ethics Board of the AP-HP. If the project participants are external to AP-HP, they have to sign a contract with the Clinical Research and Innovation Board (*Direction de la Recherche Clinique et de l'Innovation*). The project must include the goals of the research, the different steps that will be pursued, a detailed description of the data needed, of the software tools necessary for the processing, and a clear statement of the public health benefits.

Once the project is approved, the research team is granted access to the Big Data Platform (BDP), which was created by a sub-department of the IT of the AP-HP, called *Innovation and Data Division* - I&D- (in French *Pôle Innovation et Données*). The BDP is a platform internal to the AP-HP where data are collected and that external users can access to perform all their analyses, in accordance with the CNIL regulation. It is strictly forbidden to export any kind of data and each user can access only a workspace that is specific to their project. Each person of the research team can access the BDP with an AP-HP account after two-factor authentication. If the research team includes people that are not employed by the AP-HP, a temporary account associated to the project is activated.

1.1.1 Data organization within the Big Data Platform

The CDW is composed of electronic health records (EHR) gathered using different software tools installed in the hospitals (i.e. PACS for imaging data and ORBIS for clinical data). The role of the I&D is to gather all the data of the projects from the hospitals' software tools and to make them available for the users in the BDP. The I&D department of the AP-HP created an internal PACS (called "research PACS") where they copied data from each hospital's software tools.

Once the data are gathered in the research PACS, they are stored, for long-term use, in the BDP after having been pseudonymized by the AP-HP I&D department.

Technology-wise, the BDP runs under a Hadoop big data framework[1]. As such, data are stored on HDFS which is the Hadoop Distributed File System. Data on HDFS can be queried/processed using only Hadoop tools, such as HiveQL[2] or Spark[3], which were installed in the cluster machines of the BDP. The BDP cluster includes machines with CPUs and/or GPUs, where programming languages such as Python/R are available. Research teams can access the cluster machines and these tools through a JupyterLab environment. All the elements described and their interactions are presented in Figure 1.1.

1.2 The APPRIMAGE project

The APPRIMAGE project, led by the ARAMIS team (current AP-HP PI: Didier Dormont; initial AP-HP PI: Anne Bertrand, deceased March 2nd 2018) at the Paris Brain Institute, was

[1] https://hadoop.apache.org
[2] https://hive.apache.org
[3] https://spark.apache.org

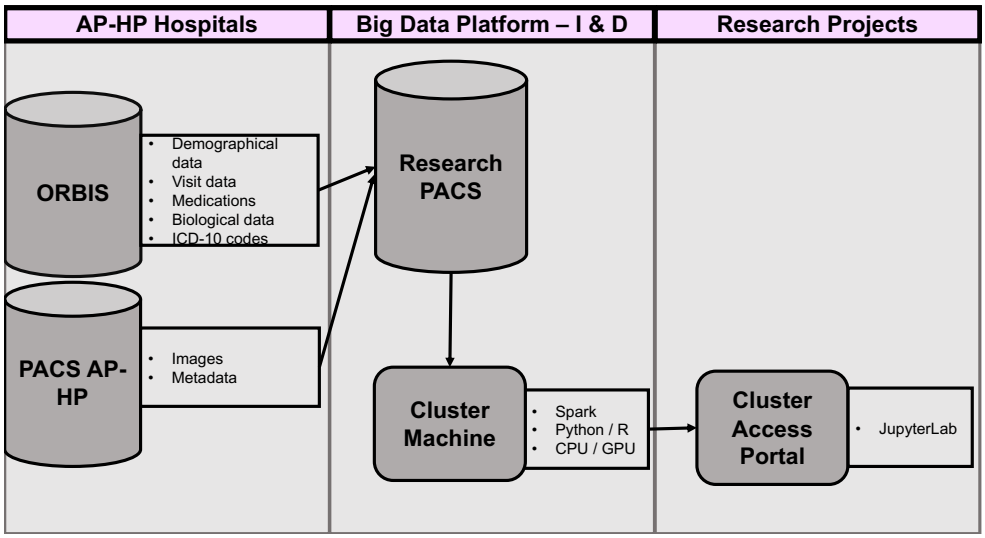

FIGURE 1.1: Workflow illustrating how data from the AP-HP hospitals are accessed by research teams passing through the Big Data Platform administered by the AP-HP I&D department. Data are gathered from hospital software tools such as ORBIS (clinical data) or PACS (imaging data), and copied to the "research PACS". Data are stored in a HDFS disk accessible through Hadoop tools in the cluster machines. Research teams can connect to cluster machines and access data through a JupyterLab.

approved by the Scientific and Ethics Board of the AP-HP in 2018. It aims at developing and validating algorithms that predict neurodegenerative diseases from structural brain magnetic resonance images (MRI), using a very large clinical data set. The project inclusion criteria were: patients aged more than 18 years and having at least one T1-weighted (T1w) brain MRI. For these different patients, the project required access to T1w and fluid-attenuated inversion recovery (FLAIR) MRI data, socio-demographic and clinical data, biological data when available, radiological and hospital reports.

In order to define the population of the project, the first step was the identification of all the images of interest. The I&D department listed all the DICOM attributes from the hospital PACS referring to MRI data. A neuroradiologist part of the APPRIMAGE project manually selected the DICOM attributes limited to those referring to 3D T1w brain MRI. More details are provided in Chapter 2. In this way a first selection of the cohort was created, which consisted of around 130,000 patients and 200,000 3D T1w brain MRIs acquired from 1980 to nowadays in the 39 hospitals of the AP-HP.

1.3 Data set and data management for the present PhD project

Data available in the cluster machines are stored in HDFS and accessible through Hadoop tools. We used HiveQL in order to collect data of interest and we saved them locally on the NAS (i.e. network-attached storage, a file-level data storage server connected to a network). HiveQL is a Hadoop tool designed to process data in HDFS in a structured form. Data in HDFS can be seen in the form of Hive tables. Once data were in the NAS, we could process them on CPUs and GPUs using Python and the software tools installed.

1.3.1 Software installation

The project was based on Python and we installed several specific libraries, the most important ones being: pytorch (Paszke et al., 2019), scikit-learn (Pedregosa et al., 2011), nilearn (Abraham et al., 2014), pydicom (Mason, 2011), ipywidgets[4]. Regarding the neuroimaging software tools, we installed in the BDP the following: Clinica[5] (Routier et al., 2021), ANTs (Avants et al., 2014), SPM standalone (Ashburner and Friston, 2005) and dcm2niix (Li et al., 2016).

1.3.2 Imaging data

Imaging data are stored in the medical PACS of the different hospitals of the AP-HP. The creation of the research PACS was necessary to preserve the medical PACS and ensure that the original images do not become corrupted. To avoid overloading the medical PACS, I&D could copy a limited number of images per day into the research PACS. Images of the APPRIMAGE project were made available by batch while stored in HDFS. They can be seen as Hive tables. In Hive tables, each line represents a single DICOM file. The columns of the Hive table are the following: `series uid` (unique id of the series representing a single image), `study uid` (unique id representing the whole study during which the sequence was acquired; for a single study one can have several series/images), `patient num` (unique id of the patient), `visit num` (unique code of the visit during which the study was undertaken), `dicom data` (binary file with all the DICOM data). DICOM in the research PACS are pseudonymized: information about the patient such as name, age, sex, weight as well as information about the physicians who requested and analysed the results of the examination are erased, and the examination date is shifted of a random amount of time (from 1 to 10 years). Note that the same shifting is applied to all the dates of the clinical data.

1.3.2.1 Difficulties encountered in obtaining exploitable data

For about a year and half, we worked closely with the I&D department in order to obtain exploitable 3D T1w brain MRI. We encountered two different types of problems: the conversion from DICOM to NIfTI format was not possible with dcm2niix (Li et al., 2016), nor the previous version of the software called dcm2nii, because the information about the position of the patient had been erased in the DICOM header, or the conversion to NIfTI worked but a large part of the brain was always missing because of missing DICOM slices (detected also by the software tools used for the conversion). The I&D department released two versions of their research PACS and three versions of the pseudonymization procedure in 18 months. Every time they did a modification, we converted and visually checked around 1,000 images to give them a feedback.

1.3.2.2 Images currently available

Once the two main problems described above were solved, two batches of images were made available.

[4] https://ipywidgets.readthedocs.io
[5] www.clinica.run

- Batch 1 contains around 11,000 3D T1W brain MRI. Images were randomly sampled from all the hospitals of the AP-HP and the different MRI machines. They were used for the study presented in Chapter 2 about quality control and for the study in Chapter 3 about feature homogenization.
- Batch 2 contains the 3D T1w brain MRI of the patients hospitalized and registered in ORBIS (more details in the next section). They were used for the study presented in Chapter 4 about the detection of patients with dementia.

1.3.2.3 Visualization of the images

One of the main limitations of the BDP is the absence of a viewer for medical images. We used the tools available in Jupyter Notebooks for the visualization. Nilearn (Abraham et al., 2014) is a popular Python package for the statistical analysis and visualization of brain imaging data: we used the function called "plot_anat" to visualize the slices of the brain in the notebooks. Nilearn allows choosing the position of the slices, so it could be adapted to our needs: we chose to visualize the central slices of the brain after the spatial normalization to the MNI space. The output was saved in PNG format to speed up the uploading of later views. Visualization of the images is essential for the project: at first, it allowed us to detect the problems in the DICOM files and to adjust the pipeline of the I&D department and then to evaluate the quality of the images or to find outliers in our data set.

To this aim, we developed a graphical interface through the Python package ipywidgets: while PNG files appear on the notebook, a text widget allows annotating the displayed PNG. The file name and the corresponding notes are saved in a text file. The interface also allows going back to the previous image in case a change in the note is needed. Figure 1.2 displays an example of the graphical interface (available on the github: `https://github.com/SimonaBottani/Quality_Control_Interface`).

1.3.3 Access to the clinical data

While DICOM files have been stored since 1980 in the medical PACS of the hospitals, following always the same structure, clinical data are stored in software tools that may vary across sites. In 2009, AP-HP decided to adopt a single software (called ORBIS) in order to store a unique electronic file for patients hospitalized at all sites. This solution has two main advantages: at first, if a patient is hospitalized in hospital A and then in hospital B, all information can be found in the electronic file, and secondly it ensures a consistency of the information that can be entered. The latter reason led to an easier extraction of the data of the research.

Thanks to this homogenization, the I&D department was able to start querying clinical data to make them available for research purposes. All clinical data are saved in a HDFS disk and they can be queried through HiveQL. They are organised as a relational database. The unique key is represented by the number identifying the patient and the number identifying the visit. Demographic data (i.e. age and sex), stored in the tables called "i2b2_visit" and "i2b2_patient", are available for all the patients registered in ORBIS (even if missing data

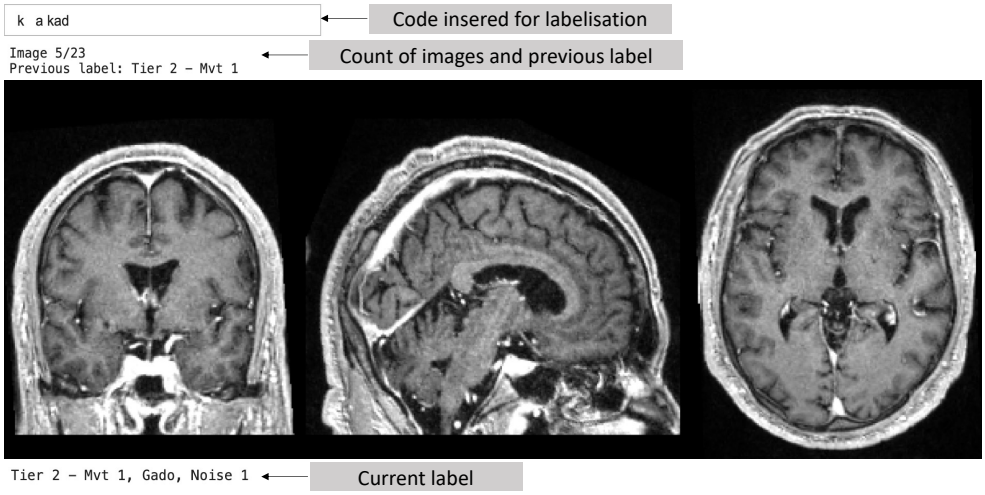

FIGURE 1.2: Example of the graphical interface for the annotation of the images in the BDP. The central slices along the coronal, sagittal and axial planes extracted from the NIfTI file were previously saved as a single PNG file to speed up download during the annotation phase. In the text box at the top the user can type the code for the labelization (a specific dictionary was created for the annotation in order to write only the essential). At the same location, the user can see their progression in the batch they are annotating, as well as the previous label. Looking at the previous label is useful to detect potential mistakes, which can be corrected by deleting the last code and going back to the previous image. At the bottom the current label is displayed. Using the space bar allows moving to the next image. Annotations are saved in a text file for further analysis.

are common), while other data (such as ICD-10 codes and list of medications) are stored in corresponding tables (such as "i2b2_observation_cim10" and "i2b2_observation_ccam") and are available only for patients registered in ORBIS and hospitalized. For outpatients (i.e. patients having a consultation at the hospital), only a fraction of the information that is stored for inpatients (i.e. hospitalized patients) are available.

For our study, we were interested only in three sociodemographic and clinical data: sex, age at the time of the visit and the ICD-10 codes to know the diagnosis related to the image. They were respectively stored in the tables called "i2b2_patient", "i2b2_visit" and "i2b2_observation_cim10". The age was calculated as the difference between the start date of the visit and the date of birth. The visit is defined as a period between the start date and the end date present in the table "i2b2_visit" and it indicates the period of hospitalization. The I&D department extracted from ORBIS all the information of the patients of the cohort. Patients were identified with the manual selection of the series corresponding to 3D T1w brain MRI.

1.3.3.1 Analysis of clinical data

Among all the patients of the cohort, clinical data were available only for 30,490 patients out of about 130,000 (i.e. 23% of the cohort). This is due to the fact that ORBIS has started being installed only since 2009 (e.g. in 2016 for the Pitié-Salpêtrière hospital) while DICOM data have been available since 1980. This fact has to be kept in mind since it represents one of the main limitations of the project.

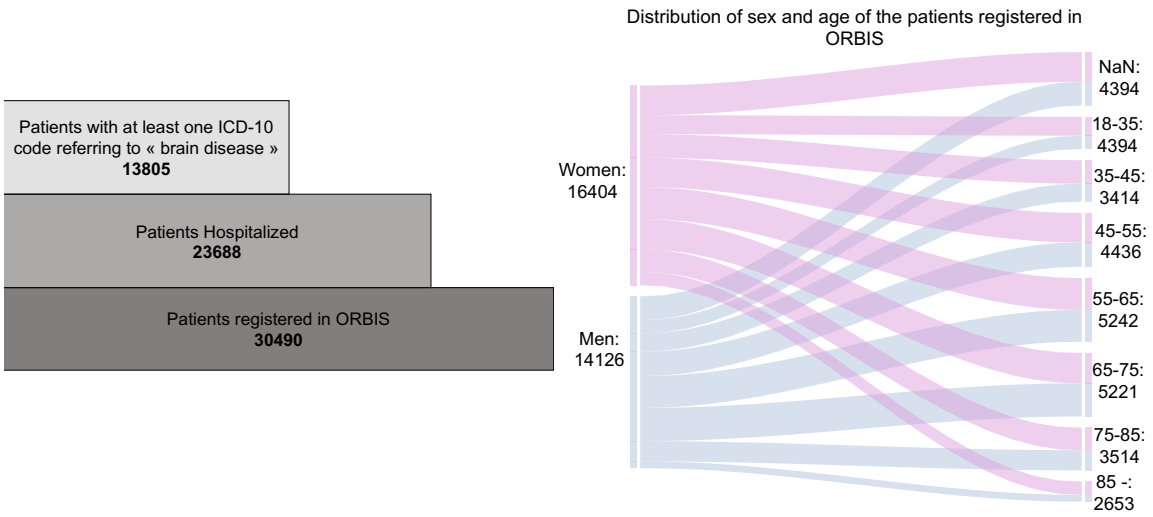

FIGURE 1.3: Left: number of patients registered in ORBIS, of hospitalized patients and of patients having at least one ICD-10 codes referring to a "brain" disease. Right: distribution of sex and age at the time of the first visit registered in ORBIS.

The right part of Figure 1.3 displays the distribution of sex and age among the patients registered in ORBIS part of our project. Each patient could have more than one visit, we thus decided to calculate the age at the first visit for this figure. Despite all the limitations in studying only this subset of data (i.e. a lot of data are lost because patients were not hospitalized in hospitals where ORBIS was installed), we can note that there are more women than men (W: 16404, M: 14126), we have more patients between 55 and 65 years old (which is consistent with the growth of neurological diseases since we calculated the age according to the first registration in ORBIS) and for 4308 of them we do not have the corresponding age because of missing data (missing date of birth or registration of the visit).

Among the 30,490 patients registered in ORBIS, 23,688 were hospitalized while the remaining went to the AP-HP hospitals just for a consultation. For the 23,688 inpatients, additional information is available, in particular the ICD-10 codes. Since our project is based on T1w MR images, with the aid of a neurologist, we made a list of the diagnoses of interest in the ICD-10 nomenclature. The purpose was to consider all the diagnoses that may lead to a T1w MRI examination. In the following, we refer to these diagnoses as the "brain" codes. Since ICD-10 is a hierarchical classification, in this part of the work we are focusing on the first letter and the first two numbers of the codes. We merged the diagnoses that are present in more than one category: F00 with G30 (Alzheimer's disease and also all the sub-categories related), and I6* and G46 (cerebrovascular disease). Among the 23,688 hospitalized patients, 13,805 patients had at least one ICD-10 code in the list of "brain" codes. The different number of patients are reported in the left of Figure 1.3.

The description of the "brain diseases", the corresponding ICD-10 codes and the number of occurrences of the diagnoses among the 13,805 patients with are reported in Table 1.1. We note that the most frequent diagnosis is vascular syndromes of brain in cerebrovascular diseases, followed by cerebral palsy, epilepsy and migraine.

These diagnoses will be in part used in the last part of our work about the classification of dementia patients in Chapter 4.

