



Deep learning for the detection of neurological diseases

Huy-Dung Nguyen

► To cite this version:

Huy-Dung Nguyen. Deep learning for the detection of neurological diseases. Image Processing [eess.IV]. Université de Bordeaux, 2023. English. NNT : 2023BORD0288 . tel-04311995

HAL Id: tel-04311995

<https://theses.hal.science/tel-04311995>

Submitted on 28 Nov 2023

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

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

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE
DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX
ÉCOLE DOCTORALE MATHÉMATIQUES ET
INFORMATIQUE

SPÉCIALITÉ : INFORMATIQUE

Par **Huy-Dung NGUYEN**

**Apprentissage profond pour la prédiction des maladies
neurologiques**

Soutenue le 08/11/2023

Membres du jury :

| | | |
|--------------------------|---|--------------------|
| Dr. Louis COLLINS | Professeur, Université de McGill | Rapporteur |
| Dr. Olivier COLLIOT | Directeur de recherche, CNRS, Université de Sorbonne | Rapporteur |
| Dr. Gwenaëlle CATHELINE | Professeure, Université de Bordeaux | Examinatrice |
| Dr. Aurélie BUGEAU | Professeure, Université de Bordeaux | Examinatrice |
| Dr. Jean-François MANGIN | Directeur de recherche, CEA, Université de Paris-Saclay | Examineur |
| Dr. Pierrick COUPÉ | Directeur de recherche, CNRS, Université de Bordeaux | Directeur de thèse |
| Dr. Michaël CLÉMENT | Maître de conférences, Bordeaux INP | Invité (Encadrant) |

Résumé : La détection des maladies neurologiques est cruciale pour améliorer la qualité de vie des patients et réduire la charge économique sur les systèmes de santé. De nos jours, l'Intelligence Artificielle (IA) joue un rôle essentiel dans l'analyse des données médicales, notamment l'Imagerie par Résonance Magnétique (IRM), qui est couramment utilisée pour diagnostiquer les maladies neurologiques. Ce type d'image est aujourd'hui produit à grande échelle, rendant l'analyse manuelle impossible. En conséquence, de nombreuses méthodes basées sur l'Apprentissage Profond (AP), la dernière et la plus puissante technique d'IA, ont été proposées pour la détection automatisée des maladies neurologiques à l'aide de données d'IRM. Cependant, les méthodes actuelles basées sur le DL présentent plusieurs limitations.

Tout d'abord, les performances des méthodes actuelles basées sur l'AP restent limitées par rapport aux approches traditionnelles d'apprentissage automatique. Deuxièmement, la capacité de généralisation des méthodes d'AP sur des données externes reste un défi. Troisièmement, d'un point de vue clinique, ces méthodes manquent souvent du niveau de compréhension souhaité, ce qui limite leur utilité pratique dans les applications cliniques. Enfin, ces méthodes se concentrent généralement uniquement sur le diagnostic des maladies individuelles, limitant ainsi la compréhension des différences entre les maladies. En conséquence, l'objectif de ce doctorat était de surmonter ces limitations et de développer une nouvelle génération de méthodes de diagnostic des maladies démontrant une bonne performance, capacité de généralisation et compréhension dans divers scénarios cliniques, comprenant à la fois le diagnostic de maladies individuelles et de multiple maladies.

Pour atteindre ces objectifs, un plan de recherche structuré a été conçu. La première étude a introduit un nouveau biomarqueur - appelé deep grading - pour le diagnostic et le pronostic précis de la maladie d'Alzheimer (AD). S'appuyant sur ce deep grading, la deuxième étude a étendu le biomarqueur pour permettre la discrimination entre les individus ayant une cognition normale (CN), l'AD et la démence frontotemporale (FTD). Dans cette optique, nous avons proposé un deep grading à canaux multiples capable de gérer la classification de plusieurs maladies. La troisième étude a proposé un nouveau biomarqueur - appelé brain structure ages - capable de traiter un grand nombre de pathologies en même temps. De plus, dans cette étude, nous avons abordé les défis liés à l'interprétabilité lors du traitement d'un plus grand nombre de maladies, en nous concentrant sur le diagnostic différentiel de la CN, de l'AD, de la FTD, de la sclérose en plaques (MS), de la maladie de Parkinson (PD) et de la schizophrénie (SZ) en tant qu'application de preuve de concept. Enfin, la quatrième étude a été menée pour explorer le potentiel des transformers pour le diagnostic de plusieurs maladies, qui ont récemment montré des résultats prometteurs dans diverses tâches de vision par ordinateur. Une comparaison entre nos méthodes de réseau de neurones convolutifs et de transformers a ensuite été effectuée selon différents critères, y compris la capacité de généralisation et la capacité de compréhension du modèle.

Les méthodes développées ont été intégrées dans la plateforme VolBrain. VolBrain est un système en ligne d'imagerie volumétrique du cerveau par IRM pour mission d'aider les chercheurs du monde entier à obtenir automatiquement des informations volumétriques sur le cerveau à partir de leurs données d'IRM sans aucun effort.

Mots-clés : Apprentissage profond, Maladies neurologiques, Diagnostic de maladie

Deep learning for the detection of neurological diseases

Abstract: The detection of neurological diseases is crucial for improving patients' quality of life and reducing the economic burden on healthcare systems. Nowadays, Artificial Intelligence (AI) plays a vital role in the analysis of medical data, including Magnetic Resonance Imaging (MRI), which is commonly used for diagnosing neurological diseases. This type of image is today produced at large scale, making the manual analysis infeasible. Consequently, numerous Deep Learning (DL)-based methods, the latest and most powerful AI technique, have been proposed for automated detection of neurological diseases using MRI data. However, the current DL-based methods present several limitations.

Firstly, the performance of current DL-based methods remains limited compared to traditional machine learning approaches. Secondly, the generalization capacity of DL methods on external data is still a challenge. Thirdly, from a clinical perspective, these methods often lack the desired level of understandability, which limits their practical utility in clinical applications. Lastly, these methods commonly focus solely on diagnosing single diseases, thereby limiting the comprehension of differences between diseases. Consequently, the objective of this PhD was to overcome these limitations and develop a new generation of methods for disease diagnosis that demonstrates high performance, generalizability and understandability in diverse clinical scenarios encompassing both single-disease and multi-disease diagnosis.

To achieve these goals, a structured research plan was designed. The first study introduced a novel biomarker - called deep grading - for the accurate diagnosis and prognosis of Alzheimer's Disease (AD). Building upon this deep grading, the second study extended the biomarker to enable the discrimination of Cognitively Normal (CN), AD and Frontotemporal Dementia (FTD). To this end, we proposed a multi-channel deep grading able to handle multi-disease classification. The third study proposed a novel biomarker - called brain structure ages - able to deal with a high number of pathologies at the same time. Moreover, in this study we addressed the interpretability challenges when dealing with a larger number of diseases, focusing on the differential diagnosis of CN, AD, FTD, Multiple Sclerosis (MS), Parkinson's Disease (PD), and Schizophrenia (SZ) as a proof-of-concept application. Lastly, the fourth study was conducted to explore the potential of transformers for multi-disease diagnosis, which have recently shown promising results in various computer vision tasks. A comparison between our CNN and transformer methods was then carried out based on different criteria, including generalization capacity and model understanding capacity.

The developed methods have been integrated in the VolBrain platform. VolBrain is an online open MRI brain volumetry system with the mission of helping researchers all over the world to obtain automatically volumetric brain information from their MRI data without effort.

Keywords: Deep learning, Neurological diseases, Disease diagnosis

Unité de recherche

Univ. Bordeaux, CNRS, LaBRI, UMR 5800, 34000 Talence, France.

Acknowledgments

I am grateful for the support and assistance of many individuals during my PhD journey. First and foremost, I would like to express my deepest appreciation to my supervisor, Dr. Pierrick Coupé, for his invaluable guidance, mentorship, and unwavering support throughout my PhD journey. His extensive knowledge, expertise, and constructive feedback have been instrumental in shaping my research project and helping me to achieve my goals.

I am also deeply grateful to my co-supervisor, Dr. Michaël Clément, for his technical expertise, insightful suggestions, and mental support throughout this journey. His guidance and encouragement have been invaluable to me, especially during challenging times.

I would like to express my appreciation to all the members of my thesis committee. I would like to thank Dr. Louis Collins from McGill University, Dr. Olivier Colliot from Sorbonne University, Dr. Gwenaëlle Catheline, Dr. Aurélie Bugeau from University of Bordeaux, and Dr. Jean-François Mangin from Paris-Saclay University for spending their valuable time reviewing my thesis.

I would also like to express my sincere gratitude to my colleagues, Mr. Boris Mansencal and Dr. Vincent Planche. Mr. Boris Mansencal provided invaluable assistance in preparing and organizing my data, and his feedback on my paper revisions was essential in improving the quality of my research. Dr. Vincent Planche provided valuable insights into the clinical aspects of my research, and his feedback was crucial in helping me to validate the results of my work.

I am also thankful to my colleagues and also my friends: PhD Réda Abdellah Kamraoui, PhD Vincent Martin and many others who made this journey exceptional.

Finally, I would like to express my gratitude to my family, for their unwavering love, support, and encouragement. Their belief in me has been a source of inspiration and motivation.

Scientific Production

JOURNAL PAPERS

- [1] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Towards better interpretable and generalizable AD detection using collective artificial intelligence”. In: *Computerized Medical Imaging and Graphics* 104 (2023), p. 102171. DOI: [10.1016/j.compmedimag.2022.102171](https://doi.org/10.1016/j.compmedimag.2022.102171).
- [2] **Nguyen, Huy-Dung**, Michaël Clément, Vincent Planche, Boris Mansencal, and Pierrick Coupé. “Deep grading for MRI-based differential diagnosis of Alzheimer’s disease and Frontotemporal dementia”. In: *Artificial Intelligence in Medicine* 144 (2023), p. 102636. DOI: [10.1016/j.artmed.2023.102636](https://doi.org/10.1016/j.artmed.2023.102636).

SUBMITTED JOURNAL PAPERS

- [3] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Brain Structure Ages - A new biomarker for multi-disease classification”. In: *HAL preprint* (2023). Under revision in *Human Brain Mapping*. eprint: [hal-04080401](https://hal.archives-ouvertes.fr/hal-04080401).

PEER-REVIEWED CONFERENCE PROCEEDINGS

- [4] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Deep Grading Based on Collective Artificial Intelligence for AD Diagnosis and Prognosis”. In: *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. IMIMIC 2021*. 2021. DOI: [10.1007/978-3-030-87444-5_3](https://doi.org/10.1007/978-3-030-87444-5_3).
- [5] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Interpretable Differential Diagnosis for Alzheimer’s Disease and Frontotemporal Dementia”. In: *Medical Image Computing and Computer Assisted Intervention. MICCAI 2022*. Vol. 13431. 2022. DOI: [10.1007/978-3-031-16431-6_6](https://doi.org/10.1007/978-3-031-16431-6_6).

- [6] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “3D Transformer based on deformable patch location for differential diagnosis between Alzheimer’s disease and Frontotemporal dementia”. In: *The 14th International Workshop on Machine Learning in Medical Imaging. MLMI 2023, Held in Conjunction with MICCAI 2023*. Vol. 14349. 2023, pp. 53–63. DOI: [10.1007/978-3-031-45676-3_6](https://doi.org/10.1007/978-3-031-45676-3_6).

PROTECTED SOFTWARE

- [7] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *AssemblyNet-AD Automatic Diagnosis of Dementia*. IDDN.FR.001.190026.000.S.C.2022.000.31230.
- [8] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *AssemblyNet-AD-FTD*. Under deposit.
- [9] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *BrainDiseaseDiagnosis*. Under deposit.
- [10] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *BrainStructureAges*. Under deposit.

TALKS AND POSTERS

1. Oral presentation - *Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2021*, 2021, Strasbourg, France.
2. Poster - *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2022, Singapore.
3. Poster - *The 14th International Workshop on Machine Learning in Medical Imaging. MLMI, held in Conjunction with MICCAI*, 2023, Vancouver, Canada.