Description	ICD-10 codes	N
Vascular syndromes of brain in cerebrovascular diseases	G46	3758
Cerebral palsy and other paralytic syndromes	G80, G81, G82, G83	2266
Epilepsy	G40	2128
Migraine, other headache syndromes, transient cerebral iscahemic attacks and related syndromes	G43, G44, G45, G47	1608
Other mental disorders due to brain damage and dysfunction and to physical disease	F06	1418
Malignant neoplasm of brain	C71	1244
Inflammatory disease of the central nervous system	G01, G02, G03, G04, G05, G06, G07, G08, G09	1171
Secondary parkinsonism	G21, G22	1015
Alzheimer's disease	F00	948
Multiple sclerosis	G35	728
Unspecified dementia	F03	716
Delirium	F05	711
Injuries to the head	S01, S02, S03, S04, S05, S06, S07, S08, S09	589
Personality and behavioural disorders due to brain disease, damage and dysfunction	F07	580
Benign neoplasm of meninges	D32	568
Vascular dementia	F01	546
Status epilepticus	G41	527
Benign neoplasm of brain and other parts of central nervous system	D33	419
Other degenerative diseases of nervous system, not elsewhere classified	G31	405
Neoplasm of uncertain or unknown behaviour of brain and central nervous system	D43	405
Organic amnesic syndrome, not induced by alcohol and other psycoactive substances	F04	379
Hydrocephalus	G91	352
Sarcoidosis	D86	280
Dementia in other diseases classified elsewhere	F02	250
Hereditary ataxia, spinal muscular atrophy and related syndromes, systemic atrophies primarly afftecting central nervous system in diseases classified elsewhere, postpolio syndrome	G10, G11, G13, G14	190

Demyelinating diseases of the central nervous system	G36, G37	155
Congenital malformations of the nervous system	Q01, Q02, Q03, Q0	154
Metabolic disorders	E75, E76, E77, E78, E79, E87	141
Glaucoma, disorders of optical nerves	H46, H47, H48	130
Toxic encaphalopathy	G92	125
Neoplasm of uncertain or unknown behavior of meninges	D42	119
Human immunodeficiency virus [HIV] disease resulting in other specified diseases	B22	101
Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system	C72	99
Malignant neoplasm of meninges	C70	93
Secondary parkinsonism	G21	90
Thiamine deficiency	E51	72
Disorders ft the automatic nervous system	G90	53
Parkinsonism in diseases classified elsewhere	G22	38
Huntington disease	G10	23
Other degenerative disorders of nervous system in diseases classified elsewhere	G32	20
Eclampsia	O15	19
Unspecified organic or symptomatic mental disorder	F09	5

TABLE 1.1: List and description of the ICD-10 codes related to "brain disease". For each of them we report the number of occurrences among the 13805 patients.

Chapter 2

Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse

---

This chapter has been published in *Medical Image Analysis*:

- **Title:** Automatic Quality Control of Brain T1-Weighted Magnetic Resonance Images for a Clinical Data Warehouse
- **Authors:** Simona Bottani, Ninon Burgos, Aurélien Maire, Adam Wild, Sebastian Ströer, Didier Dormont, Olivier Colliot, APPRIMAGE Study Group
- **DOI:** doi:10.1016/j.media.2021.102219

2.1 Introduction

Structural T1-weighted (T1w) magnetic resonance imaging (MRI) is useful for diagnosis of various brain disorders, in particular neurodegenerative diseases (Frisoni et al., 2010; Harper et al., 2016). They have thus often been used as inputs of machine learning (ML) algorithms for computer-aided diagnosis (CAD) (Falahati, Westman, and Simmons, 2014; Koikkalainen et al., 2016; Rathore et al., 2017; Burgos and Colliot, 2020).

Most ML methods are trained and validated on high-quality research data (Noor et al., 2019; Choi et al., 2019; Punjabi et al., 2019): protocols for image acquisition are standardized and a strict quality control is applied (Jack et al., 2008; Littlejohns et al., 2020). However, to be applied in the clinic, ML methods need to be validated on clinical routine images. In recent years, hospitals have constituted clinical data warehouses that can contain medical images from 100,000-1,000,000 patients (Daniel and Salamanca, 2020; Amara, Lamouchi, and Gattoufi, 2020). The quality of such images can greatly vary (see Figure 2.1), since the acquisition protocols are not standardized, scanners may not be recent and patients may have moved during the acquisition. All these factors can prevent algorithms from working properly (Reuter et al., 2015; Gilmore, Buser, and Hanson, 2019). Quality control

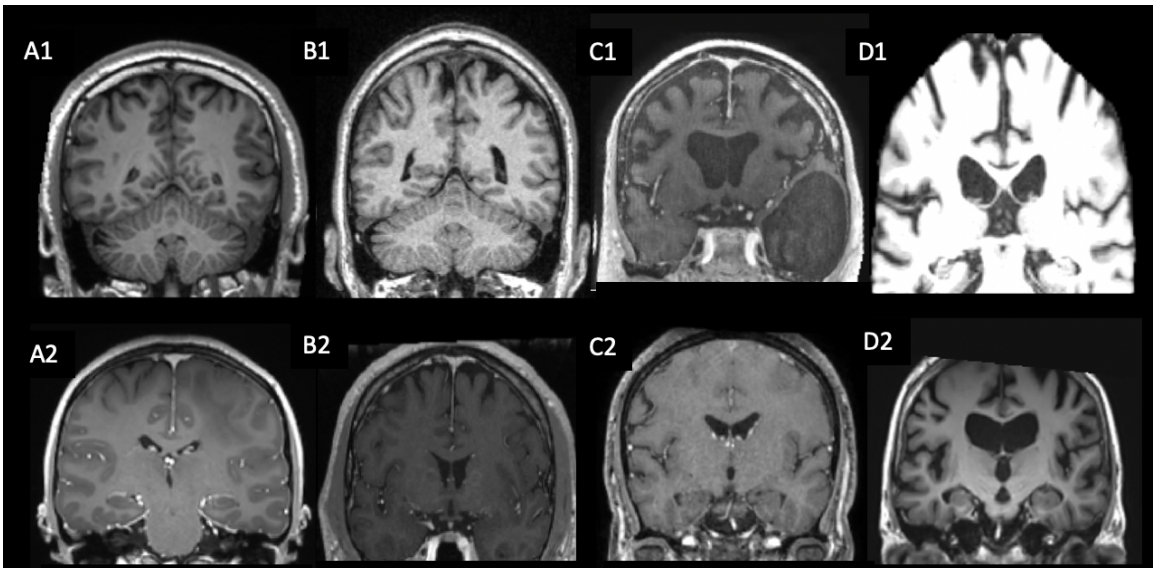

FIGURE 2.1: Examples of T1w brain images from the clinical data warehouse and the corresponding labels. A1: Image of good quality (tier 1), without gadolinium; A2: Good quality (tier 1), with gadolinium; B1: Medium quality (tier 2), without gadolinium (noise grade 1); B2: Medium quality (tier 2), with 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).

(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. Typically, clinical data warehouses can contain hundreds of thousands of samples. Even if web-based systems facilitate annotation (Kim et al., 2019; Keshavan et al., 2018), the task remains unfeasible for very large datasets. In this context, automatic QC is needed.

Several works have been proposed to enable automatic QC of cerebral MR images. The Preprocessed Connectomes Project developed a Quality Assessment Protocol[1]. The package enables the extraction of several image quality metrics (IQMs) such as the signal-to-noise ratio, the contrast-to-noise ratio or the volume of the gray and white matter. IQMs are then compared to a normative distribution obtained from three research datasets, ABIDE (Di Martino et al., 2014), CoRR[2] and NFB[3]. In the same spirit, we find (Esteban et al., 2017; Alfaro-Almagro et al., 2018; Raamana et al., 2020). These approaches propose to use the IQMs as input of a classifier for automatic QC. Esteban et al., 2017 and Alfaro-Almagro et al., 2018 developed a pipeline for the automatic QC of 3D brain T1w MRI, the first has the advantage to be an open source software (called MRIQC). Raamana et al., 2020 developed another open source software called VisualQC whose aim is the visualisation and the rating of the Freesurfer cortical segmentation output. The pipelines proposed by these works are very extensive as they require registration and segmentation steps to extract

[1] http://preprocessed-connectomes-project.org/quality-assessment-protocol
[2] http://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html
[3] http://fcon_1000.projects.nitrc.org/indi/enhanced/

features. It is not possible to assume a priori that these steps will perform well with a new unseen clinical dataset. On the contrary, it is likely that the segmentation will fail for the lowest quality images, thus making it impossible to apply the QC tool. Moreover, the extracted features may not be representative of the problems affecting clinical routine data. As proposed by Sujit et al., 2019, convolutional neural networks (CNNs) are a good option for automatic QC because they can learn features without knowing a priori which are the most adapted. A further limitation of these works is that they rely on images acquired following a well-defined research protocol. The pipeline presented in (Alfaro-Almagro et al., 2018) was developed for the large, but well-standardized, UK Biobank dataset containing mostly healthy volunteers. Esteban et al., 2017 and Sujit et al., 2019 trained their algorithms on ABIDE, a research multicenter study including patients with autism and control subjects and used another research dataset for testing. These datasets are both smaller and less realistic than a clinical dataset. In particular, Sujit et al., 2019 used 2D slices as input for the model and they classified their images only in two classes: acceptable or not acceptable.

More studies can be found if we enlarge the scope to other body parts or imaging sequences. Deep learning models have been developed for different modalities, different organs and different QC tasks: for the QC of mammograms (Kretz et al., 2020), fetal ultrasound cardiac images (Dong et al., 2019), and brain diffusion MRI (Graham, Drobnjak, and Zhang, 2018), for the detection of artefacts on cardiac MRI (Oksuz et al., 2019) and blurring on histological images (Campanella et al., 2018). Several works used a classifier trained on image quality metrics (IQM) extracted from the images: Küstner et al., 2018; Sadri et al., 2020 used this approach with a research dataset composed of different body parts and MRI sequences, Tayari et al., 2019 applied it to 3D 1H MR spectroscopy of the prostate and (Janowczyk et al., 2019) developed a tool called HistoQC for the QC of histological images. Finally, some works focused on the QC of post-processing results, mainly segmentation results. It can be done extracting IQMs from the segmented images, as proposed by Alba et al., 2018 for cardiac images, or using deep learning models as done by Robinson et al., 2018; Robinson et al., 2019 for cardiac images from the UK Biobank dataset, which contains more than 10,000 samples, and Sunoqrot et al., 2020 for prostate images.

To the best of our knowledge, there is currently no automatic QC approach dedicated to large clinical datasets of brain MRI. Our work was done using a clinical data warehouse that assembles all MRI data from all hospitals of the greater Paris area. Images come from different sites and different machines with no homogenization on the parameters, and their acquisition cover several decades. The patient may have any disease for which a brain MRI exam is required. All these factors are not present in the approaches already proposed in the literature: even when images come from different sites, the acquisition protocol is harmonized, the number of machines is limited and they are usually acquired within a few years, avoiding intrinsic problems of quality due to the progress in the technology. Additionally, the presence of different diseases such as neurodegenerative diseases, stroke, multiple sclerosis, or brain tumours, is typical of clinical datasets: they can strongly alter the structure of the brain and it may be difficult to use a specific set of features to characterize the quality of the images independently of the disease. In addition, due to security reasons,

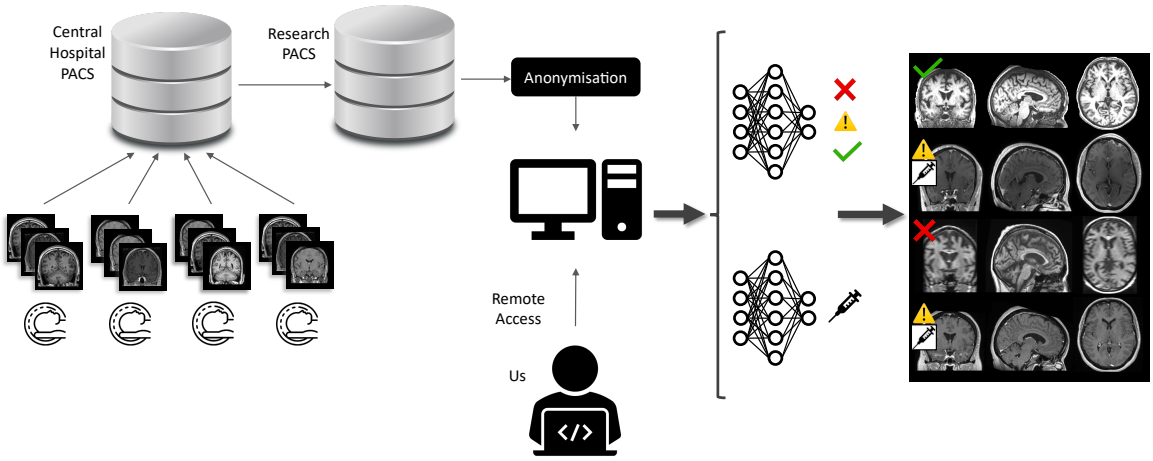

FIGURE 2.2: General workflow of the proposed QC framework. Images were acquired as part of the routine clinical care in different hospital sites and gathered in a central hospital PACS. Images relevant to our research project were copied to the research PACS and anonymized. They always remain within the hospital network that we accessed remotely. Thanks to the connection to the hospital IT network, we manually labeled the images before training and testing our deep learning models.

images from the data warehouse cannot be uploaded to a web server and we had to work in a restricted IT environment (Daniel and Salamanca, 2020).

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 data warehouses. Figure 2.2 presents an overview of our work.

2.2 Material and Methods

2.2.1 Dataset description

This work relies on a large clinical routine dataset containing all the T1w brain MR images of adult patients scanned in hospitals of the Greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]). 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.

All the images were already stored in a single central clinical PACS. Then, the data warehouse team of the AP-HP 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 pseudonymized: 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 anonymization, the date

of the exam and the date of birth were also erased from the DICOM fields. Nevertheless, as mentioned below, this information was available from another database (but not for all patients). In this other database, to increase anonymization, the date of the exam and the date of birth were also changed (they were shifted by a constant in order 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. This is summarized in Figure 2.2.

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. Standardization 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 pseudonymization 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 2.1 reports all the scanner models present in our dataset 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.

TABLE 2.1: Model name of all the scanners, grouped by manufacturer, with the corresponding magnetic field strength (T) and the number of images. Age (mean ± 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 ± std [range])	Sex (F/M)
Siemens	Aera	1.5	489	53.53 ± 18.00 [18, 95]	223 / 142
	Amira	1.5	29	47.81 ± 13.57 [19, 68]	6 / 10
	Avanto	1.5	603	52.79 ± 15.39 [18, 88]	164 / 125
	Avanto_fit	1.5	81	56.06 ± 16.64 [19, 88]	34 / 28
	Biograph mMR	3	12	-	-
	Espree	1.5	1	-	-
	Magnetom Vida	3	3	-	-
	Magnetom Essenza	1.5	11	37.2 ± 15.93 [22, 69]	1 / 9
	Sempra	1.5	3	45 ± 0 [45]	1 / 0
	Skyra	3	1851	54.31 ± 17.56 [18, 95]	708 / 692
	Spectra	3	23	55.13 ± 18.87 [22, 66]	2 / 6
	Symphony	1.5	3	-	-
	Verio	3	643	55.65 ± 17.75 [18, 92]	310 / 294
GE Healthcare	Discovery MR450	1.5	4	40.67 ± 23.57 [24, 74]	1 / 2
	Discovery MR750(w)	3	675	55.52 ± 17.49 [18, 93]	240 / 256
	Optima MR360	1.5	2	63 ± 0 [63]	0 / 1
	Optima MR450w	1.5	284	59.80 ± 18.0 [18, 95]	160 / 97
	Signa Architect	1.5	243	52.14 ± 18.63 [19, 92]	128 / 99
	Signa Artist	1.5	4	88.0 ± 1.41 [86, 89]	2 / 2
	Signa Excite	1.5	3	30.5 ± 4.5 [26, 35]	2 / 0
	Signa Explorer	1.5	1	76 ±0 [76]	1 / 0
	Signa HDx(t)	1.5	489	61.53 ±18.34 18, 94	250 / 166
	Signa Pioneer	3	1	76 ±0 [76]	0 / 1
	Signa Voyager	1.5	1	-	-
	Unknown	1.5	3	-	-
Philips	Achieva	3	21	51.0 ± 14.0 [27, 70]	5 / 2
	Ingenia	1.5	5	81.13 ± 12.20 [64, 92]	1 / 2
	Intera	1.5	7	61 ± 0 [61]	2 / 0
Toshiba	Titan	1.5	2	54.5 ± 1.5 [53, 56]	2 / 0
	Vantage Elan	1.5	3	55.5 ± 3.5 [52, 59]	1 / 1

2.2.2 Image preprocessing

The T1w MR images were converted from DICOM to NIfTI using the software dicom2niix (Li et al., 2016) and organized 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 using the SyN algorithm (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 169×208×179, 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.

2.2.3 Manual labeling of the dataset

In this section, we introduce the visual QC protocol. We describe the different characteristics noted on the images and how we created the final label for the automatic QC. Images were labeled by two trained raters and the annotation protocol was designed with the help of a radiologist.

2.2.3.1 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 (gray matter and white matter are difficult to distinguish in some parts of the image), 2: bad contrast (gray 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 2.2.

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

TABLE 2.2: Description and determination rules of the proposed quality control tiers.

2.2.3.2 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 labeled 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 optimized 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.

2.2.3.3 Consensus label

The final label used to train and validate the automatic QC is a consensus between the two raters. If the users labeled 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

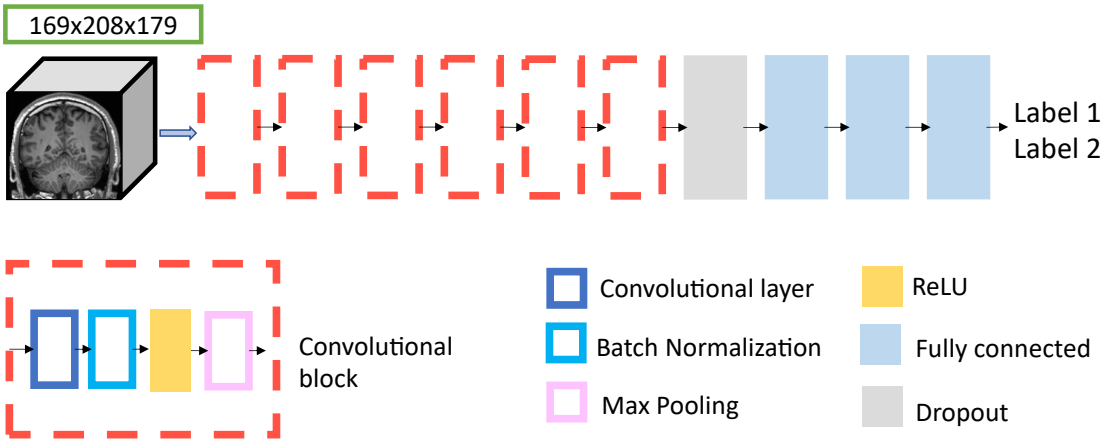

FIGURE 2.3: Architecture of the 3D CNN called Conv5_FC3. Five convolutional blocks (composed sequentially of a convolutional layer, a batch normalization layer, a ReLU and a max pooling layer) are followed by a dropout and three fully connected layers.

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.

2.2.4 Automatic quality control method

We developed an automatic QC method based on 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).

2.2.4.1 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 normalization layer, one ReLU and one max pooling. Details about architecture are represented on Figure 2.3. All the details about the parameters of the layers, i.e. the filter size, the number of filters/neurons, the stride and the padding size and the dropout rate are in the Supplementary Materials in table 2.8. 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 (Wen et al., 2020). It is available through the ClinicaDL software available on GitHub: `https://github.com/aramis-lab/ClinicaDL`.

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

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

TABLE 2.3: Weighted Cohen's kappa between the two annotators

blocks) inspired from (Jónsson et al., 2019). More details about the architectures are given Figures 2.6 and 2.7. 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 (Couvy-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 (Couvy-Duchesne et al., 2020).

2.2.4.2 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. 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).

2.3 Results

2.3.1 Manual quality control

The inter-rater agreement was evaluated using the weighted Cohen's kappa (Watson and Petrie, 2010) between the two annotators for each of the characteristics. Results are presented in Table 2.3. 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).

The distribution of the consensus labels for the 5500 patients is shown in Figure 2.4. 26% of the images are labeled as SR, 16% as tier 1, 28% as tier 2, and 30% as tier 3. Table 2.7

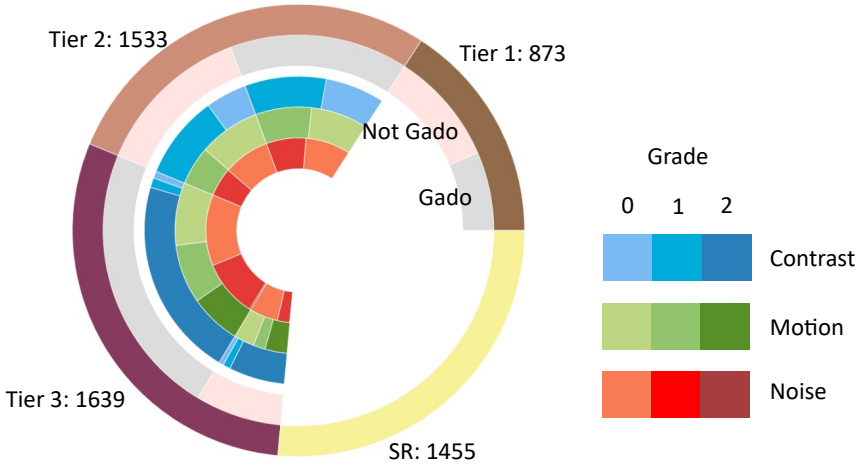

FIGURE 2.4: Distribution of the consensus labels for the whole dataset of 5500 images. Outermost circle: images in SR and in the different tiers. For every tier, we divide between images with and without gadolinium injection. For each injection status we see the grade distribution of the contrast, motion and noise characteristics.

reports the exact number of images for each category. Figure 2.1 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 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 2.4, 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

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

TABLE 2.4: 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 ± 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).

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.

2.3.2 Automatic quality control

Results obtained for the four tasks of interest by the proposed Conv5_FC3 classifier are presented in Table 2.5. 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.

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 ± 0.57	97.14 ± 0.34	83.51 ± 0.93	71.65 ± 2.15
F1 score	94.85 ± 0.41	97.04 ± 0.31	84.07 ± 1.02	74.10 ± 1.35
MCC	85.71 ± 1.11	94.00 ± 0.64	67.38 ± 2.13	42.10 ± 3.25
Sensitivity	91.83 ± 1.18	96.45 ± 0.34	79.88 ± 3.06	77.39 ± 4.29
Specificity	95.69 ± 0.53	97.82 ± 0.62	87.14 ± 3.14	65.92 ± 7.47
PPV	86.44 ± 1.43	98.33 ± 0.46	81.93 ± 3.36	83.20 ± 2.31
NPV	97.51 ± 0.35	95.39 ± 0.42	85.83 ± 1.49	57.78 ± 2.63

TABLE 2.5: 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

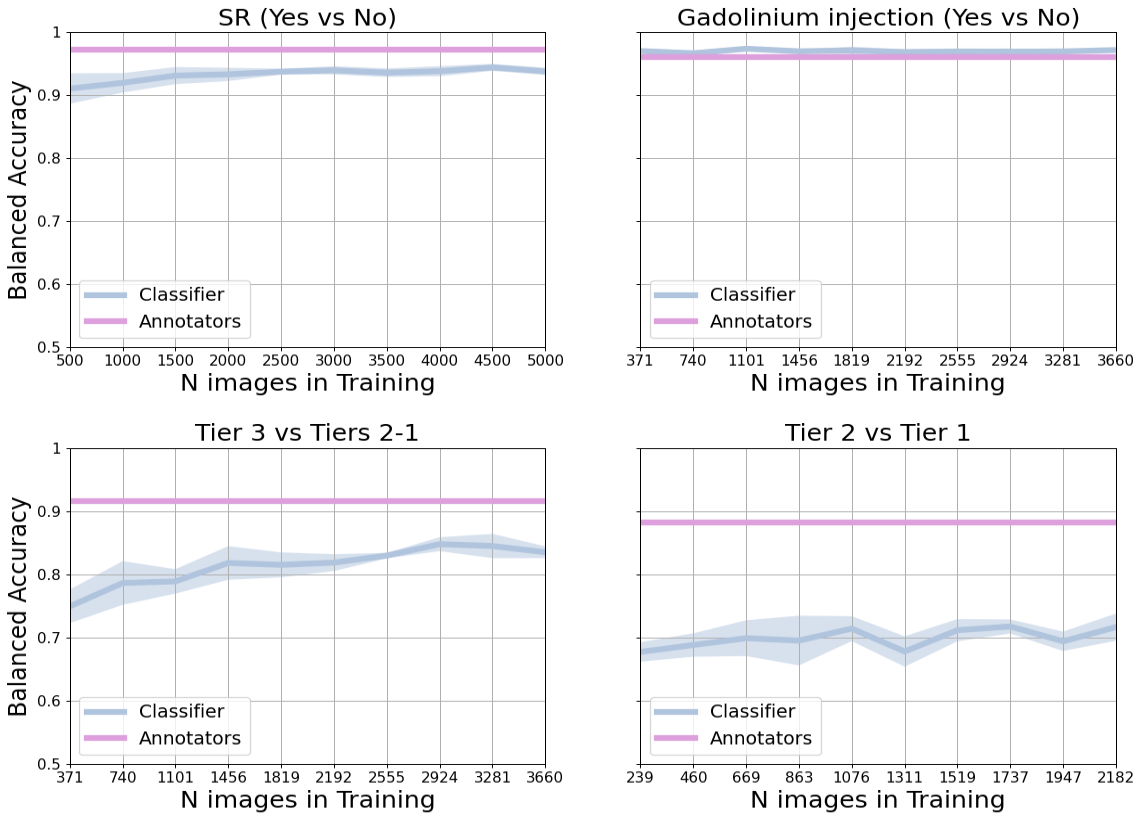

FIGURE 2.5: Learning curves for the SR (yes vs no), gadolinium injection (yes vs no), tier 3 vs tier 2-1 and tier 2 vs tier 1 tasks. Blue: balanced accuracy of the classifier across the five folds. Violet: balanced accuracy of the annotators on the testing set.

For tier 3 vs 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 2.5. 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.

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 2.6.

A. Tier 3 vs tiers 2-1

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

B. Tier 2 vs tier 1

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

TABLE 2.6: 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

2.4 Discussion

In this work, we developed a method for the automatic QC of T1w brain MRI for a large clinical data warehouse. Our approach allows: i) discarding images which are of no interest (SR), ii) recognizing gadolinium injection , iii) 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.

In the last decades, many computer-aided diagnosis systems using machine learning methods have been proposed for the detection of lesions or tumours, or for the classification of neurodegenerative or psychiatric diseases (Rathore et al., 2017; Işın, Direkoğlu, and Şah, 2016; Burgos et al., 2021). Algorithms were mainly developed and tested using research images (Samper-González et al., 2018; Noor et al., 2019; Cuingnet et al., 2011), or clinical datasets of limited size (Morin et al., 2020; Zhang et al., 2019a; Campese et al., 2019; Oh et al., 2019). Their validation on large realistic clinical datasets is crucial. To that aim, clinical data warehouses, which may gather millions of clinical routine images, offer fantastic opportunities. They also provide considerable challenges. In particular, selecting adequate images for a given analysis task can be very difficult: DICOM attributes may be unreliable, images may be of the wrong type, truncated and their quality is extremely variable. Therefore, automatic curation and QC methods are needed to fully exploit the potential of clinical data warehouses. Important efforts and achievements have been made by the scientific community to propose protocols and automatic tools for QC. MRIQC (Esteban et al., 2017) and VisualQC (Raamana et al., 2020) are two tools developed for the QC of T1w brain MRI data: they propose the extraction of image quality metrics for the detection of outliers, and a graphical interface to check the images. Alfaro-Almagro et al., 2018 proposed a pipeline for the UK Biobank dataset. Sujit et al., 2019 trained a CNN using the research dataset ABIDE. Other works focused on QC of processing results (segmentation) rather than raw data (Keshavan et al., 2018; Klapwijk et al., 2019). However, all these tools were designed for research data. Even if the data came from multiple sites, they do not cover all the images existing in a clinical PACS: they did not cover images with gadolinium and the patients presented with a limited number of diseases. Indeed, research datasets do not contain SR or tier 3 images and they may have very few tier 2 images. Protocols for the acquisition of research data are often different (in particular, scanning time is often longer) and a systematic visual QC is often performed. If the quality of an image is poor, a second scan can be acquired and information about the image quality is provided. In addition, DICOM fields are standardized among a research dataset, meaning that from the modality name it is possible to recognise whether a gadolinium-based contrast agent has been injected or not. On the contrary, in a clinical data warehouse, we may find images with or without gadolinium injection, “research quality" images, and images segmented, cropped or with so much motion that it is impossible to distinguish the brain. This heterogeneity makes it impossible to use other QC tools present in the literature. In particular, software tools such as MRIQC Esteban et al., 2017 propose an extensive image pre-processing pipeline before the calculation of image quality metrics. Classical neuroimaging software tools, such as SPM, ANTS or FSL, are typically validated only on T1w brain MRI of a good quality and without

gadolinium. The quality of our data, in particular of SR images that represent 25% of our dataset and the fact that we have about 44% of images with gadolinium injection, does not allow us to trust the metrics extracted from segmentations. To the best of our knowledge, we are the first to propose an automatic QC framework for clinical data warehouses.

To train our automatic QC algorithm, we had to manually annotate a large sample of images from the data warehouse. It was not possible to use existing protocols and software tools. In addition to the limitations mentioned above, we were also constrained by the environment of the data warehouse which only included a Jupyter notebook and a command-line interface. While constraints may vary from a data warehouse to another, it is very common that the data cannot be downloaded and thus have to be used within a specific informatics set-up (Daniel and Salamanca, 2020). We thus developed a dedicated visual QC protocol, with the assistance of a resident radiologist. We compared the annotation using 3D images and 2D slices, and we concluded that three 2D slices were sufficient and could represent a good compromise to fulfil our objectives: one being the exclusion of bad quality images that would compromise further analyses. Manual annotation results showed that our protocol is reproducible across all tasks, even though agreement was weaker for more challenging characteristics. Inter-rater agreement was strong for the SR label and the gadolinium injection and moderate for other characteristics. Manual annotation also provides 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 very low quality (Tier 3). We also confirmed that gadolinium has a strong impact on image quality, hence the critical importance of detecting it accurately, the DICOM attributes being unreliable in that regard.

For detecting straight reject, our CNN had excellent performance (BA greater than 90%). Even though the task is relatively easy, this is very important in order to automatically discard images in a very large scale study. This was also the case for detection of gadolinium, an important characteristic that strongly impacts the behavior of many image analysis methods. For the rating of image quality, the situation was different for identifying Tier 3 (low quality) images and for separating Tier 2 (medium quality) and Tier 1 (high quality). The proposed CNN classifier identified low quality images (Tier 3) with a high accuracy (83%). This is important because these are typically the images on which image processing algorithms could fail. Differentiating images of high and medium quality could also be useful but is less important as both categories can likely lead to reliable diagnostic predictions. 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. We compared several more sophisticated CNN architectures to our simple network based on five convolutional and three fully connected layers. However, these more complex networks (3D Inception and 3D ResNet) did not provide any significant improvement in performance. We could not compare our approach with the more standard ones based on the extraction of the image quality metrics since the software tools are not adapted to our data: we can trust the results of a classifier based on these types of features only if we trust the segmentation results. Our aim was to propose a framework for the QC that can be re-used on a clinical platform and so must be adapted to different tasks and have a preprocessing as light as possible. This is the reason why we

developed a CNN for all the tasks.

Thanks to the large number of hospitals in the AP-HP consortium (39 hospitals) and to the huge amount of images collected over the years (1980–now), we strongly believe that this dataset is representative of 3D T1w brain MRI that may be acquired in other hospitals. Consequently, the use of our QC framework could be generalized and it represents a first important step for the use of clinical data warehouses for the design of computer-aided diagnosis systems. Indeed, this work on quality control can help researchers to conduct studies, from observational studies that include MRI-based measurements to the development of CAD systems. First, obviously, the system will help save time by excluding SR images since they are not usable at all, both for training and testing. Even a neuroradiologist would not rely on these images for diagnostic purposes. Thereafter, the graded quality is also useful: either by controlling this confounding factor that can impact classification results or results of correlative studies, or by excluding images of bad quality (i.e. tier 3) when training the CAD. The quality grade could also contribute to building a confidence score for a classifier: when performing inference on a bad quality image we could lower the confidence in the classifier's result. Our study is going to be useful when performing research studies of different kinds (from training machine learning models to observational clinical retrospective studies). It is true that, beyond research, it could potentially be useful in a clinical routine setting. However, several steps would be needed towards that aim. First, it would obviously need to be approved as a medical device (e.g. FDA or CE approval). The most natural way to integrate it would probably be within the software provided by the MRI vendor. The computer hardware associated with the MRI machine is certainly powerful enough to perform the inference steps of our models. In a clinical routine setting, there are several potential usages of the approach. The most natural may be to associate it to automatic quantification algorithms which are more and more commonly available within the radiologist console. This would help flag exams for which, due to image quality, quantification cannot be considered reliable.

The main limitations of our study concern the annotation process. With the analysis of only three slices, we limit the chances to notice localised artefacts. Another consequence is that it may be difficult to properly distinguish the characteristics when an image is degraded: in particular the motion and the noise may be confused. This is also reflected by moderate values of the weighted Cohen's kappa obtained for these two characteristics. Additionally, even if we believe that the CNN models that were trained on data from the AP-HP data warehouse can be applied to other clinical datasets due to the large numbers of hospitals and scanner models involved in study and to the extended period of time, it would be beneficial to apply them on a public dataset for benchmarking. Furthermore, it would be interesting to study the potential association between the diagnoses of the patients and the quality of the images and the performance of the automatic QC. However, such a study is not straightforward to conduct due to the multiplicity of diagnostic codes for a given inpatient and the absence of any diagnostic information for outpatients. This is left for future work.

2.5 Conclusion

In this work, we proposed a framework for the automatic quality control of 3D brain T1w MRI for a large clinical data warehouse. Thanks to the manual annotation of 5500 images, we trained and validated different convolutional neural networks on 5000 images with a 5-fold CV and we tested them on an independent test set of 500 images. The classifier was as efficient as manual rating for the classification of images which are not proper 3D T1w brain MRI (i.e. truncated or segmented images) and for the images for which gadolinium was injected. In addition, the classifier was able to recognise low quality images with good accuracy.

Supplementary Material

QC grading	N	Gadolinium injection	N	Contrast Grade	N	Motion Grade	N	Noise Grade	N
Tier 1	873	With gado	358	-	-	-	-	-	-
		Without gado	515	-	-	-	-	-	-
Tier 2	1533	With gado	812	0	342	0	409	0	431
				1	470	1	403	1	381
		Without gado	721	0	237	0	441	0	445
				1	484	1	280	1	276
Tier 3	1639	With gado	1246	0	40	0	451	0	683
				1	56	1	425	1	549
				2	1150	2	370	2	14
		Without gado	393	0	27	0	147	0	271
				1	43	1	86	1	120
				2	323	2	160	2	2
SR	1455	-	-	-	-	-	-	-	-

TABLE 2.7: For each QC grading, we report the total number of images, the number of images with or without gadolinium injection and the number of images per grade for the contrast, motion and noise characteristics.

Layer	Filter size	Number of filters/ neurons	Stride size	Padding size	Dropout rate	Output size
Conv+BN+ReLU - 1	3x3x3	8	1	1	–	8x169x208x179
MaxPool - 1	2x2x2	–	2	adaptive	–	8x85x104x90
Conv+BN+ReLU - 2	3x3x3	16	1	1	–	16x85x104x90
MaxPool - 2	2x2x2	–	2	adaptive	–	16x43x52x45
Conv+BN+ReLU - 3	3x3x3	32	1	1	–	32x43x52x45
MaxPool - 3	2x2x2	–	2	adaptive	–	32x22x26x23
Conv+BN+ReLU - 4	3x3x3	64	1	1	–	64x22x26x23
MaxPool - 4	2x2x2	–	2	adaptive	–	64x11x13x1
Conv+BN+ReLU - 5	3x3x3	128	1	1	–	128x11x13x12
MaxPool - 5	2x2x2	–	2	adaptive	–	128x6x7x6
Dropuout	–	–	–	–	0.5	128x6x7x6
FC - 1	–	1300	–	–	–	1500
FC - 2	–	50	–	–	–	50
FC - 3	–	2	–	–	–	2
Softmax	–	–	–	–	–	2

TABLE 2.8: Hyperparameters of the 3D Conv5_FC3 CNN. BN: batch normalization; Conv: convolutional layer; FC: fully connected; MaxPool: max pooling.

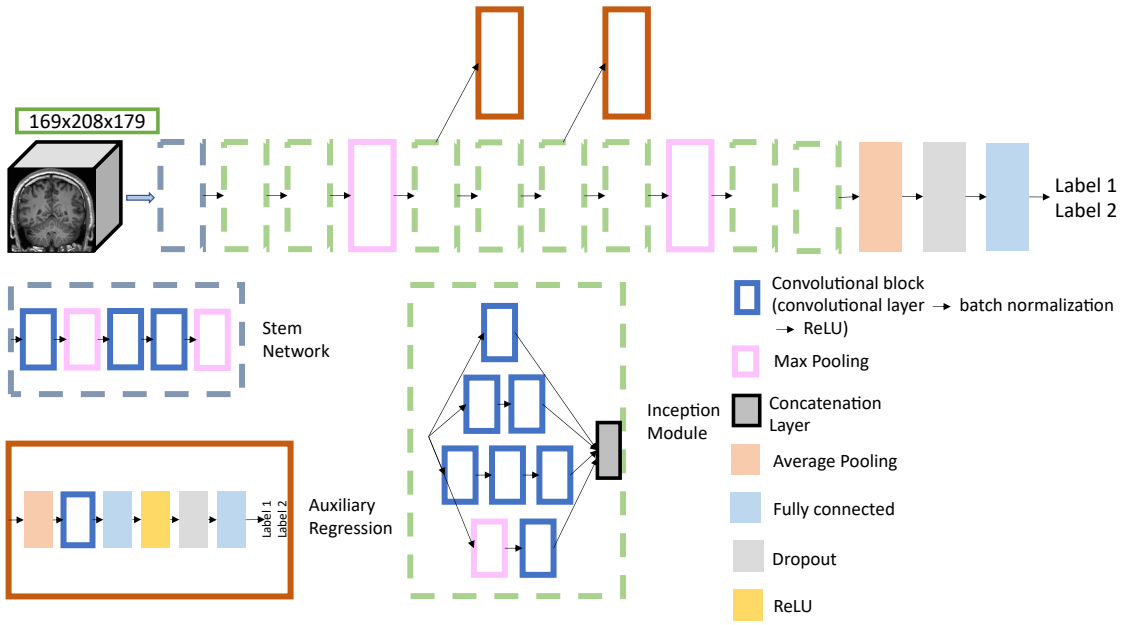

FIGURE 2.6: Architecture of the Inception 3D CNN. More information regarding the hyperparameters can be found in (Couvy-Duchesne et al., 2020).