Table of contents

| | |
|--|-----------|
| Introduction | 5 |
| Overview of neurodegenerative diseases | 5 |
| Scientific challenges | 13 |
| Contributions and outline | 15 |
| 1 Materials and methods | 20 |
| 1.1 Materials | 21 |
| 1.1.1 Datasets used in this manuscript | 21 |
| 1.1.2 Preprocessing steps for structural MRI | 23 |
| 1.2 Common automatic approaches | 26 |
| 1.2.1 Classical Machine Learning | 26 |
| 1.2.2 Deep Learning | 27 |
| 2 Deep grading for single-disease diagnosis | 29 |
| 2.1 Introduction | 30 |
| 2.1.1 Context | 30 |
| 2.1.2 Related works | 31 |
| 2.1.3 Current limitations of DL in AD classification | 33 |
| 2.1.4 Contributions | 35 |
| 2.2 Materials | 36 |
| 2.3 Methods | 36 |
| 2.3.1 Method overview | 36 |

| | | |
|----------|--|-----------|
| 2.3.2 | Deep grading | 38 |
| 2.3.3 | Collective AI | 39 |
| 2.3.4 | Feature classification | 40 |
| 2.3.5 | Implementation details | 40 |
| 2.4 | Experimental results | 41 |
| 2.4.1 | Performance study | 41 |
| 2.4.2 | Interpretation of deep grading maps | 49 |
| 2.4.3 | Consistency study | 50 |
| 2.5 | Discussion | 52 |
| 3 | Multi-channel deep grading for differential diagnosis | 55 |
| 3.1 | Introduction | 57 |
| 3.2 | Materials | 58 |
| 3.3 | Method description | 59 |
| 3.3.1 | Method overview | 59 |
| 3.3.2 | Multi-class Deep Grading-based classification | 59 |
| 3.3.3 | Atrophy-based classification | 61 |
| 3.3.4 | Implementation details | 61 |
| 3.4 | Experimental results | 62 |
| 3.4.1 | Ablation study for binary classification tasks | 63 |
| 3.4.2 | Performance for multi-disease classification | 64 |
| 3.4.3 | Comparison with state-of-the-art methods | 65 |
| 3.4.4 | Interpretation of deep grading map | 67 |
| 3.5 | Discussion | 69 |
| 4 | Brain structure ages for differential diagnosis | 71 |
| 4.1 | Introduction | 73 |
| 4.2 | Materials | 76 |
| 4.2.1 | Chronological age prediction | 76 |

| | | |
|----------|--|-----------|
| 4.2.2 | Multiple pathologies classification | 76 |
| 4.3 | Methods | 77 |
| 4.3.1 | Method overview | 77 |
| 4.3.2 | Implementation details | 79 |
| 4.3.3 | Validation Framework | 81 |
| 4.4 | Experimental results | 81 |
| 4.4.1 | Chronological age estimation | 81 |
| 4.4.2 | Disease classification | 83 |
| 4.4.3 | Predicted brain age of different populations | 86 |
| 4.4.4 | Interpretation of brain structure age gap estimation | 86 |
| 4.5 | Discussion | 87 |
| 5 | Transformer for differential diagnosis | 90 |
| 5.1 | Introduction | 91 |
| 5.2 | Materials | 92 |
| 5.3 | Method description | 93 |
| 5.3.1 | Model overview | 93 |
| 5.3.2 | Data augmentation | 96 |
| 5.3.3 | Validation framework and ensembling | 96 |
| 5.3.4 | Implementation details | 96 |
| 5.4 | Experimental results | 97 |
| 5.4.1 | Ablation study | 97 |
| 5.4.2 | Comparison with state-of-the-art methods | 98 |
| 5.4.3 | Visualization of deformable patch location | 99 |
| 5.5 | Comparison between the proposed Transformer and CNNs | 100 |
| 5.5.1 | Performance comparison | 100 |
| 5.5.2 | Comparison of Interpretability and Explainability | 103 |
| 5.6 | Discussion | 105 |

| | | |
|----------|---|------------|
| 6 | Conclusion and perspectives | 109 |
| 6.1 | Conclusion | 109 |
| 6.2 | Future works and perspectives | 111 |
| A | Appendix for Deep Grading | 113 |
| A.1 | Performance measures using the AUC metric | 113 |
| A.2 | Cross-brain regions connectivity analysis | 117 |
| B | Appendix for multi-channel deep grading | 119 |
| C | Appendix for brain structure ages | 123 |

List of Figures

| | | |
|-----|---|----|
| 1 | Examples of T1w and T2w images. | 8 |
| 2 | Visualization of a normal brain and a brain with AD using sMRI. | 9 |
| 3 | Visualization of a normal brain and a brain with FTD using sMRI. | 10 |
| 4 | Visualization of a normal brain and a brain with MS using sMRI. | 11 |
| 5 | Visualization of a normal brain and a brain with PD using sMRI. | 12 |
| 6 | Visualization of a normal brain and a brain with SZ using sMRI. | 13 |
| 1.1 | Overview of the preprocessing pipeline. | 24 |
| 1.2 | Illustration of MLP. | 27 |
| 1.3 | Illustration of CNN and Trandformer. | 28 |
| 2.1 | Overview of our pipeline for Alzheimer’s Disease classification. | 37 |
| 2.2 | UMAP visualization of test set for the AD diagnosis and prognosis. | 44 |
| 2.3 | Average grading map per group of subjects (CN, sMCI, pMCI and AD). | 50 |
| 2.4 | Typical grading maps (from individual subjects) for each state of Alzheimer’s disease with respect to age. | 51 |
| 2.5 | Consistency of grading maps between retrained grading models, and between grading models trained on different datasets. | 52 |
| 3.1 | An overview of the proposed multi-channel grading method. | 60 |
| 3.2 | UMAP visualization of grades, volumes and the ensemble on out-of-domain data (NACC). | 65 |
| 3.3 | Average grading map per group of subjects in the MNI152 space with neurological orientation | 67 |

| | | |
|-----|---|-----|
| 3.4 | Average grading map per variant of FTD in the MNI152 space with neurological orientation | 68 |
| 3.5 | Individual grading maps of each group (CN, AD and FTD) of subjects with respect to age. | 69 |
| 4.1 | An overview of the proposed chronological age prediction and multi-disease classification. | 79 |
| 4.2 | Architecture of an unit U-Net used for voxel-level age prediction. | 80 |
| 4.3 | UMAP visualization of BSAGE, volume and the combination | 86 |
| 4.4 | Predicted brain age of different populations in out-of-domain data. | 87 |
| 4.5 | BSAGE of different populations in out-of-domain data. | 88 |
| 5.1 | The architecture of our transformer model | 93 |
| 5.2 | Details of different modules in our transformer model | 94 |
| 5.3 | Visualization of deformable patch locations. The importance of each patch was estimated with GradCAM. Warmer color, larger radius mean higher importance score. | 100 |
| 5.4 | Visualization of deformable patch locations for different AD and FTD (top) and interpretable map using deep grading features (bottom). | 104 |
| 5.5 | Visualization of deformable patch locations for different subtypes of FTD (top) and interpretable map using deep grading features (bottom). | 105 |
| 5.6 | Explainable map generated for AD patients by our transformer method (top) and interpretable map using deep grading features (bottom). | 106 |
| 5.7 | Explainable map generated for sMCI, pMCI and AD patients by our transformer method (top) and interpretable map using deep grading features (bottom). | 107 |
| 5.8 | Explainable map generated for AD, FTD, MS, PD and SZ patients by our transformer method | 108 |
| A.1 | Averaged adjacency matrices of AD population and CN population. | 118 |
| B.1 | Our data split procedure | 119 |

List of Tables

| | | |
|-----|---|----|
| 3 | Overview of the different studies achieved during the PhD. | 17 |
| 1.1 | Summary of public datasets used in this manuscript. | 22 |
| 2.1 | Summary of participants used in our study | 36 |
| 2.2 | Comparison of different types of features for classification for AD diagnosis and prognosis | 43 |
| 2.3 | Comparison of different graph edge types for AD diagnosis and prognosis . | 46 |
| 2.4 | Comparison of different classifiers for AD diagnosis and prognosis | 47 |
| 2.5 | Comparison of our method with state-of-the-art methods with available code that have been retrained on our training dataset and tested on our dataset | 49 |
| 2.6 | Comparison of our method with state-of-the-art methods using published results | 50 |
| 3.1 | Summary of participants used in our study. | 59 |
| 3.2 | Ablation study of our method for binary classification tasks. | 63 |
| 3.3 | Performance of different models for the multi-disease classification CN <i>vs.</i> AD <i>vs.</i> FTD. | 64 |
| 3.4 | Comparison of our method with current state-of-the-art methods for binary classification tasks. | 66 |
| 3.5 | Comparison of our method with current state-of-the-art methods for 3-class differential diagnosis AD <i>vs.</i> FTD <i>vs.</i> CN. | 67 |
| 4.1 | Summary of participants for the age prediction task. | 77 |
| 4.2 | Number of participants (Male/Female) used for multi-class classification. . | 78 |

| | | |
|-----|---|-----|
| 4.3 | Ablation study for the chronological age estimation. | 82 |
| 4.4 | Comparison with state-of-the-art methods for the age estimation task. . . . | 83 |
| 4.5 | Ablation study for binary classification tasks. | 84 |
| 4.6 | Multi-disease classification results for CN <i>vs.</i> AD <i>vs.</i> FTD <i>vs.</i> MS <i>vs.</i> PD <i>vs.</i> SZ. | 85 |
| 5.1 | Summary of participants used in our study. | 93 |
| 5.2 | Ablation study of the model performance for the differential diagnosis CN <i>vs.</i> AD <i>vs.</i> FTD. | 98 |
| 5.3 | Ablation study of the data augmentation on our model performance for the differential diagnosis CN <i>vs.</i> AD <i>vs.</i> FTD. | 99 |
| 5.4 | Comparison with state-of-the-art methods for the differential diagnosis CN <i>vs.</i> AD <i>vs.</i> FTD. | 99 |
| 5.5 | Performance comparison between our proposed transformer and CNN for the classification CN <i>vs.</i> AD <i>vs.</i> FTD. | 101 |
| 5.6 | Comparison of the CNN and the Transformer method for AD diagnosis and prognosis. | 102 |
| 5.7 | Comparison of our BSAGE and Transformer features for multi-disease classification CN <i>vs.</i> AD <i>vs.</i> FTD <i>vs.</i> MS <i>vs.</i> PD <i>vs.</i> SZ. | 103 |
| A.1 | Comparison of different types of features for classification using AUC metric | 114 |
| A.2 | Comparison of different graph edge types using AUC metric | 115 |
| A.3 | Comparison of different classifiers using AUC metric | 116 |
| A.4 | Comparison of our method with state-of-the-art methods that have been retrained on our training dataset using the available code and tested on our dataset (AUC metric) | 117 |
| A.5 | Comparison of our method with state-of-the-art methods using published results (AUC metric) | 117 |
| B.1 | Ablation study of our method for binary classification tasks (using AUC). . | 120 |
| B.2 | Ablation study of our method for binary classification tasks (using AUC). . | 121 |
| B.3 | Comparison of our method with current state-of-the-art methods for binary classification tasks (using AUC). | 122 |

| | | |
|-----|---|-----|
| B.4 | Comparison of our method with current state-of-the-art methods for binary classification tasks (using AUC). | 122 |
| C.1 | Ablation study for binary classification tasks using ACC. | 123 |
| C.2 | Ablation study for binary classification tasks using AUC. | 124 |

List of Abbreviations

Abbreviations and their completed forms

| Abbreviations | Completed form |
|---------------------|--|
| ABIDE | Autism Brain Imaging Data Exchange |
| ACC | ACCuracy |
| AD | Alzheimer's Disease |
| ADNI | Alzheimer's Disease Neuroimaging Initiative |
| AI | Artificial Intelligence |
| AIBL | Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing |
| AUC | Area Under receiver operating characteristic Curve |
| BA | Brain Age |
| BACC | Balanced ACCuracy |
| BSA | Brain Structure Age |
| BSAGE | Brain Structure Age Gap Estimation |
| BrainAGE | Brain Age Gap Estimation |
| BrainGluSchi | Brain study based on Glutamatergic for Schizophrenia |
| C-MIND | Cincinnati MR Imaging of Neurodevelopment |
| CAM | Class Activation Maps |
| CN/HC | Cognitively Normal/Healthy Control |
| CNN | Convolutional Neural Network |
| COBRE | Center for Biomedical Research Excellence |
| DC | Disease Coordinate |
| DG | Deep Grading |
| DL | Deep learning |
| FTD | FrontoTemporal Dementia |
| GAN | Generative Adversarial Network |
| GB | Guided Backpropagation |
| GCN | Graph Convolutional Networks |
| GPU | Graphics Processing Unit |

Continued on next page

Table 1 – *Continued from previous page*

| Abbreviations | Completed form |
|------------------------|---|
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| Guided Grad-CAM | Guided Gradient Class Activation Mapping |
| ICBM | International Consortium for Brain Mapping |
| ICC | IntraCranial Cavity |
| IXI | Information eXtraction from Image |
| LRP | Layer-Wise Relevance |
| MAE | Mean Absolute Error |
| MCDG | Multi-channel Deep Grading |
| MCI | Mild Cognitive Impairment |
| MIRIAD | Minimal Interval Resonance Imaging in Alzheimer’s Disease |
| ML | Machine learning |
| MLP | Multi-layer Perceptron |
| MNI[152] | Montreal Neurological Institute and Hospital [152 space] |
| MS | Multiple Sclerosis |
| MSE | Mean Square Error |
| NACC | National Alzheimer’s Coordinating Center |
| NDAR | National Database for Autism Research |
| NIFD | Frontotemporal lobar Degeneration Neuroimaging Initiative |
| OASIS | Open Access Series of Imaging Studies |
| OFSEP | Observatoire français de la sclérose en plaques |
| PD | Parkinson’s Disease |
| PNFA | Progressive Non Fluent Aphasia |
| PPMI | Parkinson’s Progression Markers Initiative |
| ROI | Region Of Interest |
| ResNet | Residual Network |
| SGD | Stochastic Gradient Decent |
| SRPBS | Strategic Research Program for Brain Sciences |
| SV | Semantic Variant |
| SVM | Support Vector Machine |
| SZ | Schizophrenia |
| T1w [MRI] | T1-weighted [Magnetic Resonance Imaging] |
| VGG | Visual Geometry Group |
| [s]MRI | [structural] Magnetic Resonance Imaging |
| bvFTD | behavioural variant of FrontoTemporal Dementia |
| pMCI | progressive Mild Cognitive Impairment |