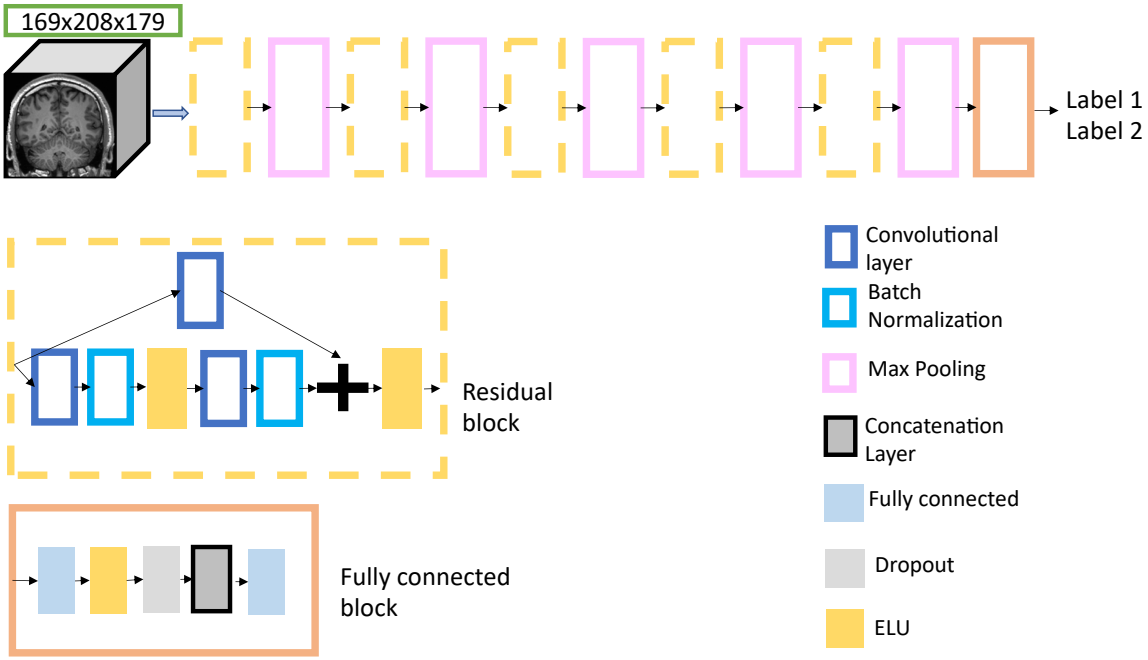

FIGURE 2.7: Architecture of the ResNet 3D CNN. More information regarding the hyperparameters can be found in (Couvy-Duchesne et al., 2020).

Chapter 3

Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation

---

This chapter is in revision to the *Journal of Medical Imaging*. A short version has been published in the Proceedings of the SPIE Medical Imaging 2022 conference.

- **Title:** Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation
- **Authors:** Simona Bottani, Elina Thibeau-Sutre, Aurélien Maire, Sebastian Ströer, Didier Dormont, Olivier Colliot, Ninon Burgos, APPRIMAGE Study Group

---

3.1 Introduction

Clinical data warehouses, gathering hundreds of thousands of medical images from numerous hospitals, offer unprecedented opportunities for research. They can for example be used to develop and validate machine learning and deep learning algorithms for the computer-aided diagnosis of neurological diseases. However, they also pose important challenges, a major challenge being their heterogeneity. Neurological diseases can result in a variety of brain lesions that are each studied with specific magnetic resonance imaging (MRI) sequences. For example, T1-weighted (T1w) brain MR images enhanced with a gadolinium-based contrast agent are used to study lesions such as tumors, and T1w images without gadolinium are used to study neurodegenerative diseases.

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. This is critical as

differential diagnosis in a clinical setting can be more challenging than in a research setting as different diseases may co-exist. Software tools such as SPM (Penny et al., 2011), ANTs (Avants et al., 2014) or FSL (Mark et al., 2012) have been widely used for feature extraction but they were largely validated using structural T1w MRI without gadolinium, 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. The U-Net and conditional generative adversarial networks (GANs) appear as the two most popular options. The U-Net was originally proposed for image segmentation (Ronneberger, Fischer, and Brox, 2015): an encoder with convolutional and downsampling blocks is followed by a decoder with upsampling and convolutional layers. The skip connections linking the encoder and decoder blocks at the same level enable the reconstruction of fine-grained details, explaining the popularity of this architecture for 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). Conditional GANs consist of a generator, which may adopt the U-Net architecture, followed by a discriminator in charge of distinguishing synthetic from real images and challenging the generator so that it improves the quality of the generated images. The good results obtained with conditional GANs explain their wide use for image translation (Chen et al., 2018; Gu et al., 2019; Kim, Do, and Park, 2018; Dinkla et al., 2018; Emami et al., 2018; Nie et al., 2018; Dar et al., 2019; Yu et al., 2019; Li et al., 2019a; Sharma and Hamarneh, 2019).

Both U-Net like models and conditional GANs have been proposed for diverse applications. Some aim to enhance the quality of the input images, for example by reducing noise in MRI (Benou et al., 2017; Jiang et al., 2018; Ran et al., 2019) or positron emission tomography (Hashimoto et al., 2019) images or by performing super-resolution (Chen et al., 2018; Du et al., 2020; Kim, Do, and Park, 2018; Pham et al., 2017; Zeng et al., 2018). Other works aim to translate an image of a particular modality into another modality, such as an MRI into an X-ray computed tomography (CT) (Han, 2017; Wolterink et al., 2017; Emami et al., 2018; Nie et al., 2018; Gong et al., 2018; Ladefoged et al., 2019) or a particular MRI sequence into another sequence (Dar et al., 2019; Yu et al., 2019; Li et al., 2019a; Sharma and Hamarneh, 2019). The U-Net architecture has also been used for the data harmonization: Dewey et al., 2019 built Deep-Harmony that aims to homogenize the contrast between images coming from different sites.

Closer to our application, various deep learning models have been developed for the synthesis of images with gadolinium from images without gadolinium: they include reinforcement learning for liver MRI (Xu et al., 2021), or Gaussian mixture modeling for CT images (Seo et al., 2021). As for the other image translation tasks, 3D U-Net like models have also been used to convert T1w-nce into T1w-ce images (Bône et al., 2021; Kleesiek et al., 2019; Sun et al., 2020). In two studies (Bône et al., 2021; Kleesiek et al., 2019), multimodal MRI sequences were used as input of the 3D U-Net that was trained and tested on patients with brain cancers. More specifically, the 3D U-Net proposed by Kleesiek et al.,

2019 predicts patches of T1w-ce, while the one proposed by Bône et al., 2021 directly predicts the full 3D T1w-ce image. The residual attention U-Net described in the last work (Sun et al., 2020) outputs synthetic T1w-ce that are used for the evaluation of cerebral blood volume in mice, instead of the real T1w-ce.

Our objective in this work was to obtain a homogeneous data set of T1w-nce images from very heterogeneous images coming from a clinical data warehouse. This homogenization 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 a very large clinical data warehouse (39 different hospitals of the Greater Paris area). 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 gray matter, white matter and cerebrospinal fluid obtained by segmenting the real T1w-nce, real T1w-ce and synthetic T1w-nce images using SPM (Ashburner and Friston, 2005) in order to verify that features extracted from synthetic T1w-nce were reliable. Preliminary work is accepted for publication in the proceedings of the SPIE Medical Imaging 2022 conference (Bottani et al., 2022b). Contributions specific to this paper include the development of additional models (a 3D U-Net like model with the addition of transformer layers, and three conditional GAN models using 3D U-Net like models as generators and a patch-based discriminator) and an extended validation of the segmentation task with a deeper analysis the tissue volume differences.

3.2 Materials and methods

3.2.1 Data set description

This work relies on a large clinical data set containing all the T1w brain MR images of adult patients scanned in one of the 39 hospitals of the Greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]). The data were made available by the AP-HP data warehouse 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.

Images were acquired as part of the routine clinical care in the different hospital sites and gathered in a central hospital PACS. Images relevant to the research project were copied to the research PACS and pseudonymized. They always remain within the hospital network that we accessed remotely. Images from this clinical data warehouse are very heterogeneous (Bottani et al., 2022a): they include 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) from 1980 up to now.

In a previous work (Bottani et al., 2022a), we developed a quality control framework to identify images that are not proper T1w brain MRIs, to identify acquisitions for which gadolinium was injected, and to rate the overall image quality defined based on three characteristics: motion, contrast and noise. We did so by manually annotating 5500 images (out of a batch of 9941 images that were available) to train and test convolutional neural network (CNN) classifiers. The graphical interface used to manually annotate the images is publicly available (`https://github.com/SimonaBottani/Quality_Control_Interface`).

The data set used in this work is composed of 307 pairs of 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., 2022a), 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., 2022a). 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).

3.2.2 Image preprocessing

All the images were organised using the Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016). 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 using the SyN algorithm (Avants et al., 2008). The registered images were further rescaled based on the min and max intensity values, and cropped to remove background resulting in images of size 169×208×179, with 1 mm isotropic voxels (Wen et al., 2020). Finally all the images were resampled to have a size of 128×128×128 using trilinear interpolation in Pytorch.

3.2.3 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).

3.2.3.1 3D U-Net like structures

We implemented three models derived from the 3D U-Net (Ronneberger, Fischer, and Brox, 2015): 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, Fischer, and Brox, 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 3.1.

- Res-U-Net: The *Res-U-Net* we implemented is based on the architecture first proposed by Milletari, Navab, and Ahmadi, 2016 and later used in (Bône et al., 2021). The five descending blocks are composed of 3D convolutional layers followed by an instance normalization 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. 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., 2016): 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 normalization 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 in order not to loose local information across depth and space. The features extracted by the CNN are the input of the transformer whose aim is to model the

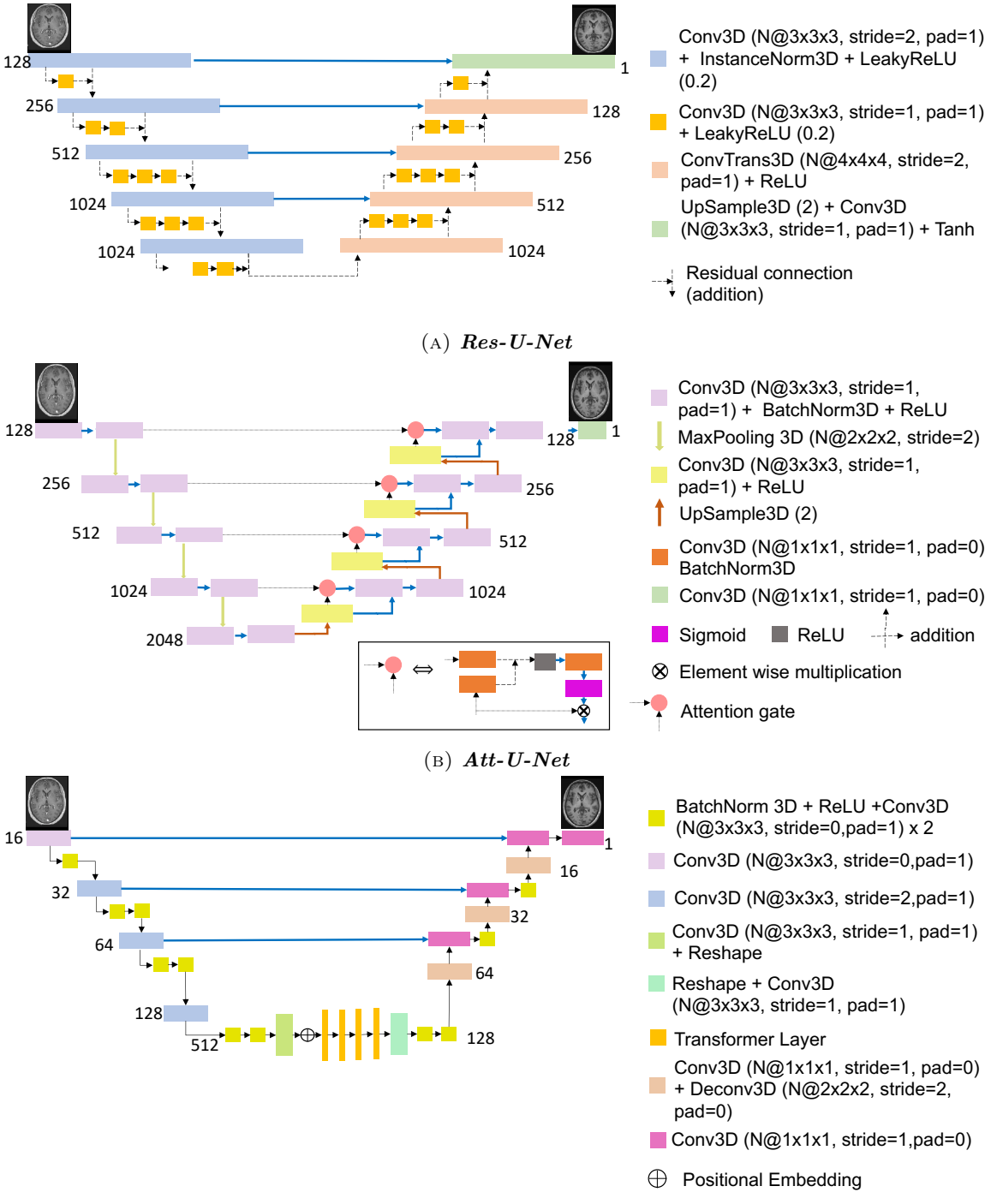

FIGURE 3.1: Architectures of the proposed 3D U-Net like models. The models take as input a real T1w-nce image of size 128×128×128 and generate a synthetic T1w-nce of size 128×128×128. *Res-U-Net*: images pass through five descending blocks, each one followed by a residual module, and then through four 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 input 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.

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 normalization 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 normalization 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 optimizer, the L1 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.

3.2.3.2 Conditional GANs

Generative adversarial networks (GANs) were firstly introduced by Goodfellow et al., 2014. They are generative deep learning models composed of two elements: a generator for synthesizing new examples and a discriminator for classifying whether examples are real, i.e. the original ones, or fake, i.e. synthesized by the generator. Conditional GANs (cGANs) (Mirza and Osindero, 2014) are a variant of GANs where the generator and the discriminator are conditioned by the true samples. They can only be used with paired data sets.

We propose three different cGAN 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 and Lee, 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 128×128×128, we created eight patches of size 64×64×64 with a stride of 50.

For the training of the discriminator we used the least-square-loss as proposed in (Mao et al., 2017) in order 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 pretrained separately. Regarding each generator, we reused the best model obtained previously. The discriminators were pretrained for the recognition of real and fake patches (fake images were obtained from each pretrained 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.

3.2.4 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}).

3.2.4.1 Synthesis accuracy

Image similarity was evaluated using the mean absolute error (MAE), peak signal-to-noise ratio (PNSR) 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 allows 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.

3.2.4.2 Segmentation fidelity

Our goal is to obtain gray 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 analyzing the tissue volumes resulting from the segmentations, which are important features when studying atrophy in the context of neurodegenerative diseases.

The volumes of the different tissues were obtained as follows. At first, synthetic T1w-nce images were resampled back to a size of 169×208×179 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 and Friston, 2005) simultaneously performs tissue segmentation, bias correction and spatial normalization. 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:

$$\text{rAD} = \frac{|V_t^I - V_t^J|}{TIV^I} \times TIV \ , \tag{3.1a}$$

$$\text{rD} = \frac{V_t^I - V_t^J}{TIV^I} \times TIV \ , \tag{3.1b}$$

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 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 not 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.

3.3 Results

We report results for the proposed 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 3.2. 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 gray and white matter is preserved in the synthetic T1w-nce (particularly visible for patients B and D). For

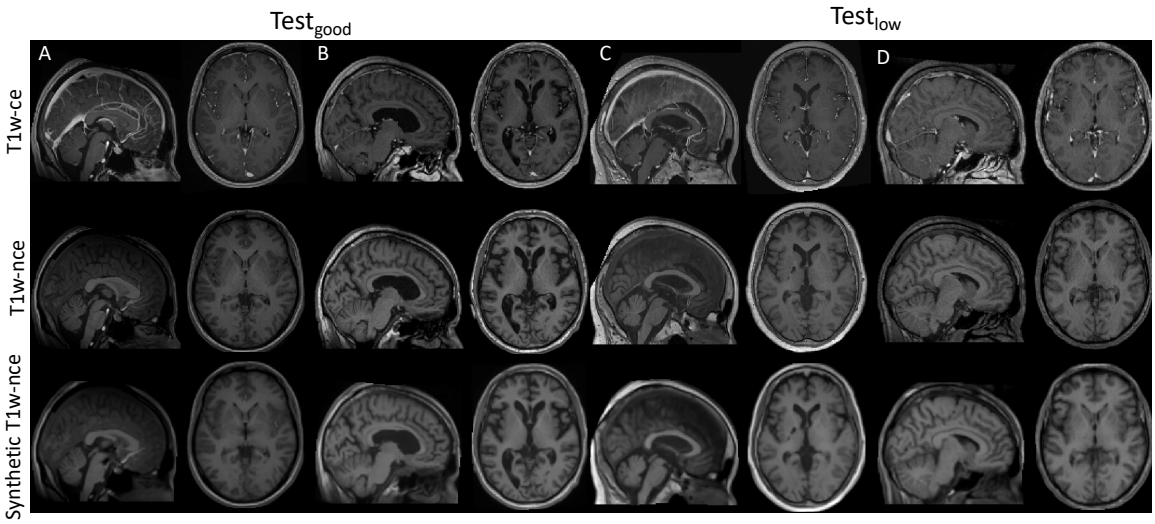

FIGURE 3.2: Examples of real T1w-ce (top), real T1w-nce (middle) and synthetic T1w-nce obtained with the *cGAN Att-U-Net* model (bottom) images in the sagittal and axial planes. Images of patients A and B belong to $\text{Test}_{\text{good}}$ (left) while images of patients C and D belong to Test_{low} (right).

Test_{low}, the contrast seems improved in the synthetic compared with the real T1w-ce image (especially for patient D).

3.3.1 Synthesis accuracy

Table 3.1 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).

Among the generators composed of 3D U-Net like models, the *Att-U-Net* performed slightly better than the others, both for $\text{Test}_{\text{good}}$ (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 were 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 $\text{Test}_{\text{good}}$ 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 *Att-U-Net* and *cGAN Att-U-Net*.

3.3.2 Segmentation fidelity

Absolute volume differences (rAD) obtained between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* model and the *cGAN Att-U-Net*) for GM, WM and CSF are reported in Table 3.2. For both test sets

TABLE 3.1: 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	*Res-U-Net*	3.06 ± 1.50	26.89 ± 4.30	0.95 ± 0.04
		Att-U-Net	2.73 ± 1.69	29.07 ± 4.53	0.96 ± 0.05
		Trans-U-Net	2.80 ± 1.42	28.00 ± 4.13	0.96 ±0.04
		cGAN Res-U-Net	3.47 ± 1.59	23.89 ± 4.30	0.95 ± 0.04
		cGAN Att-U-Net	2.69 ± 1.68	28.89 ± 4.44	0.97 ± 0.05
		cGAN Trans-U-Net	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	*Res-U-Net*	2.93 ± 1.77	26.71 ± 4.32	0.95 ± 0.05
		Att-U-Net	2.89 ± 1.85	27.15 ± 4.57	0.95 ± 0.05
		Trans-U-Net	2.98 ± 1.89	26.71 ± 4.38	0.94 ± 0.05
		cGAN Res-U-Net	3.20 ± 1.96	26.20 ± 4.42	0.93 ± 0.05
		cGAN Att-U-Net	2.86 ± 1.83	27.12 ± 4.50	0.95 ± 0.05
		cGAN Trans-U-Net	2.97 ± 1.83	26.68 ± 4.40	0.94 ± 0.05

TABLE 3.2: Absolute volume difference (mean ± standard deviation in cm^3) between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and *cGAN Att-U-Net* models) for the gray 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	$Test_{good}$ [cm^3]	$Test_{low}$ [cm^3]
Gray matter	T1w-nce / T1w-ce	-	26.68 ± 15.92	49.63 ± 49.38
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	10.36 ± 6.98 *	19.61 ± 29.54 *
		cGAN Att-U-Net	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-ce	*Att-U-Net*	7.79 ± 5.87	13.95 ± 24.74 *
		cGAN Att-U-Net	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-ce	*Att-U-Net*	13.37 ± 10.18 *	12.25 ± 7.72 *
		cGAN Att-U-Net	18.27 ± 17.20 *	17.10 ± 18.45 *

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 *Att-U-Net* on $Test_{good}$, absolute volume differences of GM and CSF between T1w-nce/T1w-ce and T1-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 significantly different (corrected p-value <0.01) for both test sets. This means that there is an advantage in using synthetic

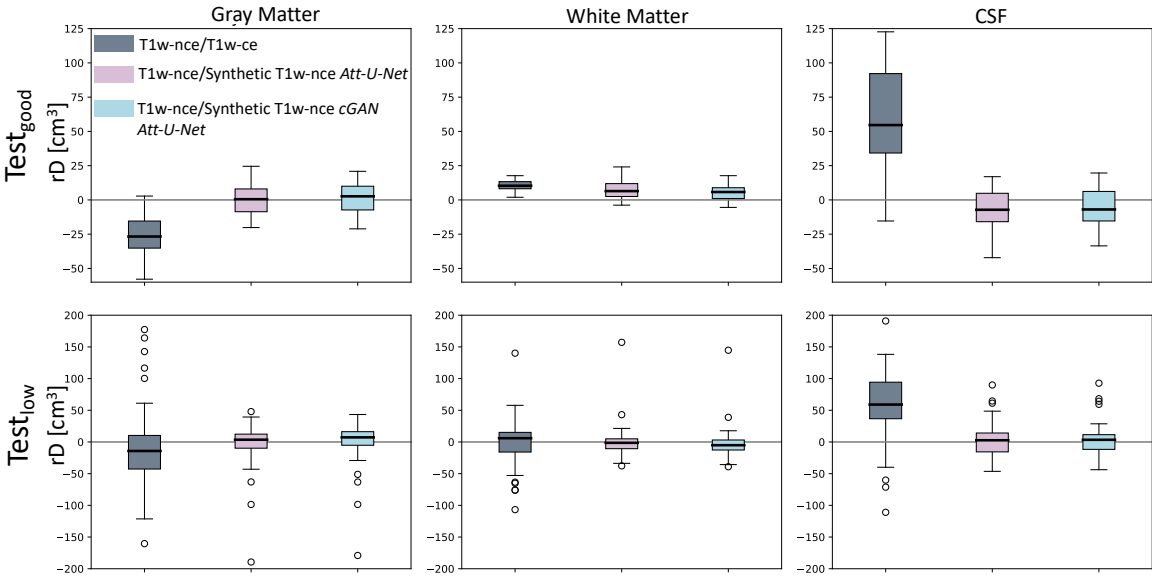

FIGURE 3.3: Volume differences (rD) in cm^3 between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and the *cGAN Att-U-Net* models) for gray matter (left), white matter (middle) and cerebrospinal fluid (CSF, right) for both $\text{Test}_{\text{good}}$ (top) and Test_{low} (bottom).

TABLE 3.3: Dice scores obtained when comparing the gray 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 *Att-U-Net* and the *cGAN Att-U-Net*)

	Compared images	Model	$\text{Test}_{\text{good}}$	Test_{low}
Gray matter	T1w-nce / T1w-ce	-	0.88 ± 0.02	0.77 ± 0.12
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	0.87 ± 0.02	0.81 ± 0.07
		cGAN Att-U-Net	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	*Att-U-Net*	0.90 ± 0.02	0.86 ± 0.04
		cGAN Att-U-Net	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	*Att-U-Net*	0.80 ± 0.05	0.78 ± 0.07
		cGAN Att-U-Net	0.80 ± 0.05	0.78 ± 0.07

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 *Att-U-Net* and *cGAN Att-U-Net*) for GM, WM and CSF are reported in Figure 3.3. 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).

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 *Att-U-Net* and the *cGAN Att-U-Net*) are displayed in Table 3.3. We observe that for both gray 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.

3.4 Discussion

The use of clinical images for the validation of computer-aided diagnosis (CAD) systems is still largely unexplored. One of the obstacles lies in the heterogeneity of the data acquired in the context of routine clinical practice. Post-acquisition homogenization is crucial because, contrary to research data, no strict acquisition protocols, that would ensure a certain homogeneity among the images, exist for clinical data. Heterogeneity originates from the fact that images are acquired with different scanners at different field strengths during a large period of time and because patients may suffer from a large variety of diseases. Homogenization of clinical data sets of 3D T1w brain MRI, and consequently of the features extracted from them, is an important step for the development of reliable CAD systems. Indeed, when training a CAD system, the algorithms must not be affected by the data set variations even though clinical images may greatly vary.

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., 2022a). To homogenize this data set, we thus proposed a framework to convert T1w-ce images into T1w-nce images using deep learning models. The choice to synthesize T1w-nce images from T1w-nce images was constrained by the fact that software tools for feature extraction in the neuroimaging community were developed for T1w-nce MRI. To the best of our knowledge, none of these tools has largely been applied to the extraction of features from T1w-ce MRI data and their performance in this scenario is thus mostly unknown.

The contribution of our work consists in the development and validation of deep learning models (U-Net models and conditional GANs) for the translation of T1w-ce to T1w-nce images coming from a clinical data warehouse. We compared three 3D U-net models differentiated by the addition of residual modules, of attention modules or of transformer layers, used as simple generators and also within a conditional GAN setting with the addition of a patch-based discriminator. These models have widely been used for the image translation of medical images (Yi, Walia, and Babyn, 2019; Burgos et al., 2021), but their application to clinical data has not been proven yet. The proposed models were trained using 230 image pairs and tested on two different test sets: 26 image pairs had both a T1w-nce and T1w-ce of good or medium quality and 51 image pairs had a T1w-nce of good or medium quality and a T1w-ce of bad quality. Having two test sets of different qualities is a key point since we are dealing with a real clinical heterogeneous data set where images of low quality, corresponding in majority to T1w-ce images with a low contrast, may represent 30% of the data (Bottani et al., 2022a).

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 synthesize T1w-nce images. Slightly better performance was reached with the addition of attention modules (*Att-U-Net* and *cGAN Att-U-Net* models), 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). We showed that the absolute volume differences of GM, WM and CSF were larger between real T1w-nce and Tw-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), 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. Even though the synthetic T1w-nce images enable the extraction of reliable features, their quality could still be improved. Many constraints exist when working with data from a clinical data warehouse. One is the fact that these data are accessible only through a closed environment provided by the IT department of the AP-HP as described in (Daniel and Salamanca, 2020). Limitations in computational resources and storage space make training deep learning models difficult and thus limits the experiments that can be performed to find the optimal model. The proposed models could be improved by better optimizing the hyperparameters (such as the learning rate or the size of the kernels), adding a perceptual loss when training the conditional GANs (Zhao et al., 2016) or adding more layers in the patch-based discriminator. Other architectures could also be explored. We have restricted our work to conditional GANs, which need paired data to be trained, but we could exploit more data working with cycle GANs (Zhu et al., 2017) as they can deal with unpaired data.

Several steps remain to be performed before using synthetic T1w-nce images for the differential diagnosis of neurological diseases. First, the performance of CAD systems trained with a mix of real T1w-nce and T1w-ce images should be compared with the performance of CAD systems trained with a mix of real and synthetic T1w-nce images. To prevent

introducing a correlation between image properties (e.g. smoothness) and pathology, which would bias the classification performance, it may be necessary to also feed the real T1w-nce images to the neural network and use the resulting images as inputs of the CAD system, as suggested in (Dewey et al., 2019).

3.5 Conclusion

Clinical data warehouses offer fantastic opportunities for computer-aided diagnosis of neurological diseases but their heterogeneity must be reduced to avoid biases. In this work we proposed to homogenize such a large clinical data set by converting images acquired after the injection of gadolinium into non-contrast-enhanced images using 3D U-Net models and conditional GANs. Validation using standard image similarity measures demonstrated that the similarity between real and synthetic T1w-nce images was higher than between real T1w-nce and T1w-ce images for all the models compared. We also showed that features extracted from the synthetic images (GM, WM, CSF volumes) were closer to those obtained from the T1w-nce brain MR images (considered as reference) than the original T1w-ce images. These results demonstrate the ability of deep learning methods to homogenize a data set coming from a clinical data warehouse.

Chapter 4

Detection of patients with dementia using T1w brain MRI in a clinical data warehouse

---

This chapter has been revised into an article submitted to *Medical Image Analysis*.

- **Title:** Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse
- **Authors:** Simona Bottani, Ninon Burgos, Aurélien Maire, Sebastian Ströer, Dario Saracino, Didier Dormont, Olivier Colliot, APPRIMAGE Study Group

---

4.1 Introduction

Dementia is a world-wide disease that is becoming more and more important due to population aging. T1-weighted (T1w) brain magnetic resonance imaging (MRI) contributes to the positive diagnosis of dementia by displaying typical spatial patterns of brain atrophy. 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) model: they could help doctors to better understand the disease and detect it early thanks to their ability to automatically extract relevant features.

CAD systems have been mainly developed using research data sets due to their ease of access (they can directly be downloaded from websites) and their ease of use: they are acquired following a research protocol whose aim is to guarantee data quality and homogenization. Several data sets originating from research studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1], the Open Access Series Of Imaging Studies (OASIS)[2], the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL)[3],

[1] http://adni.loni.usc.edu/
[2] https://www.oasis-brains.org/
[3] https://aibl.csiro.au/

the Frontotemporal lobar degeneration neuroimaging initiative (NIFD)[4] and the Parkinson's Progression Markers Initiative (PPMI)[5] are publicly available and contain various clinical and imaging data, including T1w MRI brain data. They have pushed the research on ML and DL for CAD using T1w brain MRI: in the literature we can find works focusing on Alzheimer's disease using ADNI, OASIS or AIBL data sets (Punjabi et al., 2019; Bidani, Gouider, and Travieso-González, 2019; Spasov et al., 2019; Böhle et al., 2019; Farooq et al., 2017; Wegmayr, Aitharaju, and Buhmann, 2018; Samper-González et al., 2018; Wen et al., 2020; Bron et al., 2021), or on fronto-temporal dementia using NIFD (Ma et al., 2020).

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.

In order to bring research advances to the clinic life, 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 specialized centers which are not representative of the overall clinical practice (for instance rare dementias and early-onset cases are overrepresented). Finally, they often restrict themselves to diagnosis of patients with dementia. It is thus unclear what is their specificity when dealing with MRI from patients with other diagnoses. Some works focused on the differential diagnosis, which is closer to the 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 (CDW) 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. Images of a CDW are heterogeneous (i.e. different sites, MRI sequences not harmonized) and 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 aim 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 hospitalization period. The ML model was a linear SVM using gray matter maps as features. It was then compared to several deep learning models. We compared performance obtained on a research data set to that obtained on the present clinical dataset. 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

[4] https://ida.loni.usc.edu/home/projectPage.jsp?project=NIFD
[5] https://www.ppmi-info.org/

and the presence of images of different quality). In order 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.

4.2 Materials

4.2.1 Research data set

The research data set used in this work was composed of subjects from the ADNI database (in particular ADNI 1,2, Go). We considered subjects diagnosed as cognitive normal (CN) or Alzheimer's disease (AD) at baseline and only kept subjects whose diagnosis did not change over time. We selected 800 T1w MRI corresponding to 800 subjects matching these criteria (CN: 410 subjects, 54.87 % F, age 73.20 $\pm$ 6.15 in range [55.1, 89.6]; AD: 390 subjects, 44.0 % F, age 74.88 $\pm$ 7.76 in range [55.1, 90.1]). 200 subjects (100 CN and 100 AD) composed the independent test set and the remaining subjects (310 CN and 290 CN) were used for the training/validation of the models using a 5-fold cross-validation (CV).

4.2.2 Clinical routine data set

The clinical data set comes from a large clinical database containing all the T1w brain MR images of adult patients scanned in hospitals of the Greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]). 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. All the data, both imaging and clinical, were pseudonymized by the AP-HP data warehouse and they always remained within the hospital network. We accessed it remotely for our study.

4.2.2.1 Imaging and clinical data collection

Images from this clinical data warehouse are very heterogeneous (Bottani et al., 2022a): they include 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). Imaging data were gathered in a central hospital PACS and images relevant to our research projects (i.e. 3D T1w brain MR images of patients aged more than 18 years old) were copied to the research PACS where they were pseudonymized. The selection process to obtain images of interest is described in (Bottani et al., 2022a).

At the same time, clinical data corresponding to the patients of our query are stored in a database based on the ORBIS software which is installed in the different hospitals. Clinical data gather all the information connected to the patients, i.e. date of birth, sex, ICD-10 codes, medications, biological tests, electronic health reports. As explained in (Daniel and Salamanca, 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 exam at AP-HP ($\sim$130.000 patients), only $\sim$25% were registered in ORBIS. Among

them, 23,688 patients were hospitalized. This is important because for non-hospitalized patients only sociodemographic data (sex and age) are available and not clinical data. As for the imaging data, the data warehouse was in charge to query ORBIS to provide the pseudonymized clinical data.

For our work we were interested in two sociodemographic items (age and sex) as well as one clinical item (ICD-10 codes). ICD-10 codes, from the 10^{th} revision of international classification of diseases (World Health Organization et al., 2007), were used to associate a diagnosis to each 3D T1w brain MRI. Images were labeled 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 of the hospitalization, this time window ensures that all information regarding brain disorders related to the need of a brain MRI exam are collected.

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

4.2.2.2 Definition of the different classes 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 hospitalization (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.

Firstly, 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. Thereafter, we divided the remaining codes into two groups: ICD-10 codes referring to diseases 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 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 4.1, 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 etiology, was used to identify the diseases corresponding to the dementia category as we wanted to be more specific.

Category	ICD-10 codes
D: Dementia associated to a neurodegenerative disease or a vascular disease that causes atrophy visible on T1w MRI.	• Dementia in AD with early onset (F000/G300) • Dementia in AD with late onset (F001/G301) • Dementia in AD, atypical or mixed type (F002/G308) • Dementia in AD, unspecified (F009/G309) • Dementia in Pick disease (F020/G310) • Dementia in Creutzfeldt-Jakob disease (F021/A810) • Dementia in Huntington disease (F022 + G10) • Vascular dementia of acute onset (F010) • Multi-infarct dementia (F011) • Subcortical vascular dementia (F012) • Mixed cortical & subcortical vascular dementia (F013) • Other vascular dementia (F018) • Vascular dementia, unspecified (F019) • Mild cognitive disorder (F067) • Dementia in Parkinson's disease (F023 + G20) • Lewy bodies dementia (G028 + G318)
NDL: No dementia but diagnosis that suggests presence of lesions that modify the anatomical structure of the brain visible on T1w MRI.	• 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

TABLE 4.1: 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.

4.2.2.3 Selection of patients belonging to the dementia category

Dementia is the principal category we consider since the aim of our work is to study, using clinical routine data, how well this category can be distinguished from the others using machine learning models. We thus started by selecting patients labeled as dementia. In the workflow displayed in Figure 4.1 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 the 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 their brain could appear as presenting atrophy but this would simply be due to aging and not

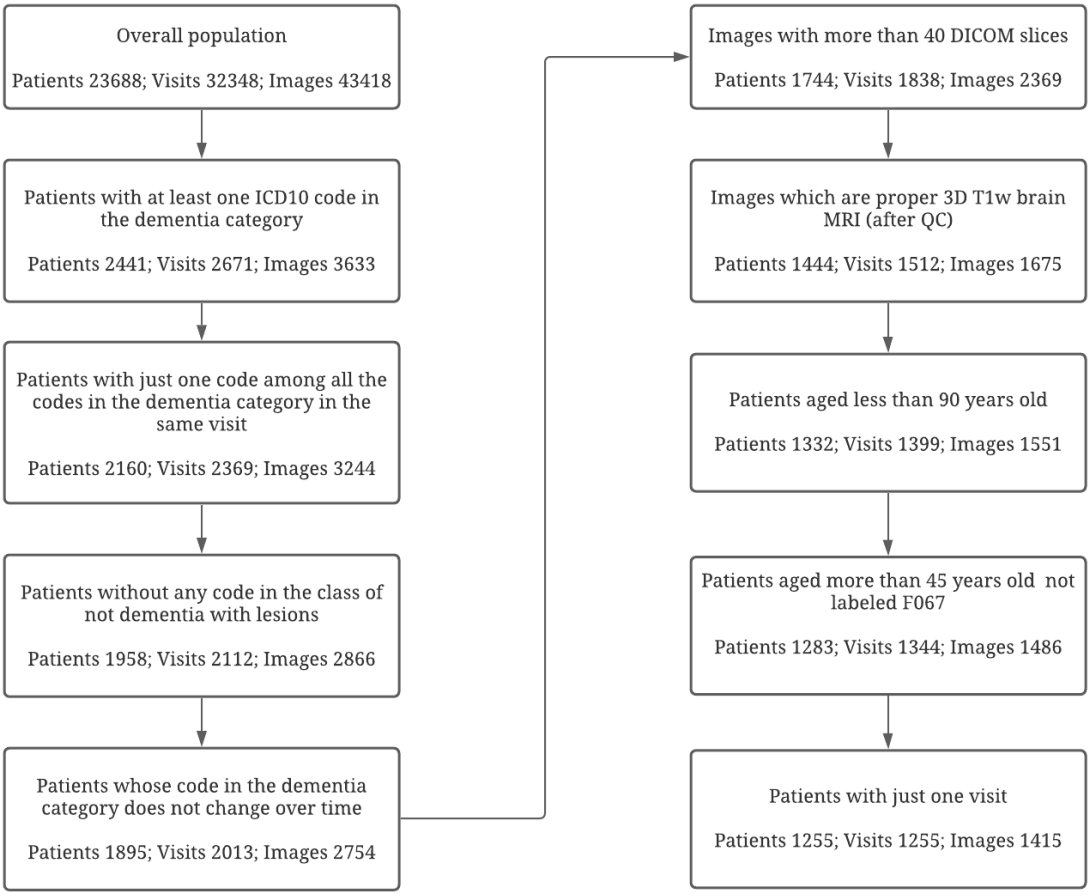

FIGURE 4.1: Workflow describing the selection of patients with belonging to the dementia category. For each selection step we report the corresponding number of patients, visits and images.

to a disease. Patients labeled F067 (mild cognitive disorder) aged less than 45 years were excluded because the diagnosis could correspond 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 labeled as straight reject by the quality control step.