Continued on next page

Table 1 – *Continued from previous page*

| Abbreviations | Completed form |
|---------------|----------------------------------|
| sMCI | stable Mild Cognitive Impairment |

Mathematical Notations

Mathematical notations and their signification

| Notations | Signification |
|--------------------------------|---|
| X | Input of a DL model (<i>e.g.</i> , an MRI image, a sub-volume of an image) |
| Y | Targeted grading map, same size as X |
| DG | Estimation of grading map, same size as X |
| m | Number of sub-volumes (patches) extracted from an MRI |
| V | Brain structure volumes (1D vector of size m) |
| A | Subject's age |
| x_i | i^{th} voxel of X |
| g_{ij} | Grading score of the voxel x_i in the local input patch X_j |
| G_i | Grading score of the voxel x_i in the global image |
| a_j | Balanced accuracy of a model of patch X_j on validation set |
| \mathcal{G} | Graph (with nodes and edges) |
| s | Number of considered brain structures |
| $\mathbb{N} = n_1, \dots, n_s$ | Set of nodes for the s brain structures |
| \mathbb{E} | Matrix $s \times s$ of edge connections |
| t | Random translated distance in voxels |
| α | Sample from the Beta distribution, used for Mixup data augmentation |
| X_{noise} | Noise sampled from normal distribution, used in test time augmentation |
| d_i | i^{th} data fold |
| dc | Targeted disease coordinate of a voxel |
| DC | Global targeted disease coordinate of a subject |
| C | Hyperparameter C in an SVM model |
| R^2 | Coefficient of determination |

Résumé en français

La détection précoce des maladies neurologiques est essentielle pour améliorer la qualité de vie des patients et réduire les coûts de santé. Aujourd'hui, l'Intelligence Artificielle (IA) joue un rôle central dans l'analyse des données médicales, en particulier dans le domaine de l'Imagerie par Résonance Magnétique (IRM), qui est largement employée pour diagnostiquer les maladies neurologiques. Toutefois, la production en grande quantité de ces images rend l'analyse manuelle quasiment impossible. En conséquence, de nombreuses méthodes basées sur l'Apprentissage Profond (AP), la dernière avancée de l'IA, ont été développées pour automatiser la détection des maladies neurologiques à partir des données IRM.

Cependant, malgré les avancées, les méthodes actuelles basées sur l'AP présentent plusieurs limitations. Leur performance est souvent inférieure aux approches traditionnelles d'apprentissage automatique, leur capacité de généralisation est un défi, et leur compréhension clinique est parfois insuffisante. De plus, les méthodes actuelles se concentrent généralement sur le diagnostic des maladies individuelles, limitant ainsi la compréhension des différences entre les maladies neurologiques.

Le but de cette thèse était de surmonter ces limitations en développant une nouvelle génération de méthodes de diagnostic des maladies neurologiques, en particulier les maladies neurodégénératives, démontrant de bonnes performances, une forte capacité de généralisation et une meilleure interprétation, tout en couvrant divers scénarios cliniques, y compris le diagnostic de maladies individuelles et de multiples maladies. Pour atteindre ces objectifs, un plan de recherche structuré consistant de 4 études a été conçu:

Deep Grading pour le diagnostic et pronostic de la maladie d'Alzheimer

Dans cette première étude, nous avons abordé le diagnostic et le pronostic de la maladie d'Alzheimer (AD).

Nous avons conçu une méthode de deux étapes. Dans la première étape, un grand ensemble de 125 U-Nets a été utilisé pour grader les images IRM structurelle. Ces U-Nets

ont permis de créer une carte 3D de grade indiquant la sévérité de la maladie au niveau des voxels, fournissant ainsi une localisation des régions cérébrales affectées par l'AD. Cette carte pourrait améliorer l'interprétabilité de notre modèle pour aider les cliniciens. La deuxième étape consistait à modéliser un graphe pour chaque individu en utilisant la carte de grade générée et d'autres informations pertinentes. Un classifieur de type Graph Convolutional Network a été utilisé pour effectuer la classification finale.

Les résultats ont montré que notre approche démontrait des performances comparables à celles des méthodes de l'état de l'art, tant pour le diagnostic que pour le pronostic de la maladie d'Alzheimer.

Multi-channel Deep grading pour le diagnostic différentiel de la maladie d'Alzheimer et de la Démence Frontotemporale

Dans la deuxième étude, notre objectif était d'étendre le biomarqueur Deep Grading pour le diagnostic différentiel entre les individus sains, la maladie d'Alzheimer et la démence frontotemporale (FTD).

Pour y parvenir, nous utilisons deux types de biomarqueurs : la grade et le volume des structures cérébrales. Tout d'abord, nous proposons d'entraîner un grand ensemble de 125 3D U-Nets pour déterminer localement les caractéristiques anatomiques des individus sains, des patients atteints de la maladie d'Alzheimer et ceux avec la démence frontotemporale à l'aide de l'IRM structurelle en entrée. La sortie de 125 U-Nets est une carte 3D à deux canaux, qui peut être transformée en une carte de grade qui est facilement interprétable pour les cliniciens. La carte à deux canaux est capable aussi de coupler à un Multi-Layer Perceptron pour différentes tâches de classification. Deuxièmement, nous proposons de combiner notre méthode avec une méthode d'apprentissage automatique traditionnelle (*i.e.*, Support Vector Machine) basée sur le volume pour améliorer la capacité discriminative et la robustesse du modèle.

Après validation croisée et validation externe, nos expériences, basées sur 3319 IRM, ont démontré que notre méthode produit des résultats compétitifs par rapport aux méthodes de l'état de l'art, à la fois pour la détection des maladies individuelles et le diagnostic différentiel.

Âge des structures cérébrales pour le diagnostic différentiel des maladies neurodégénératives

Dans la troisième étude, nous nous sommes intéressés à la capacité de l'âge apparent du cerveau pour décrire son état en fonction du vieillissement normal. L'objectif était d'estimer l'âge apparent des structures cérébrales en utilisant des données d'IRM

structurelle, ce qui pourrait aider à détecter des déviations par rapport à la trajectoire normative du vieillissement, offrant ainsi des informations sur les maladies neurodégénératives.

Pour atteindre cet objectif, un ensemble de 125 modèles U-Nets a été utilisé pour estimer une carte 3D de l'âge apparent du cerveau au niveau des voxels. Ce biomarqueur peut être utilisé dans plusieurs situations. Tout d'abord, il permet d'estimer avec précision l'âge chronologique des individus sains. Dans cette situation, notre approche surpasse plusieurs méthodes de l'état de l'art. Deuxièmement, les âges de la structure cérébrale peuvent être utilisés pour calculer la déviation par rapport à l'âge chronologique du sujet. Cette caractéristique peut être utilisée dans une tâche de classification multi-maladies pour un diagnostic différentiel précis. En effet, notre biomarqueur a démontré un bon complément avec les volumes des structures cérébrales pour le diagnostic différentiel entre les individus sains, les patients avec la maladie d'Alzheimer, ceux avec la Démence Frontotemporale, ceux avec la Sclérose en Plaque, ceux avec la maladie Parkinson et ceux avec la Schizophrénie. Enfin, les déviations de l'âge de la structure cérébrale des individus peuvent être visualisées, fournissant des indications sur les anomalies cérébrales et aidant les cliniciens dans des contextes médicaux réels.

3D Transformer pour le diagnostic différentiel de la maladie d'Alzheimer et de la Démence Frontotemporale

Dans la quatrième étude, nous avons exploré l'utilisation de modèles basés sur les transformers 3D pour adresser le diagnostic différentiel entre la maladie d'Alzheimer et la démence frontotemporale, deux affections qui partagent des symptômes cliniques similaires.

Tout d'abord, inspiré par les architectures de Vision Transformer, Swin et deformable attention, nous proposons une architecture Transformer légère pour assurer une affordable mémoire requise. Pour surmonter le défi de la rareté des données labélisées 3D en imagerie médicale, nous avons proposé une combinaison efficace de techniques d'augmentation de données pour l'entraînement des modèles basés sur les transformers. De plus, notre approche a été enrichie en combinant le modèle basé sur les transformers avec un modèle d'apprentissage automatique traditionnel (*i.e.*, Support Vector Machine) utilisant les volumes des structures cérébrales. Cette combinaison a permis d'exploiter au mieux les données disponibles. Les résultats de cette étude ont démontré l'efficacité de notre approche, offrant des performances comparables à celles des méthodes de l'état de l'art pour le diagnostic différentiel de l'AD et de la FTD. De plus, notre méthode permettait de visualiser la localisation de patch déformables, soulignant les régions cérébrales les plus pertinentes pour le diagnostic des deux maladies.

En résumé, cette thèse a apporté des contributions à l'amélioration du diagnostic des maladies neurodégénératives en utilisant des techniques avancées d'intelligence artificielle et d'IRM structurelle. Ces quatre études ont ouvert de nouvelles perspectives pour la communauté médicale en fournissant des outils plus performants, généralisables et compréhensibles pour la détection précoce des maladies cérébrales, tout en améliorant le diagnostic différentiel entre différentes affections neurologiques. Ces avancées pourraient avoir un impact important sur l'amélioration de la qualité de vie des patients et la réduction des coûts associés aux maladies neurologiques.

Les méthodes développées au cours de ces études ont été intégrées dans la plateforme VolBrain. VolBrain est un système en ligne d'imagerie volumétrique du cerveau par IRM dont la mission est d'aider les chercheurs du monde entier à obtenir automatiquement des informations volumétriques sur le cerveau à partir de leurs données d'IRM sans aucun effort.

Introduction

This thesis was conducted at LaBRI (Laboratoire Bordelais de Recherche en Informatique), specifically in the TAD (Traitement & Analyse de Données) team. Our TAD team carries out a wide range of research projects, from fundamental tasks such as filtering and reconstruction, to more complex content analysis tasks like object detection and knowledge extraction. The primary focus of this PhD research is to advance the field of neurodegenerative disease detection using brain imaging data. This work is an integral part of the ANR project DeepvolBrain, which aims to provide innovative solutions for various brain imaging analysis tasks, such as brain lesion segmentation, brain age prediction, and brain disease classification. The volBrain platform can be accessed at <https://volbrain.net/>.

The opening chapter of this manuscript serves two purposes: to provide an overview of the motivation behind this work and to introduce the field of brain disease diagnosis. Specifically, the chapter will present different neurodegenerative diseases, exploring their social and economic impacts, as well as their diagnostic strategies. Furthermore, the challenges associated with detecting these disorders will be discussed. Following this, a thorough outline of our objectives and contributions to the field will be presented. Finally, a structured outline of the manuscript will be presented, offering a roadmap to guide readers through the rest of the thesis.

Overview of neurodegenerative diseases

Neurodegenerative diseases and its impacts

Neurodegenerative diseases refer to a group of disorders characterized by the gradual loss of structure or function of neurons, a process known as neurodegeneration [67]. This damage to neurons can ultimately lead to cell death. Although the causes of neurodegenerative diseases are not yet fully understood, many works have shown that it is probably the result of a combination of various factors including genetic, environmental,

and lifestyle factors [100, 140, 165].

Neurodegenerative diseases can affect various brain functions such as movement [71], behavior [72], and cognition [22]. In the early stage of neurodegenerative disease, patients can have some minor difficulties in their basic daily activities (*e.g.*, walking, making attention, communicating) or changes in behavior (*e.g.*, apathy and agitation). Unfortunately, neurodegenerative disorders are irreversible and they get worse overtime [56]. At a late stage, individuals can experience significant neurological impairments such as complete loss of mobility and inability to recognize familiar people. Consequently, such diseases not only influence individuals' quality of life but they can also have a significant impact on their families.

Beside the impact of neurodegenerative diseases on individuals and their families, these diseases also result to a significant economic impact [252]. These diseases usually require a long process of treatment and care, leading to a considerable financial burden on healthcare systems [209]. In 2010, the costs for patients with dementia (*i.e.*, a category of neurodegenerative disease) were \$604 billion worldwide [252]. Those costs were increased to \$818 billion in 2015 [253], \$1.3 trillion in 2018 [255] and were expected to reach \$2.8 trillion by 2030 as both the number of patients with dementia and healthcare costs increase according to the World Health Organization [255].

Therefore, an accurate diagnosis of neurodegenerative diseases is crucial for reducing the social as well as economic impact. On the one hand, such diagnosis can lead to appropriate interventions which may slow down the progression of diseases and improve the quality of life for patients. On the other hand, this can help doctors/researchers better understand these diseases about its causes and associated treatment options.

Neuropsychological assessments and brain imaging

A traditional method for diagnosing and tracking the progression of neurodegenerative diseases is through neuropsychological assessments [103, 200, 86]. These assessments consist of a battery of tests and questionnaires that evaluates various cognitive abilities, including memory, language, attention, and executive function [146]. While these assessments appear to be valuable in gaining an understanding about patient's cognitive abilities, they have several limitations. One limitation is that they rely heavily on the specialized expertise of the examiner, which can introduce bias in the results and inter-rater inconsistency. Additionally, these assessments are often based on qualitative answers from the subject, which can vary depending on the patient's feelings at the time taking the questionnaire. Furthermore, these assessments may require multiple visits over sev-

eral months to make a decision¹, which can result in delays in receiving a diagnosis and accessing appropriate treatment. Lastly, these assessments may not be sensitive enough to distinguish between different types of neurodegenerative diseases, which can lead to misdiagnosis and inappropriate treatment [110, 258].