4.2.2.4 Selection of the patients belonging to the no dementia with lesions (NDL) and no dementia 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 the four diseases leading to brain lesions visible on the T1w MRI (cancer, stroke, demyelination and

hydrocephalus, see Table 4.1). 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 the one 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 to be able to study the classification performance per disease.
- 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 twice this selection process since some images were discarded after the preprocessing 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. Here we describe the criteria to match a patient with dementia.

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 or 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 twice this selection process since some images were discarded after the preprocessing steps. In total we matched 1144 patients (corresponding to 1144 visits and 1343 images).

4.2.2.5 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). In the Table 4.2 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 (Bottani et al., 2022a), which will be detailed in the Methods section.

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

TABLE 4.2: 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 (Student's T 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 image level

4.2.2.6 Training subsets

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

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

4.3 Methods

4.3.1 Image pre-processing

The T1w MR images were converted from DICOM to NIfTI using the software dicom2niix (Li et al., 2016) and organized following 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 3$^{\text{rd}}$-order spline interpolation to obtain 1 mm isotropic voxels.

A first pre-processing consisted in applying 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 using the SyN algorithm (Avants et al., 2008). The registered images were further rescaled based on the min and max intensity values. Images were then cropped to remove background resulting in images of size 169×208×179, with 1 mm isotropic voxels (Wen et al., 2020) using trilinear interpolation.

This pre-processing was used to assess the quality of the images with an automatic approach proposed in (Bottani et al., 2022a). The automatic quality control (QC) approach first identified if a given image was or not a straight reject (i.e. segmented or cropped image). If it was not a straight reject, it was further labeled by the automatic QC tool according to the tiers of quality, i.e. tier 1 (good quality), tier 2 (medium quality) or tier 3 (bad quality). In addition, the automatic QC tool determined the presence or the absence of gadolinium-based contrast agent.

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

4.3.2 Synthesis of images without gadolinium

In order 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 (Bottani et al., 2022b) 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.

4.3.3 Machine learning models used for classification

4.3.3.1 Linear SVM

A linear SVM using probability gray 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). The Gram matrix $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$ was pre-calculated using a linear kernel k for each pair of images $(\mathbf{x}_i, \mathbf{x}_j)$ for the provided subjects and was used as input for the generic SVM. When using a pre-computed Gram matrix, computing time depends on the number of subjects, and not on the number of features and it can speed up the calculations. We optimized the penalty parameter C of the error term. The optimal value of C was chosen using nested cross-validation, 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.

4.3.3.2 CNN architectures

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

The three 3D CNN models considered in the paper are denominated as follows: Conv5_FC3, ResNet, InceptionNet. The first is composed of five convolutional layers and three fully connected layers implemented in (Wen et al., 2020), 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 ResNet and the InceptionNet were implemented and used for the work of (Couvy-Duchesne et al., 2020). All the details of the architectures can be found in (Bottani et al., 2022a).

The models were trained using the cross entropy loss. We used the Adam optimizer with a learning rate of 10^-5 for the ResNet model and of 10^-4 for the InceptionNet and the Conv5_FC3 models. 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.

4.3.4 Experimental setting

We performed two tasks: dementia vs no dementia with lesions (D vs NDL) and dementia vs no dementia no lesions (D vs NDNL).

4.3.4.1 Training framework

These tasks were performed using three different set ups:

- training on research data (ADNI: CN vs AD) and tested on clinical data using the linear SVM and the CNN models;
- pretraining on research data (ADNI: CN vs AD) and fine-tuning on clinical data using the CNN models;
- training from scratch on clinical data using the linear SVM and the CNN models.

4.3.4.2 Evaluation setting

Before starting the experiments for the two tasks of interest, 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. In case several images were available per visit, the selection was made randomly. This resulted in a test set composed of 152 patients/images for the D class, 152 patients/images for NDL and 152 patients/images for NDNL. The training/validation was composed of 604 patients and 719 images for D, 604 patients and 799 images for NDL, 604 patients and 756 images for NDNL.

CN vs AD

Metric	SVM	Conv5_FC3	ResNet	InceptionNet
BA	86.80 ± 0.40	84.10 ± 1.59	85.30 ± 1.03	82.10 ± 1.77
Sensitivity	82.80 ± 0.40	79.80 ± 4.45	83.00 ± 4.52	75.80 ± 8.68
Specificity	90.80 ± 0.40	88.40 ± 7.26	87.60 ± 4.67	88.40 ± 5.16

TABLE 4.3: Dementia classification performance (AD vs CN) in a research data set (ADNI). Results were obtained with different ML models: a linear SVM using as input gray matter maps and three CNN models (Conv5_FC3, ResNet and InceptionNet) using as input minimally pre-processed T1w MR images). BA: balanced accuracy.

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.

4.4 Results

4.4.1 Performance in a research data set

To set a baseline, we first studied the ability of ML and DL classification algorithms to identify patients with dementia in a research data set. Participants from ADNI were used to derive the training/validation and test sets. The training for all the models was done using a 5-fold CV and tested on an independent test set. Results are reported in Table 4.3. Results obtained with the SVM classifier and the best performing DL models were comparable (SVM balanced accuracy 86.40 ± 0.40; ResNet balanced accuracy 85.30 ± 1.03). This is in line with performances reported in the litterature (Samper-González et al., 2018; Wen et al., 2020).

4.4.2 Performance in the clinical data set

We then studied the ability of ML and DL classification algorithms to identify patients with dementia in the clinical data set.

Results obtained for the two tasks of interest (D vs NDNL and D vs NDL) are reported in Table 4.4. Best results were obtained with the ResNet for both tasks (D vs NDNL, balanced accuracy 73.95 ± 0.96; D vs NDL: 75.07 ± 1.64) even if all the scores are very close. For the D vs NDNL task, we note that the linear SVM with T1w gray matter maps as input has a lower performance than the CNN models. Sensitivity and specificity are balanced and performance across the two tasks is similar.

A. D vs NDNL

Metric	SVM	Conv5_FC3	ResNet	InceptionNet
BA	68.75 ± 0.36	73.62 ± 1.58	73.95 ± 0.96	72.24 ± 2.18
Sensitivity	66.97 ± 0.64	73.29 ± 3.02	74.21 ± 3.59	72.24 ± 4.97
Specificity	70.53 ± 0.49	73.95 ± 2.06	73.68 ± 2.32	72.24 ± 3.44

B. D vs NDL

Metric	SVM	Conv5_FC3	ResNet	InceptionNet
BA	73.09 ± 0.32	72.24 ± 1.82	75.07 ± 1.64	72.76 ± 1.99
Sensitivity	75.92 ± 0.89	74.34 ± 7.66	74.08 ± 3.85	73.82 ± 6.53
Specificity	70.26 ± 0.49	70.13 ± 8.98	76.05 ± 2.62	71.71 ± 5.39

TABLE 4.4: Dementia classification performance (D vs NDNL and D vs NDL) in a clinical data set. Results were obtained with different ML models: a linear SVM using as input gray matter maps and three CNN models (Conv5_FC3, ResNet and InceptionNet) using as input minimally pre-processed T1w MR images). BA: balanced accuracy.

In general we note a lower classification score in the detection of dementia in a clinical data set compared to research data: this may due to the heterogeneity of the classes in the clinical data set, where many diagnoses coexist.

4.4.2.1 Influence of gadolinium injection and image quality on the classification performance

As shown in Table 4.2, the proportion of images with and without gadolinium injection and of medium/good and low quality in the dementia, NDL and NDNL categories are different. In the dementia class, there are 25% of images with gadolinium. 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.

We used the training subsets $T^{172}_{\text{no gado}}$, $T^{181}_{\text{tier 1/2}}$ and T^{172} to evaluate the existence of a potential bias resulting from such imbalance. The order of magnitude of patients per class among the training subsets is equivalent, meaning that differences observed in the classification score can not depend on the training sample size but on the characteristics of the training subset. We assume that if the presence or absence of gadolinium, or the presence of different quality of images in the training set, does not have an impact, the performance will not vary when using the different training subsets, while if the composition of the training subsets leads to an improvement of the classification score, it means that the results are actually biased.

We used SVM with probability gray matter maps as input since it is faster to train than the different CNN models and it has fewer hyper-parameters to optimize. In Table 4.5 we

A. D vs NDNL

Metric	$T^{172}_{\textbf{no gado}}$	$T^{181}_{\textbf{tier 1/2}}$	T^{172}
BA	60.33 ± 0.26	61.32 ± 2.83	68.16 ± 0.38
Sensitivity	52.76 ± 0.26	79.87 ± 2.72	73.95 ± 2.41
Specificity	67.89 ± 0.26	42.76 ± 12.70	62.37 ± 2.18

B. D vs NDL

Metric	$T^{172}_{\textbf{no gado}}$	$T^{181}_{\textbf{tier 1/2}}$	T^{172}
BA	69.74 ± 0.55	64.61 ± 1.74	72.30 ± 0.48
Sensitivity	85.13 ± 0.79	45.53 ± 4.62	66.45 ± 1.32
Specificity	54.34 ± 1.84	83.68 ± 1.47	78.16 ± 1.92

TABLE 4.5: 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 with a linear SVM using as input gray matter maps and trained on different clinical data subsets ($T^{172}_{\text{no gado}}$, $T^{181}_{\text{tier 1/2}}$ and T^{172}). BA: balanced accuracy.

report the results of the classification for the two tasks using the different training subsets. 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 a subset with the same proportion of images with and without gadolinium and of images of medium/good quality as in the original data set (i.e.T^{172}) is higher than when using $T^{172}_{\text{no gado}}$ or $T^{181}_{\text{tier 1/2}}$: for D vs NDNL, the balanced accuracy is 68.16 ± 0.38 with T^{172}, 60.33 ± 0.26 with $T^{172}_{\text{no gado}}$ and with 61.32 ± 2.83 $T^{181}_{\text{tier 1/2}}$. The same trend is present for D vs NDL. This means that results are biased by the presence of gadolinium or the differences in image quality: results increase when the bias is present. Classification is not based on the detection of the disease but on the different characteristics of the training data set.

The training subset $T^{172}_{\text{no gado}}$ still contains images of different quality and $T^{181}_{\text{tier 1/2}}$ images with and without gadolinium. The classifier may thus still be exploiting biases in the image characteristics. At this step, we want to evaluate the performance of the classifier using a training dataset without the bias of the gadolinium and of the quality of the images. This is why we used the training subset called $T^{88}_{\text{no gado, tier 1/2}}$ comparing it with the training subset T^{88}: having the same training size. The difference in performance should thus depend only on the different proportions of gadolinium and quality in the training subset. In Table 4.6 we report the results of the two classification tasks using the two different subsets as training.

For both tasks, if we delete the bias of both gadolinium and image quality, the balanced accuracy hardly reaches 50%, meaning joint influence of these two characteristics increase performance by about 20%. In fact, balanced accuracy obtained with T^{88} is 69.47 ± 2.37

A. D vs NDNL

Metric	$T^{88}_{\textbf{no gado, tier 1/2}}$	T^{88}
BA	51.51 ± 2.54	69.47 ± 2.37
Sensitivity	6.71 ± 12.44	71.97 ± 2.26
Specificity	96.32 ± 7.37	66.97 ± 2.51

B. D vs NDL

Metric	$T^{88}_{\textbf{no gado, tier 1/2}}$	T^{88}
BA	50.00 ± 0.00	73.03 ± 1.79
Sensitivity	40.00 ± 48.99	66.58 ± 4.51
Specificity	60.00 ± 48.99	79.47 ± 1.13

TABLE 4.6: 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 with a linear SVM using as input gray matter maps and trained on two clinical data subsets ($T^{88}_{\text{no gado, tier 1/2}}$ and T^{88}). BA: balanced accuracy.

for D vs NDNL and 73.03 ± 1.13 for D vs NDL, which is almost equivalent to the performance for both tasks obtained with T^{172}: both training subsets contain the same biases. Therefore, when it cannot exploit biases in image characteristics, the classifier is not better than random.

4.4.2.2 Classification performance obtained after gadolinium removal using image translation

We showed in the previous chapter (Chapter 3) that gadolinium could be removed from contrast-enhanced T1w MR images using a DL-based image translation approach. We created a training subset composed of 88 synthetic images obtained from images of medium/good quality acquired with and without gadolinium injection as described in section 4.3.2. If the gadolinium is successfully removed, training with this subset should be equivalent to training with the $T^{88}_{\text{no gado, tier 1/2}}$ subset that includes only images without gadolinium. This is what we study in this section.

Results of these experiments are reported in Table 4.7. Balanced accuracy scores are equivalent when using the training subset Synthetic $T^{88}_{\text{tier 1/2}}$ and $T^{88}_{\text{no gado, tier 1/2}}$: it means that the effect of gadolinium has been deleted using synthetic images since the performance is the same when there are no images with gadolinium.

A. D vs NDNL

Metric	$T^{88}_{\text{tier 1/2}}$	Synthetic $T^{88}_{\text{tier 1/2}}$	$T^{88}_{\text{no gado, tier 1/2}}$
BA	60.26 ± 5.41	51.71 ± 1.15	51.51 ± 2.54
Sensitivity	58.68 ± 30.44	75.66 ± 34.75	6.71 ± 12.44
Specificity	61.84 ± 22.63	27.76 ± 34.98	96.32 ± 7.37

B. D vs NDL

Metric	$T^{88}_{\text{tier 1/2}}$	Synthetic $T^{88}_{\text{tier 1/2}}$	$T^{88}_{\text{no gado, tier 1/2}}$
BA	68.29 ± 3.55	54.08 ± 5.19	50.00 ± 0.00
Sensitivity	69.34 ± 7.71	52.50 ± 41.55	40.00 ± 48.99
Specificity	67.24 ± 14.43	55.66 ± 45.67	60.00 ± 48.99

TABLE 4.7: Classification performance obtained after gadolinium removal using image translation. Results were obtained for the D vs NDNL and D vs NDL classification tasks with a linear SVM using as input gray matter maps and trained on three clinical data subsets ($T^{88}_{\text{tier 1/2}}$, $T^{88}_{\text{tier 1/2}}$, $T^{88}_{\text{no gado, tier 1/2}}$). BA: balanced accuracy.

4.4.2.3 Classification performance when training on a research data set or on an unbiased clinical data set

We demonstrated that the characteristics of the training set can bias the performance of the classifier. In order to obtain unbiased results, there must be no correlation between the output and the characteristics of the images such as image quality or presence of gadolinium. One way to reach this situation is to make the training data set must be homogeneous, i.e. containing only images of the similar quality and without constrast agent. In our work, this can be obtained using either the research data set (ADNI contains only images without gadolinium and of good quality) or a clinical data set composed only of images of medium/good quality without gadolinium injection (meaning the so called Synthetic $T^{181}_{\text{tier 1/2}}$). In the previous section we demonstrated that synthetic images can suppress the effect of the gadolinium injection. We studied the performance of the classifier when using these training sets and in addition we evaluated if a CNN pre-trained on research data and fine-tuned on Synthetic $T^{181}_{\text{tier 1/2}}$ could lead to a better classification performance.

In Table 4.8, we report the results when using the two training sets for the two classification tasks, using a SVM with probability gray matter maps or a ResNet with minimally pre-processed T1w MRI. In the ResNet case, we also report the results with pre-training on research data. The best results with the ResNet were obtained using the clinical data set Synthetic $T^{181}_{\text{tier 1/2}}$ trained from scratch for D vs NDNL (balanced accuracy 63.22 ± 3.47) and pre-trained on research data for D vs NDL (balanced accuracy 68.03 ± 2.44). There was no substantial advantage in pre-training the CNN models on research data compared to training from scratch using clinical data. When using the linear SVM combined with

A. D vs NDNL

Metric	SVM		ResNet		
	Training on research data	**Synthetic $T^{181}_{\text{tier 1/2}}$**	**Training on research data**	**Pre-training on research data**	**Synthetic $T^{181}_{\text{tier 1/2}}$**
BA	64.08 ± 0.82	61.91 ± 1.34	61.84 ± 4.07	62.96 ± 2.40	63.22 ± 3.47
Sensitivity	62.76 ± 0.53	81.32 ± 2.45	60.92 ± 8.28	55.58 ± 13.93	52.24 ± 10.65
Specificity	65.39 ± 1.29	42.50 ± 4.59	62.76 ± 6.55	59.34 ± 13.18	74.21 ± 7.22

B. D vs NDL

Metric	SVM		ResNet		
	Training on research data	**Synthetic $T^{181}_{\text{tier 1/2}}$**	**Training on research data**	**Pre-training on research data**	**Synthetic $T^{181}_{\text{tier 1/2}}$**
BA	69.47 ± 0.32	64.61 ± 1.74	61.78 ± 4.35	68.03 ± 2.44	67.50 ± 0.98
Sensitivity	62.76 ± 0.53	45.53 ± 4.62	60.92 ± 8.28	59.61 ± 4.83	64.47 ± 10.47
Specificity	76.18 ± 0.49	83.68 ± 1.47	62.63 ± 4.43	76.45 ± 6.02	70.53 ± 10.05

TABLE 4.8: Classification performance when training on a research data set or on an unbiased clinical data set. Results were obtained for the D vs NDNL and D vs NDL classification tasks using a linear SVM with probability gray matter maps or a ResNet with minimally pre-processed T1w MR images. Three training setups are compared: training on research data (ADNI), training from scratch on unbiased clinical data, i.e. synthetic images without gadolinium obtained from images of medium/good quality (synthetic $T^{181}_{\text{tier 1/2}}$), and pre-training on ADNI with a fine-tuning on synthetic $T^{181}_{\text{tier 1/2}}$ (ResNet only). BA: balanced accuracy.

probability gray matter maps, models trained on research data performed slightly better than models trained on the clinical data set Synthetic $T^{181}_{\text{tier 1/2}}$. For both tasks the highest balanced accuracy was reached with the linear SVM.

These results are more reliable than those in Table 4.4 since unbiased: we can conclude that balanced accuracy scores are lower than those obtained using research data in Table 4.3.

4.5 Discussion

Research on computer vision applied to the detection of neurodegenerative diseases has been propelled by the availability of T1w brain MRI from public research data sets. In the literature, we can find several works that show promising results in the field (Samper-González et al., 2018; Falahati, Westman, and Simmons, 2014; Manera et al., 2021; Bron et al., 2021). All these studies share the same limitation: they only develop and validate machine learning and deep learning models using research data.

While translation to the clinic may seem straightforward when data are available, the difficulties are numerous. They mainly concern the definition of the different classes of interest that will represent the classification tasks, and the heterogeneity of the images.

The aim of our work is to show how the translation to the clinic can be pursued and what are the main results when applying machine learning and deep learning models to images of a clinical data set.

The first part of our work consisted in defining the cohorts that would allow us to identify patients with dementia from the other patients included in the CDW database. While in a research data set, they must be distinguished from cognitive normal subjects, in a clinical data set they had to be differentiated from patients having lesions visible in the T1w MR images and patients without any lesions. The three classes of interest (D, NDL, NDNL) were defined according to the ICD-10 codes assigned during the patients' hospitalization. When defining the criteria used to determine the D and NDL categories, we decided to be very precise in order to avoid errors in coding, so we excluded all the patients whose codes changed over the visits. Despite this, there may be some remaining errors and biases in the diagnosis as defined by the ICD-10 codes. For instance, ICD-10 codes are used for the billing of the expenses by the hospitals which may lead to biases. Assessing to which extent the diagnostic codes are biased would require to have a neurologist check the medical record of each individual patient, which is beyond the scope of our study.

The second part of our work consisted in training and applying several ML and DL classification algorithms to both research and clinical data for various scenarios.

Performance of the classifiers on research data for the detection of AD subjects were useful to set a baseline: best results for the task AD vs CN were obtained using SVM with probability gray matter maps (balanced accuracy: 86.80 $\pm$ 0.40). When the same model was applied to clinical data, the balanced accuracy decreased by about 15 percent points. Thus, ML/DL models that lead to high classification performance in a research framework, do not necessarily generalize to clinical data set. More analyses were performed in order to dissect these results.

There is a clear correlation between the diagnostic groups and the different proportions of images of bad quality and of images with gadolinium in the three classes (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 hypothesized 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 ($T^{172}_{\text{no gado}}$) or images of medium/good quality ($T^{181}_{\text{tier } 1/2}$) 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}). Thanks to these comparisons, we showed that when we used unbiased training subsets the balanced accuracy score was lower, meaning that improvements when using T^{172} was not due to the features characterising the diseases, but more to the different quality scores or the presence/absence of gadolinium.

Biased results are due to the bias in the training data set. We proposed two different solutions in order to overcome this problem. The fist was to create an unbiased training set using the clinical routine dataset: we included only images of medium/good quality and without gadolinium injection. In order to pass from images with gadolinium to images without gadolinium we applied the models proposed in (Bottani et al., 2022b): we validated

it verifying if the balanced accuracy obtained with a training subset with only images of medium/good quality without gadolinium ($T^{88}_{\text{no gado, tier 1/2}}$) was the same as that obtained with a training subset with synthetic images of medium/good quality (Synthetic $T^{88}_{\text{tier 1/2}}$). The use of synthetic images allowed us to delete the bias of gadolinium while keeping a larger number of images in the training sample.

The second solution was to use as training set the research data of ADNI: they do not contain any of the biases described above. No matter the solution implemented, balanced accuracy scores were lower than those obtained with the research data set for AD vs CN: best balanced accuracy for D vs NDNL was 64.08 ± 0.82 and for D vs NDL 69.47 ± 0.32.

These results show that translation from research to clinical routine data is not straightforward. First of all, we demonstrated the importance of having a proper training set at the expense of reducing the number of samples. Future works on images of CDW should focus more on the quality control and homogenization of the images: this could allow to obtain larger usable training set. Thereafter, the quality control should not be limited to the images, but it should include also clinical data in order to make them reliable for the labelling.

This experimental study presents some limitations. Unlike 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 hospitalization 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. 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 in order to assess if 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.

4.6 Conclusion

Computer-aided diagnosis systems for the detection of patients with dementia using T1w brain MRI data have not been validated yet using large clinical data sets. In this experimental work, we have evaluated the performance of ML/DL models in the detection of patients with dementia in a large clinical data set coming from a clinical data warehouse including 39 hospitals. In particular, we used different CNN models with minimally pre-processed T1w MRI and linear SVM with probability gray matter maps. At first we defined the classes of interest using ICD-10 codes. Then we compared the performance of the models with that obtained using a research data set. We found out that the balanced accuracy is 15 percent point lower with models trained with clinical data set if compared to that obtained with research data. Furthermore, we demonstrated that the difference in the proportions of images with gadolinium and of images of medium/good quality among the classes of the training set could bias the results. We proposed two solutions to overcome this problem: training the models using only a research data set, which does not present these biases, rather than

clinical data set, or using only images of good/medium quality without gadolinium in the training set. The latter could be deleted using a deep learning model of that translates contrast-enhanced into non-contrast-enhanced images.

Conclusion and Perspectives

Conclusion

Availability of research data set, ease of access and use, have propelled the development of CAD for the detection of neurodegenerative diseases using T1w brain MRI in the last years. The results are promising but these approaches have only been developed using research data or small clinical data sets: the usefulness of these CAD tools has yet not been demonstrated in a real clinical setting. This PhD work is a step towards the validation of ML/DL models to assist diagnosis using a very large clinical data set.

Translation from research to clinical practice is possible only if enough data are available. The development of a CDW for the Paris Greater Area Hospitals allowed the birth of the APPRIMAGE project whose final aim is to validate algorithms developed in a research context in a clinical environment. This PhD work aimed at showing how one can deal with such data, at describing the challenges and proposing methods to overcome them in order to assess the performance of ML and DL algorithms for the classification of dementia in a clinical context.

Before the classification of patients, we worked on the usability of T1w brain MRI coming from a CDW which includes 39 different hospitals (in Chapters 2 and 3). What distinguishes them from images acquired in a research context is their heterogeneity in terms of quality and sequences.

The development of an automatic quality control of the images was an essential step to continue the project. Indeed, we found out that more than 25% of the images were not proper 3D T1w brain MRI, that the DICOM header describing the sequences was not reliable and that about 30% of the images had a very low quality score. All the approaches described in the literature were not adequate for our data: they are all based on the extraction of features that are only reliable if the quality of data is good enough. The experimental approach that we developed satisfied our needs and its good performance shows that DL models can be useful in this field. Indeed the CNN models trained were able to identify images which are not proper T1w brain MRI, images of low quality and acquisitions for which gadolinium was injected with a performance score compared to that of the manual raters.

Once images had been classified according to their quality, the other problem we faced was the heterogeneity of the MRI sequences. Unlike a research data set where the acquisition protocol is well defined, CDW include images acquired using a wide range of parameters. In particular, among the T1w brain MRI of the AP-HP CDW we found images acquired both with and without gadolinium injection. These MRI sequences are used for different purposes: T1w brain MRI enhanced with gadolinium may highlight lesions such as brain tumors, while non-contrast enhanced T1w brain MRI may be useful to study the atrophy characteristic of

neurodegenerative diseases. Sequences must be homogenized in order to ensure consistency among the features extracted which could allow a CAD system to correctly recognise all the diagnoses. We developed 3D U-Net and conditional GAN models that are able to correctly synthesize images without gadolinium from contrast-enhanced images part of the AP-HP CDW. In addition, we showed that the presence of gadolinium could lead to errors when processing the images with classical neuro-imaging software tools.

The first two contributions of this PhD can already shed light on the difficulties of creating a proper training data set for the development of ML/DL models using images of a CDW. Besides, we ran into another obstacle for the development of a CAD system able to detect patients with dementia among the other patients included in the CDW: the definition of the diagnostic classes. In fact, unlike research data sets where diagnoses are defined following strict criteria identical for all the sites, in our work we used ICD-10 codes to label patients. Several choices were made in order to limit possible coding errors: this resulted in the exclusion of a large number of patients with the purpose of creating as clean a training set as possible.

The last contribution was an experimental study of the performance of CAD tools on the clinical routine dataset. We highlighted that, inhomogeneities between the classes due to differences in image quality or in the proportion of images with gadolinium were still present. We showed that such imbalances can greatly bias classification results. To overcome such bias, the training set must be homogeneous across the diagnostic classes. This is possible using images of a research data set or using clinical images only of medium/good quality passed in the image translation model to delete gadolinium. In any case, the balanced accuracy of the classifiers applied to clinical data is lower by at least 15 percent points compared to that obtained in a research framework.

Overall, our work has demonstrated the challenges posed by the design and validation of CAD algorithms on clinical routine datasets. We developed approaches for automatic quality control and image homogeneization that were critical for the rest of the project and which shall be useful for other studies using CDWs. Our experimental study of CAD algorithms demonstrated that their performance can be biased upwards by exploiting heterogeneities in the dataset. Furthermore, we demonstrated a huge drop in performance when moving from a research context to that of clinical routine datasets. This highlights the remaining challenges for bringing CAD tools to the clinic in the domain of neuroimaging of dementia.

Perspectives

This PhD is the first work, to the best of our knowledge, based on the study of T1w brain MRI from a CDW to validate ML/DL models for the computer-assisted diagnosis of neurodegenerative dementias. Despite all the difficulties encountered, it opens the way to several research directions that could be pursued in the future in order to succeed in the development of CAD in a real life environment.

Translation from research to clinical practice is not straightforward. More work should be done to improve the automatic quality control and feature homogenization frameworks.

In particular, we should improve the classification among images of good and medium quality in order to keep just the first category when training CAD systems. To this end we could work on the detection of the different artefacts that characterise the quality tiers. According to our results, motion artefact is the most difficult to detect, but its presence may also degrade the contrast of the images and it can make them appear noisy. Automatic approaches have been proposed to detect motion artefacts (Mohebbian et al., 2021; Fantini et al., 2021; Oksuz, 2021; Zhao et al., 2020; Godenschweger et al., 2016). Once detected, the quality of these images could be improved with models for image enhancement. We refer for instance to models developed for image denoising (Tamada, 2020; Manjón and Coupe, 2018) or motion artefact reduction (Higaki et al., 2019; Pawar et al., 2018; Parkes et al., 2018; Oksuz, 2021).

Image translation from contrast-enhanced to non-contrast-enhanced MRI is necessary, but it does not represent the only homogenization step that could be performed. Data set inhomogeneities can have multiple origins: images acquired with scanners of different magnetic fields (1.5 or 3 Tesla), scanners coming from different sites, different scanner models or different sequence parameters. Some strategies have been proposed in the literature for these types of homogenization (Zuo et al., 2021; Cackowski et al., 2021) and could be applied to our data. An analysis similar to that proposed in Chapter 4 could also be done to study the effects of the inhomogeneities of these characteristics.

Images are not the only data whose quality should be controlled. As mentioned in Chapter 4, the ICD-10 codes used to define the diagnostic labels may not be trustworthy. To systematically evaluate the potential biases of the ICD-10 codes, a possibility would be to systematically review the medical record by neurologists as well as the imaging data by radiologists. We could not only study the potential biases present in the ICD-10 codes as well as the inter-rater reliability of the diagnoses of the neurologists and radiologists. This is beyond the scope of the present thesis but is definitely an important avenue to assess and potentially control for biases in ICD-10 codes and thus make CDWs more reliable for computer-aided diagnosis.

Another interesting avenue is to develop an automated labelization using natural language processing models, as done also in (Wood et al., 2020; Wood et al., 2022b; Sorin et al., 2020; Senders et al., 2019). For the training of this automatic approach, we will need a training data sets where each report (i.e. the inputs of the model) correspond to a disease (i.e. the target of the model). At this aim, we should work with one or two neuro-radiologist taking into account a representative sample of the reports.

In order to be closer to clinical practice, multiple imaging modalities could be used as input of the CAD system. T2-weighted fluid-attenuated inversion recovery MR images would be particularly relevant to assess white matter hyperintensities that are characteristic of certain diseases such as vascular dementia. The addition of this MRI sequence could help the ML/DL models to be more specific towards these diseases (Wood et al., 2022a).

In conclusion, in the era of big data, there are high expectations for the development of tools that will help in diagnosing certain diseases. A lot of work is still needed to develop them in a clinical setting. While clinical data warehouses offer fantastic opportunities, they also pose considerable challenge to their use for the design and validation of computer-aided

diagnosis systems.

Appendix A

Computer-aided diagnosis of neurodegenerative diseases using machine learning and deep learning – PubMed query

This appendix provides the two PubMed queries used to obtain the graph displayed in Figure 1.

A.1 Machine learning query

("neurodegenerative" [Title] OR "dementia" [Title] OR alzheimer [Title] OR "Cognitive Impairment" [Title] OR "MCI" [Title] OR "Parkinson" [Title] OR "Huntington" OR "Posterior cortical atrophy" [Title] OR Pick [Title] OR "frontotemporal dementia" [Title] OR "Frontotemporal lobar degeneration" [Title] OR "Primary Progressive Aphasia" [Title] OR PPA [Title] OR "semantic dementia" [Title] OR "Lewy Body Dementia" [Title] OR LBD [Title] OR "vascular dementia" [Title] OR "Progressive supranuclear palsy" [Title] OR "Amyotrophic lateral sclerosis" [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])

A.2 Deep learning query

("neurodegenerative" [Title] OR "dementia" [Title] OR alzheimer [Title] OR "Cognitive Impairment" [Title] OR "MCI" [Title] OR "Parkinson" [Title] OR "Huntington" OR "Posterior cortical atrophy" [Title] OR Pick [Title] OR "frontotemporal dementia" [Title] OR "Frontotemporal lobar degeneration" [Title] OR "Primary Progressive Aphasia" [Title] OR PPA [Title] OR "semantic dementia" [Title] OR "Lewy Body Dementia" [Title] OR LBD [Title] OR "vascular dementia" [Title] OR "Progressive supranuclear palsy" [Title] OR "Amyotrophic lateral sclerosis" [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])

Bibliography

Abraham, A. et al. (2014). "Machine learning for neuroimaging with scikit-learn". In: *Frontiers in neuroinformatics* 8, p. 14.

Acharya, U. R. et al. (2018). "Deep Convolutional Neural Network for the Automated Detection and Diagnosis of Seizure Using EEG Signals". In: *Computers in Biology and Medicine* 100, pp. 270–278. DOI: 10.1016/j.compbiomed.2017.09.017.

Afonso, L. et al. (2019). "A Recurrence Plot-Based Approach for Parkinson's Disease Identification". In: *Future Generation Computer Systems* 94, pp. 282–292. DOI: 10.1016/j.future.2018.11.054.

Akkus, Z. et al. (2017). "Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence". In: *Journal of Digital Imaging* 30.4, pp. 469–476. DOI: 10.1007/s10278-017-9984-3.

Alba, X. et al. (2018). "Automatic initialization and quality control of large-scale cardiac MRI segmentations". In: *Medical image analysis* 43, pp. 129–141.

Alfaro-Almagro, F. et al. (2018). "Image Processing and Quality Control for the First 10,000 Brain Imaging Datasets from UK Biobank". In: *Neuroimage* 166, pp. 400–424.

Amara, N., O. Lamouchi, and S. Gattoufi (2020). "Design of a Breast Image Data Warehouse Framework". In: *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*. IEEE, pp. 1–13.

Aoe, J. et al. (2019). "Automatic Diagnosis of Neurological Diseases Using MEG Signals with a Deep Neural Network". In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-41500-x.

Ashburner, J. and K. J. Friston (2005). "Unified segmentation". In: *NeuroImage* 26.3, pp. 839–851.

Avants, B. B. et al. (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. et al. (2014). "The Insight ToolKit image registration framework". In: *Frontiers in Neuroinformatics* 8, p. 44.

Banerjee, D. et al. (2019). "A Deep Transfer Learning Approach for Improved Post-Traumatic Stress Disorder Diagnosis". In: *Knowledge and Information Systems* 60.3, pp. 1693–1724. DOI: 10.1007/s10115-019-01337-2.

Benou, A. et al. (2017). "Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences". In: *Medical Image Analysis* 42, pp. 145–159.

Bidani, A., M. S. Gouider, and C. M. Travieso-González (2019). "Dementia detection and classification from MRI images using deep neural networks and transfer learning". In: *International Work-Conference on Artificial Neural Networks*. Springer, pp. 925–933.

Böhle, M. et al. (2019). "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification". In: *Frontiers in aging neuroscience* 11, p. 194.

Bône, A. et al. (2021). "Contrast-enhanced brain MRI synthesis with deep learning: key input modalities and asymptotic performance". In: *2021 IEEE ISBI*.

Bottani, S. et al. (2022a). "Automatic Quality Control of Brain T1-weighted Magnetic Resonance Images for a Clinical Data Warehouse". In: *Medical Image Analysis* 75, p. 102219.

Bottani, S. et al. (2022b). "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*.

Bron, E. E. 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.

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. et al. (2021). "Deep learning for brain disorders: from data processing to disease treatment". In: *Briefings in Bioinformatics* 22.2, pp. 1560–1576.

Cackowski, S. et al. (2021). "ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization". In: *arXiv preprint arXiv:2109.06756*.

Campanella, G. et al. (2018). "Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology". In: *Computerized Medical Imaging and Graphics* 65, pp. 142–151.

Campese, S. et al. (2019). "Psychiatric Disorders Classification with 3D Convolutional Neural Networks." In: *INNSBDDL*, pp. 48–57.

Chagué, P. 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, Y. et al. (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.

Chien, Y.-W. et al. (2019). "An Automatic Assessment System for Alzheimer's Disease Based on Speech Using Feature Sequence Generator and Recurrent Neural Network". In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-56020-x.

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.

Choi, H. et al. (2017). "Refining Diagnosis of Parkinson's Disease with Deep Learning-Based Interpretation of Dopamine Transporter Imaging". In: *NeuroImage. Clinical* 16, pp. 586–594. DOI: 10.1016/j.nicl.2017.09.010.

Choi, H. et al. (2019). "Deep learning only by normal brain PET identify unheralded brain anomalies". In: *EBioMedicine* 43, pp. 447–453.

Choi, H. et al. (2020). "Cognitive Signature of Brain FDG PET Based on Deep Learning: Domain Transfer from Alzheimer's Disease to Parkinson's Disease". In: *European Journal of Nuclear Medicine and Molecular Imaging* 47.2, pp. 403–412. DOI: 10.1007/s00259-019-04538-7.

Couvy-Duchesne, B. et al. (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.

Cuingnet, R. et al. (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.

Daniel, C. and E. Salamanca (2020). "Hospital Databases". In: *Healthcare and Artificial Intelligence*. Springer, pp. 57–67.

Dar, S. U. et al. (2019). "Image synthesis in multi-contrast MRI with conditional generative adversarial networks". In: *IEEE Transactions on Medical Imaging* 38.10, pp. 2375–2388.