Brain imaging is also an important tool for diagnosing and tracking the progression of neurodegenerative diseases [259]. On one hand, brain imaging allows for the visualization of the brain tissue and can reveal changes in patients with neurodegenerative diseases. On the other hand, brain imaging can be more objective than neuropsychological assessments, as it is less dependent on the expertise of the examiner and the mental state of the subject making the test. Various brain imaging modalities are available, including structural Magnetic Resonance Imaging (sMRI), functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET), Computed Tomography (CT), Electroencephalography (EEG), Diffusion Tensor Imaging (DTI), and Single-Photon Emission Computed Tomography (SPECT). While various modalities may hold interest in different contexts, this PhD research specifically focuses on sMRI for several compelling reasons.

One of the primary advantages of sMRI is its ability to generate high-resolution images that can effectively differentiate between different types of brain tissue. This capability allows researchers and clinicians to carefully examine the size, shape and density of interesting brain structures such as the cerebral cortex and hippocampus. By comparing structural differences between individuals with neurodegenerative disorders and healthy subjects, valuable insights into these conditions can be gained.

Furthermore, sMRI is a non-invasive and safe imaging technique. Unlike modalities that utilize ionizing radiation, sMRI employs powerful magnets and radiofrequency pulses to capture detailed images without exposing patients to harmful radiation. This significantly reduces the associated risks, making sMRI preferable in clinical practice.

Additionally, sMRI is widely available in clinical and research settings, making it easily accessible for both diagnostic purposes and scientific investigations. This availability ensures efficient and timely diagnoses, as well as facilitating its integration into research studies, which further advances our understanding of the brain.

Finally, it is important to note that there are different contrasts of sMRI: T1-weighted (T1w) and T2-weighted (T2w) images. Figure 1 shows examples of T1w and T2w images. Each sMRI contrast highlights different properties of the brain. While T2w images can be useful when observing the fluid-filled spaces in the brain, T1w images are usually used to visualize the gray matter and white matter structures and appear to be more suitable for detecting neurodegenerative diseases. Thus, the T1w images have been chosen as the only type of image used during this PhD.

¹<https://www.nhs.uk/conditions/alzheimers-disease/diagnosis/>

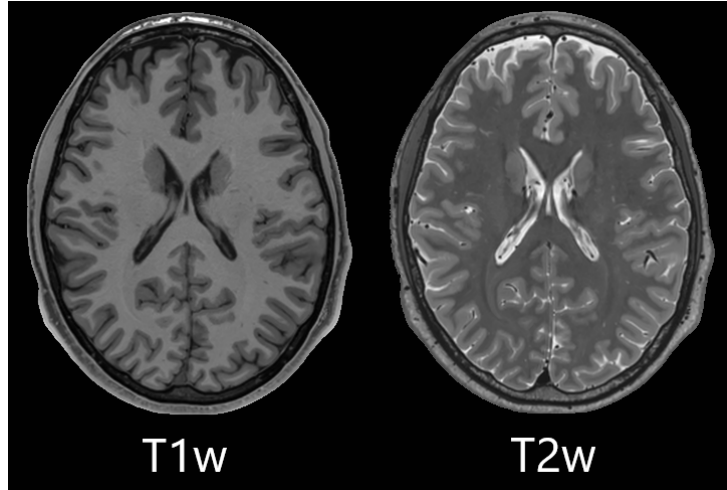


Figure 1: Examples of T1w and T2w images. The T1w image (left) highlights the gray matter and white matter structures, while the T2w image (right) highlights the fluid-filled spaces in the brain.

Common neurodegenerative diseases

Before describing the challenges of detecting neurodegenerative diseases, it is beneficial to establish a fundamental knowledge about these pathologies, such as their potential causes and their effects on the brain. Therefore, in this section, we present a concise introduction to several prevalent neurodegenerative diseases that served as the focal point for the methods developed throughout this PhD research: Alzheimer’s Disease, Frontotemporal Dementia, Multiple Sclerosis, Parkinson’s Disease and Schizophrenia. This overview may not be exhaustive, however, it can provide a basic understanding and hopefully help readers to better follow the rest of this manuscript.

Alzheimer’s Disease

Alzheimer’s Disease (AD) stands as the most prevalent form of neurodegenerative disease, categorized as a type of dementia which is characterized by memory loss and cognitive impairment. It is estimated that AD contributes to approximately 60-70% of dementia cases [100]. According to the Alzheimer’s Association, the majority of individuals with AD are 65 years old or older. The risk of AD doubles every 5 years after 65 years old and the risk reaches nearly one-third at the age of 85². In the early stages of the disease, patients may experience mild forgetfulness, personality changes, and increased anxiety. As the disease progresses, communication abilities decline, and patients become entirely reliant on others for their care, significantly impacting their daily lives and their families. In 2006, there were 26.6 million AD patients worldwide [35] which increased to

²<https://www.alz.org/alzheimers-dementia/what-is-alzheimers/causes-and-risk-factors>

46.8 million in 2015. This number is expected to reach 131.5 million in 2050 [96].

The causes of AD are multifaceted and not yet fully understood, but they are believed to involve a combination of genetic and environmental factors [154]. The apolipoprotein E (APOE) gene can play an important role in the development of AD. It manages the production and clearance of amyloid beta protein. Some APOE gene variants may lead to the accumulation of amyloid beta which damage brain cells [106]. Furthermore, environmental and lifestyle factors, such as bad diet, lack of exercise, and low level of education, have also been identified as potential risks for AD [161].

In Alzheimer’s disease, the typical brain structures presenting volume atrophy can include hippocampus, entorhinal cortex and parahippocampal cortex [115, 187, 198]. These changes can be observed on MRI scans. Figure 2 shows a visualization of a normal brain and a brain with AD using sMRI. For simplicity, only the morphological changes of hippocampus are highlighted. In this specific case, we can observe the atrophy of the hippocampus in a patient with AD.

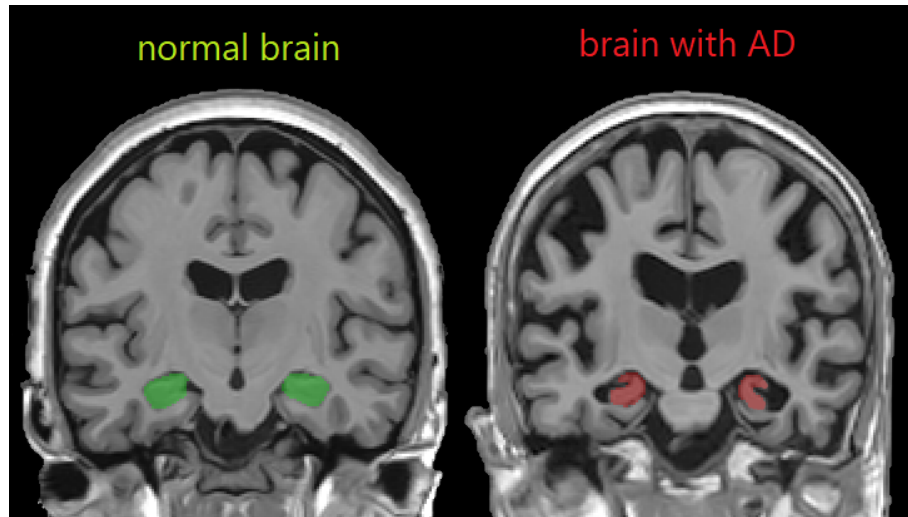


Figure 2: Visualization of a normal brain (left) and a brain with AD (right) using sMRI. Images taken from Alzheimer’s Disease Neuroimaging Initiative dataset. The regions of hippocampus are highlighted.

Frontotemporal Dementia

Frontotemporal dementia (FTD) is also a common cause of neurodegenerative disease after Alzheimer’s disease [75]. Compared to AD, FTD is less important, presenting less than 5% of dementia cases [145]. Patients with FTD also have problems of memory loss and cognitive impairment. FTD typically occurs earlier than AD, from 45 to 65 years old [122]. Patients with FTD can experience unusual or antisocial behavior as well as loss of speech or language. At the final stages of FTD, people cannot care for

themselves [224]. Although FTD and AD have some mentioned differences, they also share numerous common symptoms, which presents significant challenges in clinical diagnosis.

The analysis of FTD remains an area of ongoing research, and our understanding of its causes is still incomplete. However, as AD, FTD is believed to be influenced by a combination of genetic and environmental factors. For instance, mutations in the microtubule-associated protein tau gene can result in the buildup of abnormal tau protein in the brain, potentially disrupting the function of brain cells [87]. Additionally, certain environmental factors such as head injuries [207] and exposure to toxins [11] may be implicated in the development of FTD.

In Frontotemporal dementia, there may be a reduction in volume of structures in the frontal or temporal lobe [210, 249], which may also be observed using MRI scans. Figure 3 shows a visualization of a normal brain and a brain with FTD using sMRI. For simplicity, only the morphological changes of the superior frontal gyrus are highlighted. In this specific case, we can observe the atrophy of the frontal lobe in a patient with FTD.

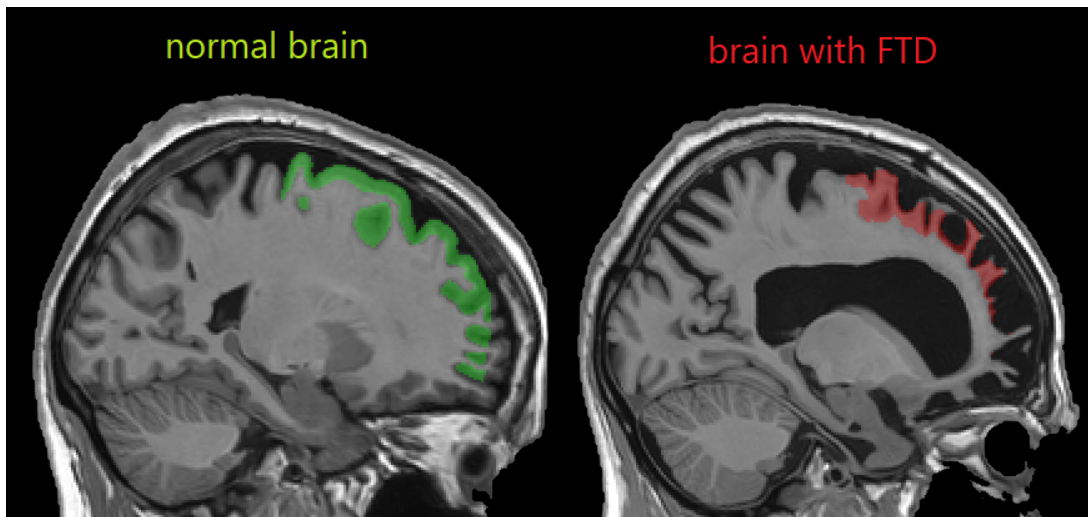


Figure 3: Visualization of a normal brain (left) and a brain with FTD (right) using sMRI. Images taken from the Frontotemporal Lobar Degeneration Neuroimaging Initiative dataset. The regions of superior frontal gyrus are highlighted.

Multiple Sclerosis

Multiple sclerosis (MS) is a demyelinating and chronic disease that attacks the central nervous system. In MS, the insulating covers of nerve cells in the brain and spinal cord are damaged [176]. This disrupts the ability of the nervous system to transmit signals [137]. MS can occur at any age, but typically affects individuals between 20 and 40 years old [212]. The symptoms of multiple sclerosis may depend on the location of

the damaged nerve fibers. In 2023, an estimated 2.8 million people live with MS worldwide [82].

The primary cause of Multiple Sclerosis (MS) is believed to be related to an autoimmune response, wherein the immune system mistakenly attacks the body’s own tissues. Specifically, this immune response damages the myelin sheath, which covers the nerve fibers in the brain and spinal cord [89]. Additionally, environmental factors such as viral infections, cigarette smoke and low levels of vitamin D may also related to the development of MS [190].

The impact on brain morphology in MS can include lesions in the white matter and gray matter, as well as brain structure atrophy such as thalamus, putamen, ventral diencephalon, and brainstem [55]. Figure 4 shows a visualization of a normal brain and a brain with MS using sMRI. For simplicity, only the morphological changes of the thalamus are highlighted. In this specific case, we can observe the atrophy of the thalamus in a patient with MS.

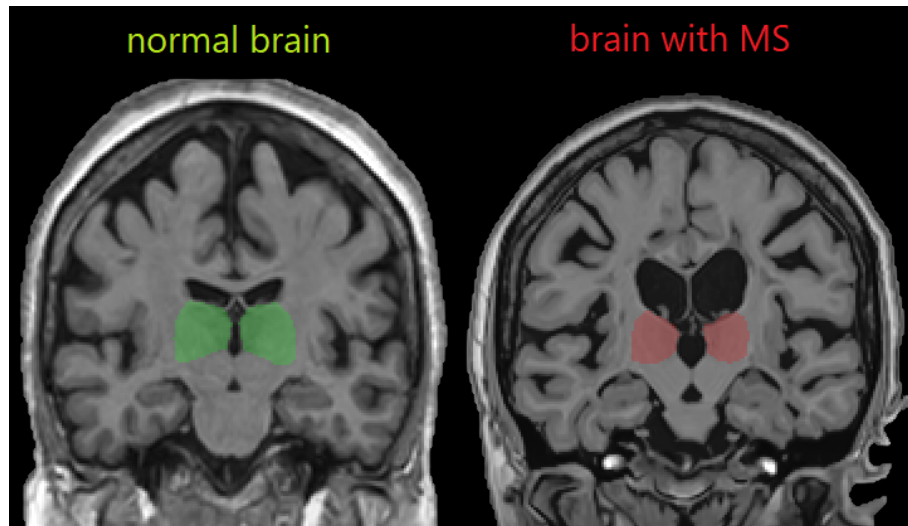


Figure 4: Visualization of a normal brain (left) and a brain with MS (right) using sMRI. Images respectively taken from the Alzheimer’s Disease Neuroimaging Initiative and the Observatoire Français de la Sclérose en Plaques dataset. The thalamus is highlighted.

Parkinson’s Disease

Parkinson’s disease (PD) is a neurodegenerative disorder that affects the nervous system and primarily parts that control the body movement. PD commonly affects individuals over the age of 50 [74]. Patients with PD usually experience symptoms including tremors, muscle stiffness and slowness of movement [70]. According to the World Health Organisation, an estimation of 8.5 million people currently living with PD, which is an

increase of 81% since 2000³.

Similar to the aforementioned diseases, the exact cause of PD is still unclear. However, genetic factor is thought to play the most important role in the development of PD. A lot of genes may be related to PD, however, inheritance patterns of these genes are complex and not fully understood [134].

The impact on brain morphology in PD can include changes in the cortical thickness, and also regions associated with movement [260] such as substantia nigra and striatum. Figure 5 shows a visualization of a normal brain and a brain with PD using sMRI. For simplicity, only the morphological changes of the caudate, a sub-region of striatum are highlighted. In this specific case, we can observe the atrophy of the striatum in a patient with PD.

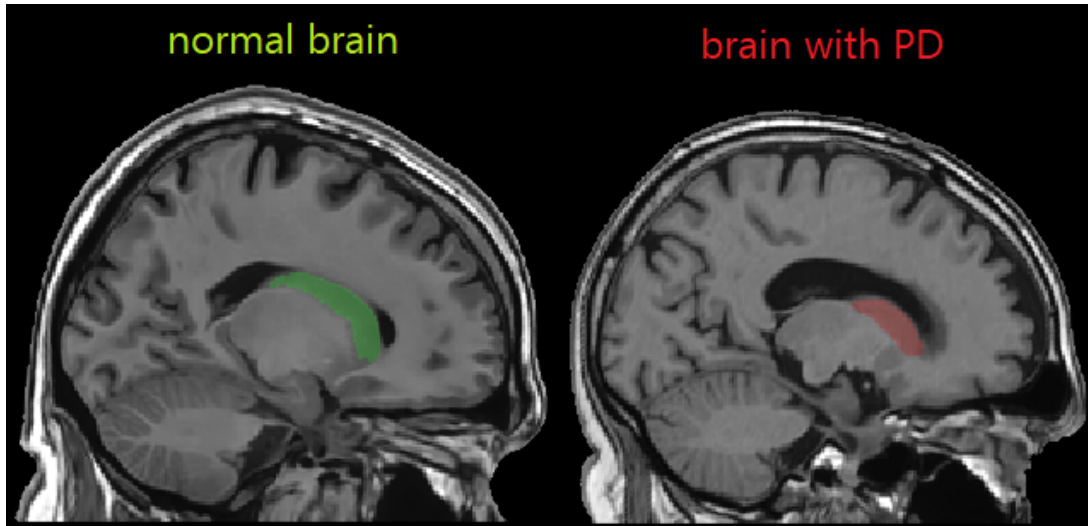


Figure 5: Visualization of a normal brain (left) and a brain with PD (right) using sMRI. Images taken from the Parkinson’s Progression Markers Initiative dataset. The caudate, a sub-region of striatum, is highlighted.

Schizophrenia

Schizophrenia (SZ) is a neurodevelopmental disorder with neurodegenerative processes that emerge in later stages. The initial symptoms of SZ are commonly observed in patients aged between 10 and 35 years old [204]. Individuals with SZ may experience abnormal interpretations of reality [182]. According to the World Health Organization, SZ symptoms can include a combination of delusions, hallucinations, passivity, disorganized thinking, and behavioral changes. It is estimated that approximately 24 million people were living with SZ in 2022⁴.

³<https://www.who.int/news-room/fact-sheets/detail/parkinson-disease>

⁴<https://www.who.int/news-room/fact-sheets/detail/schizophrenia>

The precise causes of SZ are not fully understood. However, it is believed that a combination of genetic factors and various environmental influences contribute to SZ [208]. For example, dopamine and glutamate neurotransmission are strongly implicated in SZ [178]. Several works have states that factors such as prenatal infections [17], maternal stress [130] can also be related to SZ.

Some of the most common brain morphological changes in schizophrenia include gray matter reductions in several regions of the brain, such as the frontal cortex, hippocampus, and temporal lobes [245]. Figure 6 shows a visualization of a normal brain and a brain with SZ using sMRI. For simplicity, only the morphological changes of the hippocampus are highlighted.

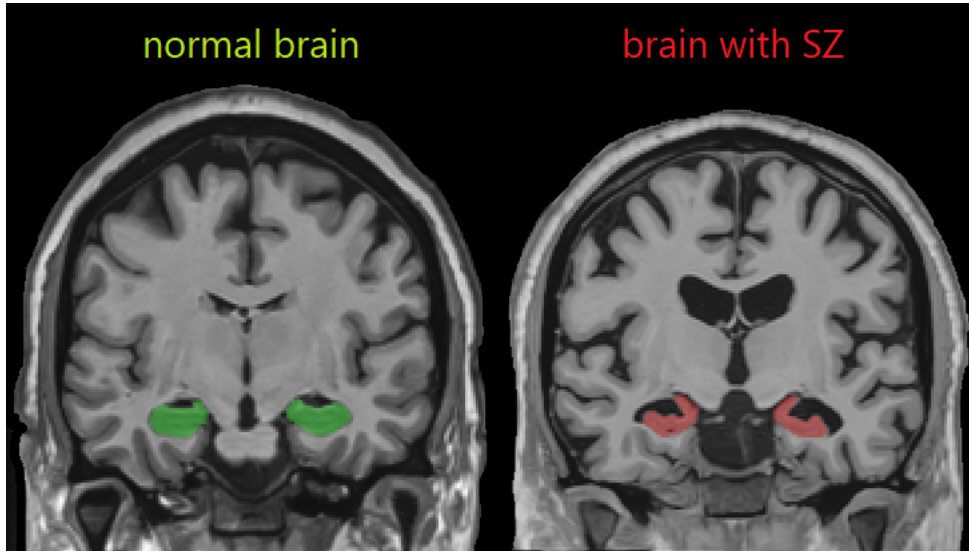


Figure 6: Visualization of a normal brain (left) and a brain with SZ (right) using sMRI. Images taken from the Center for Biomedical Research Excellence dataset. The regions of hippocampus are in the circles.

Scientific challenges

As previously mentioned, sMRI can serve as an effective tool for identifying neurodegenerative diseases. Furthermore, this type of brain imaging is widely used and accessible in most neurological testing centers. Consequently, the accurate detection of neurodegenerative diseases through sMRI is highly valuable in clinical settings. However, the large volume of sMRI data being generated, along with the 3D nature of the images containing millions of voxels, presents a challenge for manual analysis. This raises the need for automatic methods for analyzing sMRI data such as machine learning (ML) and deep learning (DL). Compared to ML, DL requires less feature engineering process and offers greater

capability for learning complex patterns from data. However, existing DL methods face several challenges:

Performance challenge

Over the last decade, several DL-based methods have been proposed with the aim of improving the accuracy and efficiency of diagnosis for various neurodegenerative diseases using sMRI data. However, despite these efforts, the performance of DL methods still exhibits certain limitations. For instance, when addressing the long-standing problem of AD diagnosis, DL approaches, particularly CNN-based methods, have not demonstrated superior performance compared to traditional ML methods such as support vector machine (SVM) [250]. Similarly, Bron *et al.* found comparable performance between CNN and SVM models when employing state-of-the-art CNN designs for the same application [34]. In the domain of FTD detection, Termine *et al.* observed that DL methods showed similar performance to SVM methods [228]. However, these authors also suggested that improvements in the framework design may yield better performance, indicating the potential for further advancements in DL-based methods for neurodegenerative disease diagnosis.

Generalization challenge

An important consideration in the field of neurodegenerative disease diagnosis is the limited knowledge about the generalization of existing methods, as they are often trained and evaluated on the same datasets. This practice limits our understanding of the generalization capacity of these models. For instance, in the case of AD diagnosis, approximately 90% of studies rely on a single dataset (*i.e.*, ADNI dataset, see more about this cohort in Section 1.1.1) for both training and evaluating their model [68]. Similarly, in the case of FTD diagnosis, each study tends to employ its own unique dataset for training and assessing the model performance [229, 58, 65, 33, 263]. This practice raises concerns about the applicability of these developed models in real-world scenarios where the data may differ from the training datasets in terms of MRI protocols, age ranges, country of origin, or inclusion criteria. Indeed, for AD classification, numerous studies have demonstrated that current deep learning methods perform well when applied to similar datasets but exhibit poor performance when faced with datasets that possess differences in these factors [34, 250].

Interpretability/explainability challenge

Considering the clinical perspective, the practical utility of DL models in clinical applications is also limited due to their lack of interpretability and explainability, making it difficult for clinicians to understand the reasons behind the model’s predictions. In this thesis, we rely on the definitions of [19] where interpretability refers to a model that can be directly understood by a human, while explainability involves external procedures to uncover its internal workings. Compared to explainability, interpretability seems to be more valuable as it provides clinicians with direct insights into the model’s functioning and predictions, without compromising important information through additional actions. However, only a few methods consider the model’s interpretability, preventing clinicians from trusting and comprehending DL models. In this PhD, both interpretability and explainability are considered in all developed methods, with a preference for interpretability to enhance the clinical utility of DL models.

Application scope challenge

Another important limitation of current DL methods in clinical applications is their limited focus on the diagnosis of single diseases, neglecting the discrimination of multiple diseases. This limitation becomes particularly significant when there are overlapping symptoms between different conditions, such as the need to distinguish between AD and FTD. However, only a few methods have been proposed to address this challenge [167, 104]. This lack of approaches that effectively handle multi-disease classification limits our knowledge about the differences between diseases and can hinder accurate diagnosis and appropriate treatment planning selection.

Contributions and outline

Based on the challenges discussed, the objective throughout this PhD is to develop a new generation of methods for disease diagnosis that demonstrates high performance, generalizability and interpretability/explainability. Furthermore, this research investigates various clinical scenarios, encompassing both single-disease and multi-disease diagnosis, with the aim of identifying specific abnormalities associated with individual diseases and revealing the differences between diseases in terms of their impact on the brain. Following the chapter 1, which provides an overview of the data and background knowledge related to this thesis, the four subsequent chapters will be presented to reach the stated objectives. These chapters are organized as follows:

- Chapter 2: The focus of the first study was to introduce a novel biomarker – called deep grading – with the mentioned objective for single-disease classification, with a specific emphasis on AD diagnosis. Additionally, we conducted an evaluation to assess the robustness of the framework by examining its generalization capacity on an unseen task (*i.e.*, AD prognosis).
- Chapter 3: Building upon the first study, the second study expanded the scope by extending our deep grading to multi-disease diagnosis. This study specifically targeted the differential diagnosis of CN *vs.* AD *vs.* FTD. To better understand these disease, we investigated whether the patterns exhibited by these diseases on brain structures are distinguishable or if they share similarities. Additionally, we examine how the different subtypes of FTD impact the brain.
- Chapter 4: Recognizing the need for interpretability when the number of diseases becomes higher, the third study aimed to overcome the issues of deep grading facing a large number of classes. To this end, we proposed a novel biomarker called brain structure ages. Specifically, the focus was on differentiating between several groups, which is CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ.
- Chapter 5: In our previous methods, we introduced two-stage frameworks that prioritized interpretability enhancements. However, in this particular study, our focus shifted slightly away from interpretability, as we explored the feasibility of an end-to-end (*i.e.*, one-stage) framework that could deliver comparable performance and generalizability. While our previous two-stage frameworks generally outperformed existing end-to-end methods, which predominantly relied on CNN architectures, recent advancements in DL indicated that Transformers could offer competitive or even superior results compared to CNN methods. Therefore, we conducted here an investigation into the potential of Transformers for multi-disease diagnosis. The differential diagnosis of CN *vs.* AD *vs.* FTD was chosen as the targeted application. Finally, the explainability of the proposed method was assessed and compared to our previous interpretable methods.

For each of these experimental studies, we further detail in Table 3 the associated observations, contributions and associated publications, offering a clear understanding of the research progression throughout this PhD.

Table 3: Overview of the different studies achieved during the PhD. Color indicators: ■ accuracy, ■ generalization capacity, ■ interpretability or explainability and ■ multi-disease handling.

Common objectives: providing methods with high accuracy, generalizability and interpretability (or explainability) for neurodegenerative disease diagnosis.