De Filippis, R. et al. (2019). "Machine Learning Techniques in a Structural and Functional MRI Diagnostic Approach in Schizophrenia: A Systematic Review". In: *Neuropsychiatric Disease and Treatment* 15, pp. 1605–1627. DOI: 10.2147/NDT.S202418.

Dewey, B. E. et al. (2019). "DeepHarmony: a deep learning approach to contrast harmonization across scanner changes". In: *Magnetic Resonance Imaging* 64, pp. 160–170.

Di Martino, A. et al. (2014). "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". In: *Molecular Psychiatry* 19.6, pp. 659–667.

Dinkla, A. M. et al. (2018). "MR-only brain radiation therapy: dosimetric evaluation of synthetic CTs generated by a dilated convolutional neural network". In: *International Journal of Radiation Oncology* Biology* Physics* 102.4, pp. 801–812.

Dong, J. et al. (2019). "A generic quality control framework for fetal ultrasound cardiac four-chamber planes". In: *IEEE journal of biomedical and health informatics* 24.4, pp. 931–942.

Du, J. et al. (2020). "Brain mri super-resolution using 3d dilated convolutional encoder–decoder network". In: *IEEE Access* 8, pp. 18938–18950.

Emami, H. et al. (2018). "Generating synthetic CTs from magnetic resonance images using generative adversarial networks". In: *Medical physics* 45.8, pp. 3627–3636.

Eslami, T. et al. (2019). "ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data". In: *Frontiers in Neuroinformatics* 13. DOI: 10.3389/fninf.2019.00070.

Esteban, O. et al. (2017). "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites". In: *PLOS One* 12.9, e0184661.

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.

Fantini, I. et al. (2021). "Automatic MR image quality evaluation using a Deep CNN: A reference-free method to rate motion artifacts in neuroimaging". In: *Computerized Medical Imaging and Graphics* 90, p. 101897.

Farooq, A. et al. (2017). "A deep CNN based multi-class classification of Alzheimer's disease using MRI". In: *2017 IEEE International Conference on Imaging systems and techniques (IST)*. IEEE, pp. 1–6.

Frisoni, G. B. et al. (2010). "The clinical use of structural MRI in Alzheimer disease". In: *Nature Reviews Neurology* 6.2, pp. 67–77. DOI: 10.1038/nrneurol.2009.215.

Fu, S. et al. (2019). "Natural Language Processing for the Identification of Silent Brain Infarcts from Neuroimaging Reports". In: *Journal of Medical Internet Research* 21.5. DOI: 10.2196/12109.

Gautam, R. and M. Sharma (2020). "Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis". In: *Journal of medical systems* 44.2, pp. 1–24.

Ge, C. et al. (2018). "Deep Learning and Multi-Sensor Fusion for Glioma Classification Using Multistream 2D Convolutional Networks". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Honolulu, HI: IEEE, pp. 5894–5897. DOI: 10.1109/EMBC.2018.8513556.

Ghafouri-Fard, S. et al. (2019). "Application of Single-Nucleotide Polymorphisms in the Diagnosis of Autism Spectrum Disorders: A Preliminary Study with Artificial Neural Networks". In: *Journal of Molecular Neuroscience* 68.4, pp. 515–521. DOI: 10.1007/s12031-019-01311-1.

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.

Godenschweger, F. et al. (2016). "Motion correction in MRI of the brain". In: *Physics in Medicine & Biology* 61.5, R32.

Gong, K. et al. (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.

Goodfellow, I. et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems* 27.

Gorgolewski, K. J. et al. (2016). "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". In: *Scientific data* 3.1, pp. 1–9.

Graham, M. S., I. Drobnjak, and H. Zhang (2018). "A supervised learning approach for diffusion MRI quality control with minimal training data". In: *NeuroImage* 178, pp. 668–676.

Gu, J. et al. (2019). "Deep generative adversarial networks for thin-section infant MR image reconstruction". In: *IEEE Access* 7, pp. 68290–68304.

Han, X. (2017). "MR-Based Synthetic CT Generation Using a Deep Convolutional Neural Network Method". In: *Medical Physics* 44.4, pp. 1408–1419.

Harper, L. et al. (2016). "MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases". In: *Brain* 139.4, pp. 1211–1225.

Hashimoto, F. et al. (2019). "Dynamic PET image denoising using deep convolutional neural networks without prior training datasets". In: *IEEE Access* 7, pp. 96594–96603.

He, K. et al. (2016). "Identity mappings in deep residual networks". In: *European conference on computer vision*. Springer, pp. 630–645.

Higaki, T. et al. (2019). "Improvement of image quality at CT and MRI using deep learning". In: *Japanese journal of radiology* 37.1, pp. 73–80.

Hollon, T. et al. (2020). "Near Real-Time Intraoperative Brain Tumor Diagnosis Using Stimulated Raman Histology and Deep Neural Networks". In: *Nature Medicine* 26.1, pp. 52–58. DOI: 10.1038/s41591-019-0715-9.

Huang, K.-Y., C.-H. Wu, and M.-H. Su (2019). "Attention-Based Convolutional Neural Network and Long Short-term Memory for Short-term Detection of Mood Disorders Based on Elicited Speech Responses". In: *Pattern Recognition* 88, pp. 668–678. DOI: 10.1016/j.patcog.2018.12.016.

Isensee, F. et al. (2019). "Automated brain extraction of multisequence MRI using artificial neural networks". In: *Human Brain Mapping* 40.17, pp. 4952–4964.

Işın, A., C. Direkoğlu, and M. Şah (2016). "Review of MRI-based brain tumor image segmentation using deep learning methods". In: *Procedia Computer Science* 102, pp. 317–324.

Isola, P. et al. (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. et al. (2008). "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of Magnetic Resonance Imaging* 27.4, pp. 685–691.

Janowczyk, A. et al. (2019). "HistoQC: an open-source quality control tool for digital pathology slides". In: *JCO clinical cancer informatics* 3, pp. 1–7.

Jiang, D. et al. (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.

Jónsson, B. A. et al. (2019). "Brain age prediction using deep learning uncovers associated sequence variants". In: *Nature Communications* 10.1, pp. 1–10.

Keshavan, A. et al. (2018). "Mindcontrol: A web application for brain segmentation quality control". In: *NeuroImage* 170, pp. 365–372.

Kim, H. et al. (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.

Kiryu, S. et al. (2019). "Deep learning to differentiate parkinsonian disorders separately using single midsagittal MR imaging: a proof of concept study". In: *European radiology* 29.12, pp. 6891–6899.

Klapwijk, E. T. et al. (2019). "Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data". In: *NeuroImage* 189, pp. 116–129.

Kleesiek, J. et al. (2019). "Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study". In: *Investigative Radiology* 54.10, pp. 653–660.

Koikkalainen, J. et al. (2016). "Differential diagnosis of neurodegenerative diseases using structural MRI data". In: *NeuroImage: Clinical* 11, pp. 435–449.

Kretz, T. et al. (2020). "Mammography image quality assurance using deep learning". In: *IEEE Transactions on Biomedical Engineering* 67.12, pp. 3317–3326.

Küstner, T. et al. (2018). "A machine-learning framework for automatic reference-free quality assessment in MRI". In: *Magnetic resonance imaging* 53, pp. 134–147.

Ladefoged, C. N. et al. (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.

Li, H. et al. (2019a). "DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 795–803.

Li, J. et al. (2019b). "Classifying ASD Children with LSTM Based on Raw Videos". In: *Neurocomputing*. DOI: 10.1016/j.neucom.2019.05.106.

Li, X. et al. (2016). "The first step for neuroimaging data analysis: DICOM to NIfTI conversion". In: *Journal of Neuroscience Methods* 264, pp. 47–56.

Li, Z. et al. (2017). "Deep Learning Based Radiomics (DLR) and Its Usage in Noninvasive IDH1 Prediction for Low Grade Glioma". In: *Scientific Reports* 7.1, pp. 1–11. DOI: 10.1038/s41598-017-05848-2.

Littlejohns, T. J. et al. (2020). "The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions". In: *Nature Communications* 11.1, pp. 1–12.

Ma, D. 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.

Manera, A. L. 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.

Manjón, J. V. and P. Coupe (2018). "MRI denoising using deep learning". In: *International Workshop on Patch-based Techniques in Medical Imaging*. Springer, pp. 12–19.

Mao, X. et al. (2017). "Least squares generative adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802.

Mark, J. et al. (2012). "FSL". In: *NeuroImage* 62.2, pp. 782–790.

Marzullo, A. et al. (2019). "Classification of Multiple Sclerosis Clinical Profiles via Graph Convolutional Neural Networks". In: *Frontiers in Neuroscience* 13.JUN. DOI: 10.3389/fnins.2019.00594.

Mason, D. (2011). "SU-E-T-33: pydicom: an open source DICOM library". In: *Medical Physics* 38.6Part10, pp. 3493–3493.

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.

Mirza, M. and S. Osindero (2014). "Conditional generative adversarial nets". In: *arXiv:1411.1784*.

Mohebbian, M. et al. (2021). "Classifying MRI motion severity using a stacked ensemble approach". In: *Magnetic Resonance Imaging* 75, pp. 107–115.

Moon, S. et al. (2019). "Accuracy of Machine Learning Algorithms for the Diagnosis of Autism Spectrum Disorder: Systematic Review and Meta-Analysis of Brain Magnetic Resonance Imaging Studies". In: *Journal of Medical Internet Research* 21.12. DOI: 10.2196/14108.

Morin, A. 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.

Nakao, T. et al. (2018). "Deep Neural Network-Based Computer-Assisted Detection of Cerebral Aneurysms in MR Angiography". In: *Journal of Magnetic Resonance Imaging* 47.4, pp. 948–953. DOI: 10.1002/jmri.25842.

Naseer, A. et al. (2020). "Refining Parkinson's Neurological Disorder Identification through Deep Transfer Learning". In: *Neural Computing and Applications* 32.3, pp. 839–854. DOI: 10.1007/s00521-019-04069-0.

Neppl, S. 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. et al. (2018). "Medical image synthesis with deep convolutional adversarial networks". In: *IEEE Transactions on Biomedical Engineering* 65.12, pp. 2720–2730.

Noor, M. B. T. et al. (2019). "Detecting neurodegenerative disease from MRI: A brief review on a deep learning perspective". In: *International Conference on Brain Informatics*. Springer, pp. 115–125.

Oh, K. et al. (2019). "Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization". In: *Schizophrenia Research* 212, pp. 186–195.

Oksuz, I. (2021). "Brain MRI artefact detection and correction using convolutional neural networks". In: *Computer Methods and Programs in Biomedicine* 199, p. 105909.

Oksuz, I. et al. (2019). "Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning". In: *Medical image analysis* 55, pp. 136–147.

Oktay, O. et al. (2018). "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999*.

Parkes, L. et al. (2018). "An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI". In: *Neuroimage* 171, pp. 415–436.

Paszke, A. et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32.

Pawar, K. et al. (2018). "Motion correction in MRI using deep convolutional neural network". In: *Proceedings of the ISMRM Scientific Meeting & Exhibition, Paris*. Vol. 1174.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Pellegrini, E. et al. (2018). "Machine Learning of Neuroimaging for Assisted Diagnosis of Cognitive Impairment and Dementia: A Systematic Review". In: *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring* 10, pp. 519–535. DOI: 10.1016/j.dadm.2018.07.004.

Penny, W. D. et al. (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Pham, C.-H. et al. (2017). "Brain MRI super-resolution using deep 3D convolutional networks". In: *2017 IEEE ISBI*, pp. 197–200.

Punjabi, A. et al. (2019). "Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks". In: *PloS one* 14.12, e0225759.

Raamana, P. R. et al. (2020). "Visual QC Protocol for FreeSurfer Cortical Parcellations from Anatomical MRI". In: *bioRxiv*.

Ran, M. et al. (2019). "Denoising of 3D magnetic resonance images using a residual encoder–decoder Wasserstein generative adversarial network". In: *Medical image analysis* 55, pp. 165–180.

Rathore, S. et al. (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.

Reuter, M. et al. (2015). "Head motion during MRI acquisition reduces gray matter volume and thickness estimates". In: *NeuroImage* 107, pp. 107–115.

Robinson, R. et al. (2018). "Real-time prediction of segmentation quality". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 578–585.

Robinson, R. et al. (2019). "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study". In: *Journal of Cardiovascular Magnetic Resonance* 21.1, pp. 1–14.

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.

Routier, A. et al. (2021). "Clinica: An Open Source Software Platform for Reproducible Clinical Neuroscience Studies". In: *hal-02308126*.

Sadri, A. R. et al. (2020). "MRQy—An open-source tool for quality control of MR imaging data". In: *Medical Physics* 47.12, pp. 6029–6038.

Samper-González, J. et al. (2018). "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data". In: *NeuroImage* 183, pp. 504–521.

San-Segundo, R. et al. (2019). "Classification of Epileptic EEG Recordings Using Signal Transforms and Convolutional Neural Networks". In: *Computers in Biology and Medicine* 109, pp. 148–158. DOI: 10.1016/j.compbiomed.2019.04.031.

Senders, J. T. et al. (2019). "Natural language processing for automated quantification of brain metastases reported in free-text radiology reports". In: *JCO Clinical Cancer Informatics* 3, pp. 1–9.

Seo, M. et al. (2021). "Neural Contrast Enhancement of CT Image". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3973–3982.

Sharma, A. and G. Hamarneh (2019). "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network". In: *IEEE transactions on medical imaging* 39.4, pp. 1170–1183.

Shinde, S. et al. (2019). "Predictive Markers for Parkinson's Disease Using Deep Neural Nets on Neuromelanin Sensitive MRI". In: *NeuroImage. Clinical* 22, p. 101748. DOI: 10.1016/j.nicl.2019.101748.

Shiri, I. et al. (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.

Silva, I. R. et al. (2019). "Model based on deep feature extraction for diagnosis of Alzheimer's disease". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–7.

Sorin, V. et al. (2020). "Deep learning for natural language processing in radiology—fundamentals and a systematic review". In: *Journal of the American College of Radiology* 17.5, pp. 639–648.

Spasov, S. et al. (2019). "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease". In: *Neuroimage* 189, pp. 276–287.

Spuhler, K. D. et al. (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.

Sujit, S. J. et al. (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. et al. (2020). "Substituting Gadolinium in Brain MRI Using DeepContrast". In: *2020 IEEE ISBI*, pp. 908–912.

Sunoqrot, M. R. et al. (2020). "A quality control system for automated prostate segmentation on T2-weighted MRI". In: *Diagnostics* 10.9, p. 714.

Szegedy, C. et al. (2016). "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

Tamada, D. (2020). "Noise and artifact reduction for MRI using deep learning". In: *arXiv preprint arXiv:2002.12889*.

Tayari, N. et al. (2019). "Simple and broadly applicable automatic quality control for 3D 1H MR spectroscopic imaging data of the prostate". In: *Magnetic resonance in medicine* 81.5, pp. 2887–2895.

Tustison, N. J. et al. (2010). "N4ITK: improved N3 bias correction". In: *IEEE Transactions on Medical Imaging* 29.6, pp. 1310–1320.

Ueda, D. et al. (2019). "Deep Learning for MR Angiography: Automated Detection of Cerebral Aneurysms". In: *Radiology* 290.1, pp. 187–194. DOI: 10.1148/radiol.2018180901.

Wada, A. et al. (2019). "Differentiating Alzheimer's Disease from Dementia with Lewy Bodies Using a Deep Learning Technique Based on Structural Brain Connectivity". In: *Magnetic Resonance in Medical Sciences* 18.3, pp. 219–224. DOI: 10.2463/mrms.mp.2018-0091.

Wang, W. et al. (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. et al. (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 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.

Wegmayr, V., S. Aitharaju, and J. Buhmann (2018). "Classification of brain MRI with big data and deep 3D convolutional neural networks". In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. International Society for Optics and Photonics, 105751S.

Wei, W. et al. (2019). "Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis". In: *Medical image analysis* 58, p. 101546.

Wen, J. et al. (2020). "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation". In: *Medical Image Analysis*, p. 101694.

Wolterink, J. M. et al. (2017). "Deep MR to CT synthesis using unpaired data". In: *International workshop on simulation and synthesis in medical imaging*. Springer, pp. 14–23.

Wood, D. A. et al. (2020). "Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM)". In: *Medical Imaging with Deep Learning*. PMLR, pp. 811–826.

Wood, D. A. et al. (2022a). "Accurate brain-age models for routine clinical MRI examinations". In: *NeuroImage*, p. 118871.

Wood, D. A. et al. (2022b). "Deep learning to automate the labelling of head MRI datasets for computer vision applications". In: *European Radiology* 32.1, pp. 725–736.

World Health Organization et al. (2007). "International classification of diseases and related health problems, 10$^{\text{th}}$ revision". In: *http://www. who. int/classifications/apps/icd/icd10online*.

Xiao, Z. et al. (2018). "SAE-based Classification of School-Aged Children with Autism Spectrum Disorders Using Functional Magnetic Resonance Imaging". In: *Multimedia Tools and Applications* 77.17, pp. 22809–22820. DOI: 10.1007/s11042-018-5625-1.

Xu, C. et al. (2021). "Synthesis of gadolinium-enhanced liver tumors on nonenhanced liver MR images using pixel-level graph reinforcement learning". In: *Medical Image Analysis* 69, p. 101976.

Yang, J. et al. (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, Q. et al. (2018). "MRI cross-modality neuroimage-to-neuroimage translation". In: *arXiv:1801.06940*.

Yang, Z. et al. (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, H. et al. (2019). "Precise Diagnosis of Intracranial Hemorrhage and Subtypes Using a Three-Dimensional Joint Convolutional and Recurrent Neural Network". In: *European Radiology* 29.11, pp. 6191–6201. DOI: 10.1007/s00330-019-06163-2.

Yi, X., E. Walia, and P. Babyn (2019). "Generative adversarial network in medical imaging: A review". In: *Medical image analysis* 58, p. 101552.

Yu, B. et al. (2019). "Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis". In: *IEEE transactions on medical imaging* 38.7, pp. 1750–1762.

Zeng, K. et al. (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.

Zhang, J. et al. (2019a). "Three dimensional convolutional neural network-based classification of conduct disorder with structural MRI". In: *Brain imaging and behavior*, pp. 1–8.

Zhang, X. et al. (2019b). "Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study". In: *Scientific Reports* 9.1, p. 797. DOI: 10.1038/s41598-018-37545-z.

Zhao, H. et al. (2016). "Loss functions for image restoration with neural networks". In: *IEEE Transactions on computational imaging* 3.1, pp. 47–57.

Zhao, Y. et al. (2020). "Localized motion artifact reduction on brain MRI using deep learning with effective data augmentation techniques". In: *arXiv preprint arXiv:2007.05149*.

Zhu, J.-Y. et al. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zou, L. et al. (2017). "3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI". In: *IEEE Access* 5, pp. 23626–23636. DOI: 10.1109/ACCESS.2017.2762703.

Zuo, L. et al. (2021). "Information-Based Disentangled Representation Learning for Unsupervised MR Harmonization". In: *International Conference on Information Processing in Medical Imaging*. Springer, pp. 346–359.