Chapter 2

CNN for single-disease diagnosis: CN *vs.* AD

Observations:

In AD diagnosis, a long-standing problem, current DL methods present several limitations:

- Performance: Limited accuracy compared to traditional ML methods (*e.g.*, SVM).
- Generalization: Limited knowledge about the generalization capacity of DL methods on external datasets.
- Interpretability: Few interpretable methods for understanding model decisions in AD classification.

Contributions:

- Deep Grading biomarker ■: Improves the interpretability of deep model outputs.
 - Collective Artificial Intelligence ■: Better generalizes on external data and unseen tasks.
 - Graph-based classification ■: Better captures AD signature to improve the classification performance.
-

Publications:

- Journal paper: [1].
 - Conference paper: [4].
 - Software: AssemblyNet-AD [7].
-







Continued on next page

Table 3 – *Continued from previous page*

| Chapter 3 CNN for multi-disease diagnosis: CN <i>vs.</i> AD <i>vs.</i> FTD | |
|--|---|
| Observations: <p>Only a few studies tackled this problem even though it presents valuable outcomes. The limitations of current methods are:</p> <ul style="list-style-type: none"> • Performance: Limited accuracy in many studies. This task is challenging due to overlapping symptoms of the 3 classes. • Generalization: No study conducted to assess the generalization on external datasets. • Interpretability: <ul style="list-style-type: none"> ◦ Few methods exists. ◦ Deep Grading biomarker is not well-suited for multi-disease diagnosis. | Contributions: <ul style="list-style-type: none"> • Multi-channel Deep Grading biomarker ■ ■: Extends the deep grading framework to produce 3D grading maps capable of detecting brain regions with specific disease-related patterns (AD-like or FTD-like patterns). • Combination of the structure grading and the structure atrophy ■ ■: Improves the model classification performance and its generalization capacity. <hr/> Publications: <ul style="list-style-type: none"> • Journal paper: [2]. • Conference paper: [5]. • Software: [8] (Under deposit). |
| Chapter 4 CNN for multi-disease classification: CN <i>vs.</i> AD <i>vs.</i> FTD <i>vs.</i> MS <i>vs.</i> PD <i>vs.</i> SZ | |
| Observations: <p>Multi-disease differential diagnosis with a high number of class is rarely studied.</p> <p>Multi-channel Deep Grading biomarker may solve this problem but the grading map is hard to visualize due to the high number of class, reducing the interpretable ability.</p> <p>Brain age biomarker shows promising results in disease diagnosis but seems to only be useful with binary classification problems.</p> | Contributions: <ul style="list-style-type: none"> • Brain Structure Ages biomarker ■ ■: Extends the notion of brain age to estimate the brain age at structural level. This biomarker can be used for: <ul style="list-style-type: none"> ◦ Age prediction: Accurately estimates of the chronological age of healthy people. ◦ Interpretation and Explanation: Provides quantitative visualization of abnormal brain regions showing the severity between diseases. • Combination with the structure atrophy ■ ■: Improves the performance and generalization capacity. <hr/> Publications: <ul style="list-style-type: none"> • Journal paper: [3] (Under revision). • Softwares: [9, 10] (Under deposit). |

Continued on next page

Table 3 – *Continued from previous page*

| Chapter 5 Transformer for multi-disease diagnosis: CN <i>vs.</i> AD <i>vs.</i> FTD | |
|--|---|
| Observations: <p>Transformer has recently appeared to be a promising alternative to CNN-based models in computer vision tasks. Despite their potential, its application in disease diagnosis is still sparse.</p> <p>In the domain of brain imaging analysis, there are two key challenges to address when utilizing transformers:</p> <ul style="list-style-type: none"> • High computational requirements, particularly when dealing with 3D sMRI data. • Limited availability of labeled data, typically consisting of a relatively small number of images per class (e.g., around 100-200 images). | Contributions: <ul style="list-style-type: none"> • Transformer-based model using Deformable Patch Location module    : <ul style="list-style-type: none"> ◦ Novel architecture: Improves model performance and generalization capacity. ◦ Interpretation and Explanation: Provides qualitative visualization of abnormal brain regions related to the model’s decision. • Novel data augmentation pipeline  : Provides a data augmentation pipeline for transformer models for disease diagnosis using sMRI data, allowing to boost the model performance and its generalization capacity. <hr/> Publications: <ul style="list-style-type: none"> • Conference paper: [6]. |

After the presentation of our methods, in Chapter 6, we summarize the main results and discuss about future research directions.

Lastly, we present supplementary materials A, B and C for Chapter 2, Chapter 3 and Chapter 4 in the appendices, respectively.

Chapter 1

Materials and methods

| | | |
|-------|--|----|
| 1.1 | Materials | 21 |
| 1.1.1 | Datasets used in this manuscript | 21 |
| 1.1.2 | Preprocessing steps for structural MRI | 23 |
| 1.2 | Common automatic approaches | 26 |
| 1.2.1 | Classical Machine Learning | 26 |
| 1.2.2 | Deep Learning | 27 |

Before going into the details of our proposed methods, this chapter provides an overview of the data and background related to this thesis. We begin by presenting general information about the datasets used in this PhD research, along with the image processing procedures employed. Subsequently, we introduce the fundamentals of classical machine learning and deep learning as part of the background.

1.1 Materials

In this section, we begin by providing an overview of the datasets used in this PhD research. To assess the generalizability of our proposed methods, we employ different datasets for each study, and the differences in image characteristics between these datasets are highlighted in this section. Following that, we introduce the image preprocessing procedure employed in our studies, which plays a crucial role in enhancing the data quality for improved model performance.

1.1.1 Datasets used in this manuscript

To facilitate the comparison with existing methods in the literature, we employ a set of publicly available datasets for the development and validation of our methods. Detailed information about these datasets is presented in Table 1.1.

It is important to note that all experiments in our studies are conducted using the T1w images acquired at the baseline. By focusing on the baseline T1w images, our purpose is to identify subtle structural changes in the brain and to develop biomarkers at the earliest possible stage.

It is also noteworthy that some of the datasets mentioned in Table 1.1 consist of multiple phases. For instance, ADNI encompasses ADNI1, ADNI2, ADNI3, and ADNIGO. However, for our studies, we specifically utilize ADNI1 (Chapters 2 and 4) and ADNI2 (Chapters 2, 3 and 5). Similarly, ABIDE and PPMI datasets have two distinct phases, while OFSEP comprises data from multiple data centers. Detailed information regarding each specific phase or subcategory is presented in the respective studies.

Additionally, in the ADNI1 dataset mentioned in Chapter 2, our focus extends beyond individuals classified as CN or AD. We also take a particular interest in patients with Mild Cognitive Impairment (MCI). This MCI category can be further divided into two groups based on the selection criteria from ¹:

¹<https://aramislab.paris.inria.fr/clinicadl/tuto/2020/html/Notebooks-AD-DL/clinical.html>

Table 1.1: Summary of public datasets used in this manuscript.

| Dataset Name | Male/Female | Age (mean \pm std) | Machine type | Labels | Related Chapters |
|---------------------|-------------|----------------------|--------------|--------------------|------------------|
| ABIDE [61] | 811/261 | 16.1 \pm 8.8 | 3T | CN | 4 |
| CamCAN [227] | 75/85 | 63.2 \pm 18.5 | 3T | CN | 4 |
| C-MIND ² | 107/129 | 8.4 \pm 4.3 | 3T | CN | 4 |
| ICBM [177] | 112/182 | 33.7 \pm 14.3 | 1.5T | CN | 4 |
| IXI ³ | 242/307 | 48.8 \pm 16.5 | 1.5T, 3T | CN | 4 |
| NDAR [193] | 208/174 | 12.4 \pm 6.0 | 1.5T, 3T | CN | 4 |
| UKBioBank [37] | 14917/14334 | 64.2 \pm 7.9 | 3T | CN | 4 |
| ADNI [115] | 628/491 | 74.9 \pm 7.0 | 1.5T, 3T | CN, AD, pMCI, sMCI | 2, 3, 4, 5 |
| AIBL [69] | 151/159 | 72.9 \pm 7.2 | 1.5T, 3T | CN, AD, pMCI, sMCI | 2, 4 |
| OASIS [139] | 360/466 | 70.0 \pm 9.2 | 1.5T | CN, AD | 2, 4 |
| MIRIAD [169] | 31/38 | 69.4 \pm 7.0 | 1.5T | CN, AD | 2, 4 |
| NIFD ⁴ | 151/135 | 63.7 \pm 7.2 | 3T | CN, FTD | 3, 4, 5 |
| NACC [24] | 939/1765 | 68.9 \pm 10.8 | 3T | CN, AD, FTD | 3, 4, 5 |
| OFSEP [246] | 746/1936 | 42.7 \pm 11.6 | 1.5T, 3T | MS | 4 |
| PPMI [113] | 337/202 | 62.2 \pm 9.9 | 1.5T, 3T | CN, PD | 4 |
| COBRE ⁵ | 65/21 | 39.2 \pm 13.3 | 3T | CN, SZ | 4 |
| SRPBS [226] | 172/118 | 39.4 \pm 13.1 | 3T | CN, SZ | 4 |
| BrainGluSchi [36] | 132/36 | 38.0 \pm 12.6 | 3T | CN, SZ | 4 |

- Progressive MCI (pMCI): patients who was diagnosed as MCI at baseline, and progressed to dementia in the following 36 months.
- Stable MCI (sMCI): patients who was diagnosed as MCI at baseline, and the diagnosis remains unchanged in the following 36 months.

Lastly, in the NIFD dataset, patients with FTD can be further classified into three subtypes (more details are provided in Chapter 3):

- Behavioral variant of frontotemporal dementia (bvFTD): Patients presenting changes in behavior and personality
- Progressive Non Fluent Aphasia variant (PNFA): Patients with difficulties in producing or comprehend language.
- Semantic Variant (SV): Patients with difficulties in recognizing objects and faces, and a loss of vocabulary

1.1.2 Preprocessing steps for structural MRI

In the medical imaging domain, a preprocessing procedure is necessary to improve the robustness of the classification methods by reducing irrelevant information and minimizing differences in orientation, size, and resolution between images. Moreover, it could increase the signal-to-noise ratio and improve the contrast between different tissue types, enhancing the quality of the data used for classification. The used preprocessing procedure consisted of 6 steps: denoising, inhomogeneity correction, affine registration into MNI152 space, intensity standardization, intracranial cavity (ICC) extraction and optional segmentation. Figure 1.1 shows an overview of the preprocessing pipeline. In the following, we provide a brief overview of each of these steps.

Denoising

Structure MRI is a 3D image where each voxel has an intensity value obtained from a MRI machine. However, during the acquisition process, random noise are generated, reducing the quality of image. Several methods have been proposed for alleviating the effect of random noise [173, 53]. During this PhD, the denoising technique used refers to [172]. This technique is based on non-local mean (NLM) filter method which was initially proposed for denoising natural images. In principle, NLM filter updates the intensity of each voxel with a weighted average of all other voxels in the image. The weights are estimated using the similarity between local neighborhood of the considered voxel and other voxels.

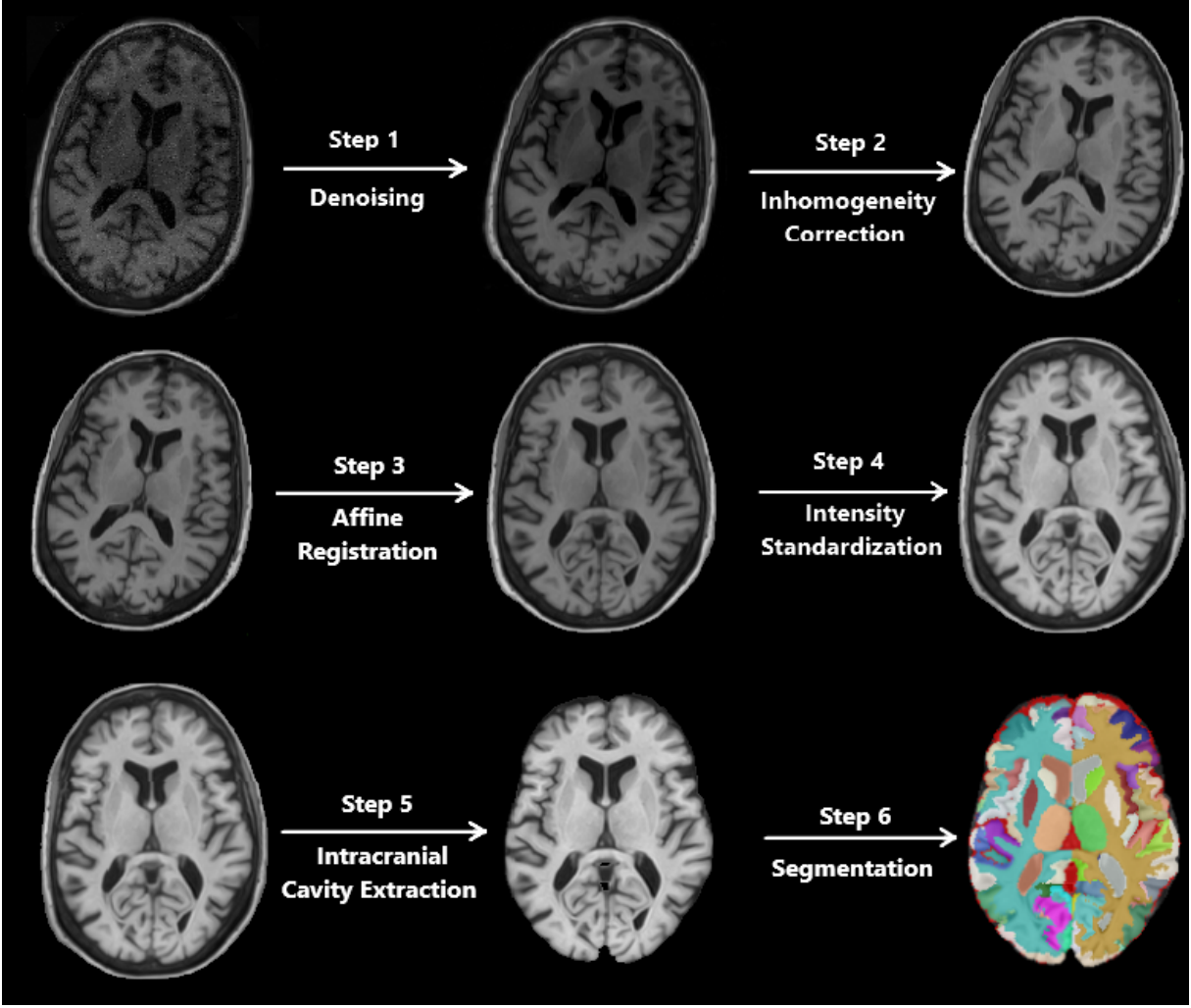


Figure 1.1: Overview of the preprocessing pipeline utilized in this PhD. The image of the step 2 is adapted from [84] for the purpose of better visualization. All images of other steps are taken from our data.

Inhomogeneity correction

Inhomogeneity correction is a preprocessing step aiming to correct for intensity non-uniformity (bias field). This phenomenon can be caused by various factors, such as magnetic field inhomogeneity in MRI [225] or posture of the subject [73]. The method used for this preprocessing step is the N4ITK algorithm [239]. This algorithm uses a series of low-pass filtering, B-spline fitting and an iterative estimation procedure to correct the bias field in the MRI.

Affine registration into MNI152 space

Affine registration is a preprocessing step used to correct for global differences between brain MRI images by applying a series of transformations including translation,

rotation, scaling and shearing⁶. The MNI152 space is a standard template constructed in 2001 at the Montreal Neurological Institute using brain MRI images from 152 normal subjects [170]. This space represents a brain that is considered representative of the population. The main purpose of this preprocessing step is to transform individual brain images to a standardized and common coordinate space, allowing for better comparison between different brains and analysis across multiple studies. Moreover, it may reduce the bias between different datasets. The method used for this preprocessing step refers to [16].

Intensity standardization

Due to variations in scanner configurations, the obtained sMRI may not consistently exhibit the same intensity levels. Furthermore, even when employing the same scanner and settings, intensity patterns in the captured images may display variability across different sessions. A correction of this intensity variation is important, particularly for DL methods where the different data distribution can result in unexpected model behavior. This preprocessing step refers to [171].

Intracranial cavity extraction

Intracranial cavity extraction or skull stripping is a preprocessing step in MRI that removes non-brain tissue (*e.g.*, skull, scalp, and meninges) from the MRI images, leaving only the brain tissue. This step is necessary as it can remove potential sources of noise and error that could interfere further analysis. The method used for this preprocessing step refers to [174].

Segmentation

Segmentation, a well-established domain of research with a large community, focuses on identifying the locations of brain structures at the voxel level. While segmentation is a challenging task, we simplify the discussion to maintain focus. In our study, segmentation serves as an optional preprocessing step, employed when specific information about the location of brain structures is required, such as calculating the volumes of different brain regions. For this step, we use the AssemblyNet method [50]. By considering the segmentation as a preprocessing step, we aim to direct the readers' attention towards our proposed methods and their contributions, while acknowledging the importance of segmentation in the medical imaging research landscape.

⁶<https://neuroimaging-data-science.org/content/006-image/003-registration.html>

1.2 Common automatic approaches

In this section, we provide a brief overview of the most common automatic approaches for neurodegenerative disease diagnosis. We first introduce the classical machine learning (ML) methods, followed by deep learning (DL) methods with their basic components such as multi-layer perceptron (MLP), convolutional neural network (CNN), and transformer.

1.2.1 Classical Machine Learning

ML techniques have been used in medical imaging domain for a long time, enabling the development of various approaches for tasks such as classification, regression, and segmentation. ML-based methods typically consist of two steps: feature extraction and mapping of the extracted features to the desired output.

In the feature extraction step, we simplify the learning process by utilizing input features that are more relevant to the task, rather than using raw voxel intensities from MRI scans. These extracted features, often referred to hand-crafted features, are manually selected by domain experts based on their knowledge and understanding of the underlying data. For neurodegenerative disease diagnosis, some important features can be brain structure volumes and the cortical thickness as these disorders often cause brain atrophy. In the context of neurodegenerative disease diagnosis, critical features may include brain structure volumes and cortical thickness, as these disorders often manifest as brain atrophy [221, 196]. Overall, the goal of this step is to transform the raw data into a set of meaningful features that capture essential information for a specific task.

Once the relevant features are extracted, a chosen ML algorithm proceeds to the second step, which involves correlating these features with the desired output. This mapping is done through a training process. Many ML algorithms have been used for neurodegenerative disease diagnosis, such as the bayesian approach [218, 216], support vector machine [13, 162, 214, 85, 264], logistic regression [205], linear discriminant analysis [265, 102] and K-means clustering [242].

In the context of medical imaging, certain ML methods can be easy to interpret. However, it is important to acknowledge that ML methods also have certain limitations. Indeed, ML methods often rely on domain expertise to select relevant input features. This may prevent the discovery of novel knowledge and hinder the ability to effectively handle new disorders or conditions where domain expertise is limited.

1.2.2 Deep Learning

Deep Learning is a specific subset of ML that is based on deep neural networks. DL models employ multiple layers of processing to progressively extract higher-level features from input data. In recent years, DL methods have gained significant popularity, mainly due to advancements of computing capacity and the availability of large-scale training datasets. As a data-driven approach, the effectiveness of DL models heavily relies on the quantity and quality of the training data. With more representative and diverse data, DL models have the potential to deliver improved performance and generalization capabilities.

In general, a deep neural network is composed of two main components: a feature extractor and a classification (or regression) head. The feature extractor plays a crucial role in capturing meaningful representations from the input data, while the classification head is responsible for making predictions based on these extracted features.

Two of the most commonly used feature extractors in deep neural networks are the convolutional neural network (CNN) and the transformer. The head of a deep neural network in most cases is a multi-layer perceptron (MLP). In the following, we will describe more about how these components work.

Multi-Layer Perceptron: MLPs are the most basic type of neural networks. They can directly take the hand-crafted features (similar to ML approaches) or the output representations of the feature extractor to perform classification. They consist of multiple hidden layers, each comprising a set of neurons with weighted connections. The neurons in each layer receive inputs from the previous layer and apply an activation function (*i.e.*, introduce non-linearity) to produce an output. Figure 1.2 shows an illustration of an MLP.

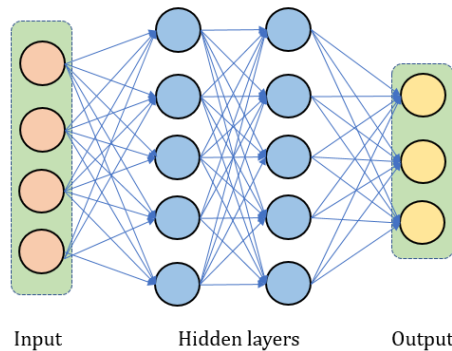


Figure 1.2: Illustration of MLP. Only the neurons (circle) is shown for simplicity.

Convolutional Neural Networks: CNNs are a type of deep neural networks that have been widely used in computer vision tasks. The core of CNNs is convolutional layers that convolve learnable filters or parameters with input data, allowing for the extraction

of local patterns and features (see Figure 1.3). In the initial convolutional layers, the model focuses on capturing low-level patterns, including edges, textures, and shapes. As the CNN progresses deeper into subsequent layers, it gradually learns more abstract and high-level representations of the input data. This hierarchical learning process enables the network to capture increasingly complex and meaningful information from the input.

Transformers: Transformers are built upon the concept of self-attention, initially designed for sequential data, but have recently shown impressive results in natural image applications. In image processing, the input is divided into patches (see Figure 1.3). Transformers treat these patches as a sequential input and utilize multiple attention layers for feature extraction. Within each layer, all patches attend to each other to progressively update the imaging features. By attending to all patches simultaneously, transformers is capable of capturing long-range dependencies and modeling complex relationships between different regions of the image.

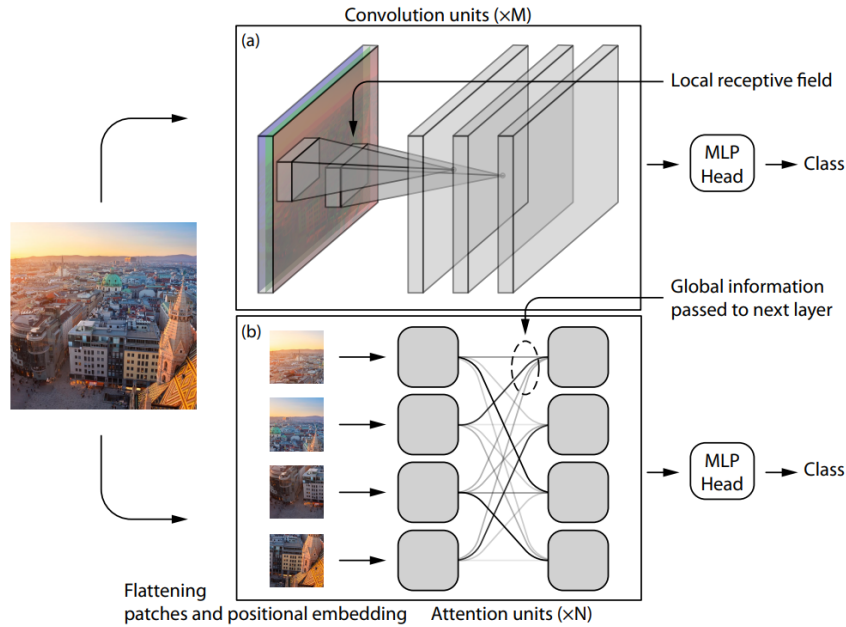


Figure 1.3: Illustration of CNN and Transformer (image from [238])

In contrast to classical ML methods, DL minimizes or removes the requirement for manual feature engineering, as both feature extraction and correlation are learned concurrently during training. However, a potential drawback of DL methods is their reduced interpretability, which can hinder understanding the decision-making process. This challenge of interpretability poses particular difficulties in the integration of AI methods into clinical practice.

Chapter 2

Deep grading for single-disease diagnosis

| | | |
|-------|--|----|
| 2.1 | Introduction | 30 |
| 2.1.1 | Context | 30 |
| 2.1.2 | Related works | 31 |
| 2.1.3 | Current limitations of DL in AD classification | 33 |
| 2.1.4 | Contributions | 35 |
| 2.2 | Materials | 36 |
| 2.3 | Methods | 36 |
| 2.3.1 | Method overview | 36 |
| 2.3.2 | Deep grading | 38 |
| 2.3.3 | Collective AI | 39 |
| 2.3.4 | Feature classification | 40 |
| 2.3.5 | Implementation details | 40 |
| 2.4 | Experimental results | 41 |
| 2.4.1 | Performance study | 41 |
| 2.4.2 | Interpretation of deep grading maps | 49 |
| 2.4.3 | Consistency study | 50 |
| 2.5 | Discussion | 52 |

The method presenting in this chapter is related to the following publications and softwares:

- [4] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Deep Grading Based on Collective Artificial Intelligence for AD Diagnosis and Prognosis”. In: *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. IMIMIC 2021*. 2021. DOI: [10.1007/978-3-030-87444-5_3](https://doi.org/10.1007/978-3-030-87444-5_3).
 - [1] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Towards better interpretable and generalizable AD detection using collective artificial intelligence”. In: *Computerized Medical Imaging and Graphics* 104 (2023), p. 102171. DOI: [10.1016/j.compmedimag.2022.102171](https://doi.org/10.1016/j.compmedimag.2022.102171).
 - [7] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *AssemblyNet-AD Automatic Diagnosis of Dementia*. IDN.FR.001.190026.000.S.C.2022.000.31230.
-

In this chapter, we will present a new biomarker for single-disease diagnosis problems. This biomarker is designed to be interpretable and possesses a high discriminative capacity for the purpose of neurodegenerative disease classification. Alzheimer’s disease has been chosen for analysis, as it is the most prevalent neurodegenerative disorder and has a large research community, offering valuable insights and opportunities for a good benchmark against existing literature.

2.1 Introduction

2.1.1 Context

As previously discussed, AD is the most common form of dementia in individuals over the age of 65. In addition to the impact on patients’ quality of life, the costs associated for carrying individuals with AD have been rapidly increasing over the past decade, as more treatments and services are required over time. Therefore, early and accurate detection

of AD is crucial for the development of new therapies, slowing disease progression, and reducing associated costs.

The prodromal stage of AD is Mild Cognitive Impairment or MCI [175]. People may experience minor changes in cognitive abilities at this stage but there is still no impact on their daily lives [14]. Statistically, 10 – 17% of people with MCI will progress to AD over a few years while other MCI patients will remain stable [92]. The first group refers to pMCI and the second one refers to sMCI (see Section 1.1.1 for more details). Besides the need of distinguishing AD patients from CN (*i.e.*, AD diagnosis), identifying pMCI patients from sMCI patients (*i.e.*, AD prognosis) is even more crucial to apply appropriate therapies and slow down the transition from MCI to AD. Therefore, a fast and accurate tool for both AD diagnosis and prognosis is expected to help clinician to take care of the patient as soon as possible.

Brain atrophy is an important biomarker of AD. Many studies state that this morphological change may occur before the first cognitive symptoms of AD [88, 126, 49, 34]. Those anatomical changes can be identified with the help of sMRI [32]. Recently, with the advances of DL, a large number of methods have been proposed for automatic AD diagnosis and prognosis using sMRI [250, 220, 68]. However, the large size of 3D sMRI and the limited GPU memory has required to adapt DL methods to medical imaging. These methods can be categorized into: 2D slice-based methods, 3D subject-based methods, 3D region-of-interest (ROI) methods and 3D patch-based methods.

2.1.2 Related works

2D slice-based methods

The concept behind slice-based approaches is that a minimal number of 2D slices may accurately depict the disease status. Some methods employ their own strategy to extract the most appropriate 2D slices from 3D sMRI, while others use standard image projections (*i.e.*, coronal, sagittal, axial plane) [68]. Valliani *et al.* considered only the median axial slice of sMRI and used ResNet for AD diagnosis [241]. Pan *et al.* trained 123 classifiers for 123 2D slice positions of three projection planes for both AD diagnosis and prognosis [192]. The 15 models with the highest accuracy on validation set were chosen to form the final ensemble model. Qiu *et al.* manually chose three slice positions to analyze well-known regions associated with Alzheimer’s disease: lateral ventricles, inferior temporal, and middle temporal cortices [201]. Three CNN models (one per region) were then trained for the problem of classification CN vs. MCI. The final result was based on the majority vote. Entropy-based sorting is another method to select slice positions. It is based on the hypothesis that a slice with higher intensity variation is more informative.

This strategy was used in [101, 117] to select the most 32 informative slices for various AD classification tasks. All of these slice-based methods have the advantage of being based on well-known CNNs architectures dedicated to natural image classification. However, a comparative study showed that 2D slice-based methods were less efficient than 3D methods [250]. This study explained that spatial information is not fully exploited by 2D slice-based methods which limits their performance.

3D subject-based methods

Recently, more methods using the whole 3D MRI have been proposed for AD classification (3D subject-based methods). In general, these models have fewer layers than 2D slice-based approaches due to the limited computing capacity. Backström *et al.* used a 3D CNN with 8 layers for AD diagnosis [18]. Yee *et al.* used dilated convolution to increase the model depth to 11 layers [257]. In doing so, they improved the receptive field while keeping a reasonable number of parameters. VGG and ResNet are usually employed by many authors for classification tasks in natural images. In [135], the authors implemented 3D version of these two architectures and showed comparable performance of both models for different AD classification tasks. Modern architecture like inception module was also proposed in [191]. Li *et al.* proposed a multi-model for AD diagnosis [148]. As each model had a different receptive field, the ensemble model was expected to be able to capture both global and local features. Overall, 3D subject-based methods have the advantage of preserving spatial information. However, since the 3D architectures are shallower, with current memory limitations these models do not yet offer optimal performance.

3D regions-of-interest (ROI) methods

With a limited computing capacity, reducing the input dimension is a good way to increase the model complexity. Many methods focused on particular parts of the brain known to be related to AD. Only one or a few small 3D cubic sub-volumes located at specific brain structures are used as input. Consequently, deeper models can be used and more complex patterns can be captured. The hippocampal region is a ROI well-known to be affected by AD [215]. Huang *et al.* cropped a region centered at the hippocampi from sMRI. They used a VGG-like architecture for classification [107]. Cui *et al.* used two cubic sub-volumes surrounding the left and right hippocampus to exploit also their adjacent regions for accurate AD classification [57]. They suggested that these areas, including the parahippocampus and amygdala, may be involved in AD. The main drawback of this type of method is that they only use the information around a priori defined anatomical regions. In contrast, alterations caused by AD can affect other brain areas [247]. Therefore, relevant information outside of the selected ROIs is not used, limiting model performance.

3D patch-based methods

Another way to reduce the input dimension is to use 3D patch-based methods. An MRI is simply divided into multiple smaller patches, all of them are then used for training. Cheng *et al.* extracted 27 overlapping patches that were uniformly distributed across the whole brain. They then trained 27 models (one per patch) and an ensemble model aggregating patch-level results to make the final decision [40]. Li *et al.* divided the original MRI into 27 non-overlapping patches [149]. These patches were grouped into different clusters and one CNN was trained per cluster for the AD diagnosis problem. The final decision was made by ensembling these models. In several studies, Liu *et al.* used a landmark detection algorithm to locate the most informative patches in sMRI [158, 157, 156]. In [158], the authors trained 27 different models (one per patch) for the classification problem. The final decision was obtained by majority voting strategy. In [157], they designed an end-to-end CNN model with multiple branches, each one analyzing one patch. The learned features were concatenated and forwarded through a final CNN for AD classification. In [156], they constructed multi-channels input from extracted patches and used a simple CNN for AD classification. Lian *et al.* performed a voxelwise anatomical correspondence across all available images [152]. They then selected 120 voxel locations and used them as centers for extracting 120 patches. They built a single end-to-end CNN model in which feature representations learned from patch-level was concatenated at regional-level, feature representation at regional-level was then concatenated to provide the decision at subject-level. From a literature review, it appears that a single model is not enough to capture the diverse patterns of all patch locations [250]. Indeed, methods using multiple models [40, 149, 158] offer better AD classification accuracy. Compared to previously detailed strategies, 3D patch-based methods enable to fully exploit the 3D information, to drastically reduce memory requirement and to analyze the entire MRI.

2.1.3 Current limitations of DL in AD classification

Although many efforts were made to adapt deep learning methods to AD classification, existing methods still present several limitations. Indeed, current approaches have limited prognosis performance and usually suffer from a lack of generalization and interpretability.

Limited Performance

At the time of writing this manuscript, CNN based-models seemed not to perform better than traditional machine learning methods (*e.g.*, SVM). Bron *et al.* showed similar

performance between CNN and SVM models while carefully following the state-of-the-art CNN designs [34]. In another study, Wen *et al.* [250] even found that their linear SVM model was at least as good as the best CNN model for AD diagnosis and better for AD prognosis. However, they both suggested that a more sophisticated DL architecture may help for better performance.

Limited Generalization

A recent survey showed that about 90% of studies use the same dataset (*i.e.*, ADNI dataset) to evaluate their model performance which limits our knowledge of CNN performance on other databases [68]. Moreover, most of the studies mentioned above used the same dataset for training and testing. Such validation framework is known to overestimate method performance. Indeed, in-domain validation is dangerous as methods showing high performance on a single dataset might just better capture the particular characteristics of that dataset and might poorly perform in another dataset [231]. As a consequence, current DL literature offers limited knowledge about the generalization capability of DL methods on external datasets. This limitation does not only apply to AD classification application but also to other diseases (*e.g.*, Frontotemporal Dementia [229], Parkinson’s disease [181], etc.). A general cause leading to a low generalization capacity is overfitting on the training set [250]. Especially, this often occurs when the size of the training domain is too small. To alleviate this problem, we applied several data augmentation techniques during the training process (see Section 2.3.5), making the model more robust to heterogeneity. Furthermore, our use of a large number of models (see Section 2.3.3) that can be seen as an ensemble model allows the reduction of generalization error [136].

Limited Interpretability and Explainability

Besides the need for an accurate and generalizable AD classification model, understanding the model decision is also vital. Here, we consider two terminologies: interpretability and explainability as in [19]. Interpretability refers to the passive characteristic of a model that can be directly understood by humans. By contrast, explainability refers to external procedures applied to a model to discover its internal functionalities. The majority of current deep learning methods use an external explainable method including Class Activation Mapping (CAM), Gradient Class Activation Mapping (Grad-CAM), Guided Backpropagation (GB) to study their model decision. However, some explainable methods (*i.e.*, Guided Backpropagation and Guided Gradient Class Activation Mapping) produce visually and quantitatively similar explanations between a model randomly-initialized and a trained model. This makes analysis based on the produced explanations suspicious [12].

In [257], two explainable methods were applied to the same model but different results were obtained. Moreover, in [34], the obtained saliency map showed regions known to be little affected by AD. Indeed, each explainable method works differently, so the discovery may not be unique or little informative. For an interpretable model instead, humans can directly infer its characteristic without losing information due to additional actions. Thus, this kind of method seems to be more valuable to understand the model decision. However, to the best of our knowledge, there is currently few interpretable methods for AD classification, with our definition [21, 20].

2.1.4 Contributions

To address these current major limitations of DL methods, we propose a novel interpretable, generalizable and accurate deep framework for both AD diagnosis and prognosis. This clinical tool is available at <https://volbrain.net>.

First, we propose a novel Deep Grading (DG) biomarker to improve the interpretability of deep model outputs. Inspired by the patch-based grading frameworks [49, 51, 234, 99, 48, 97], this new biomarker can capture CN, AD patterns from MRI input and provides a grading map with a score between -1 and 1 at each voxel that reflects the disease severity. This interpretable biomarker may assist clinicians in localizing brain regions affected by AD, allowing them to make more informed decisions.

Second, we propose to extend the concept of Collective Artificial Intelligence (AI) to AD diagnosis and prognosis. The collective AI consists of using a large number of communicating neural networks, each of them is specializing in a unique brain location. The global result is then obtained by fusing the local results. For the brain segmentation application, it has demonstrated a better generalization capacity against domain shift [50, 127]. In this study, we propose a robust fusion strategy in the generation of the global deep grading map using validation accuracy. Our experiments show an improvement of model performance using this strategy. Moreover, this could also help to emphasize the brain locations related to AD, making the global deep grading map more reliable.

Finally, we propose to use graph-based modeling to better capture AD signature. Concretely, we propose to use graph convolutional network (GCN) model for AD classification problems. As a result, this shows state-of-the-art in performance for both AD diagnosis and prognosis.

Table 2.1: Summary of participants used in our study. Data used for training are in bold

| Dataset | Statistic | CN | AD | sMCI | pMCI |
|---------|----------------------|----------------------------------|----------------------------------|----------------|----------------|
| ADNI1 | No. subjects | 170 | 170 | 129 | 171 |
| | Age (Mean \pm Std) | 75.9 ± 5.2 | 75.1 ± 7.2 | 74.6 ± 7.5 | 74.5 ± 7.0 |
| ADNI2 | No. subjects | 149 | 181 | | |
| | Age (Mean \pm Std) | 74.1 ± 6.6 | 74.0 ± 7.2 | | |
| AIBL | No. subjects | 232 | 47 | 12 | 30 |
| | Age (Mean \pm Std) | 72.3 ± 6.7 | 72.7 ± 8.6 | 72.5 ± 6.2 | 73.9 ± 8.0 |
| OASIS | No. subjects | 658 | 98 | | |
| | Age (Mean \pm Std) | 68.6 ± 8.9 | 76.8 ± 8.4 | | |
| MIRIAD | No. subjects | 23 | 46 | | |
| | Age (Mean \pm Std) | 69.6 ± 7.0 | 69.3 ± 7.0 | | |

2.2 Materials

The data used in this study, consisting of 2106 subjects, were obtained from multiple cohorts: ADNI, OASIS, AIBL and MIRIAD. We used the baseline T1-weighted MRI available in each of these studies. Each dataset contains AD patients and CN subjects. ADNI1 and AIBL datasets also include pMCI and sMCI patients. Further details about these datasets and the selection criteria of pMCI and sMCI refers to Section 1.1.1. Table 2.1 summarizes the number of participants and their age distribution for each dataset used in this study. During our experiments, AD and CN subjects from ADNI1 were used for training and all the other subjects as testing set. To minimize possible bias learned through training, we selected the same number of AD/CN subjects from ADNI1 for training without significant differences between the two age distributions ($p_{value} = 0.27$). The evaluation consisted of two different tasks: Diagnosis (main task) and Prognosis (unseen task).

2.3 Methods

2.3.1 Method overview

An overview of our proposed pipeline is shown in Figure 2.1. Our pipeline is designed based on different blocks, each of which serves a distinct purpose. First, the role of the collective AI block is to simulate a big model that cannot fit into a GPU by a large ensemble of smaller models. This strategy may help to capture more disease-related patterns

than a single model. Indeed, it shows an improvement in generalization (see Section 2.4.1) compared to other techniques. Second, the deep grading map provides a quantitative and interpretable assessment of the progression of AD. This 3D map can show AD-related regions, providing insight into the model prediction and helping clinicians in making reliable decisions. We use a segmentation here for a better visualization of the grading map and to reduce the data dimensionality in a meaningful way for experts. Finally, we use GCN to capture the relationship between brain structures. We demonstrate that GCN is well-adapted with grading features for AD detection (see Section 2.4.1).

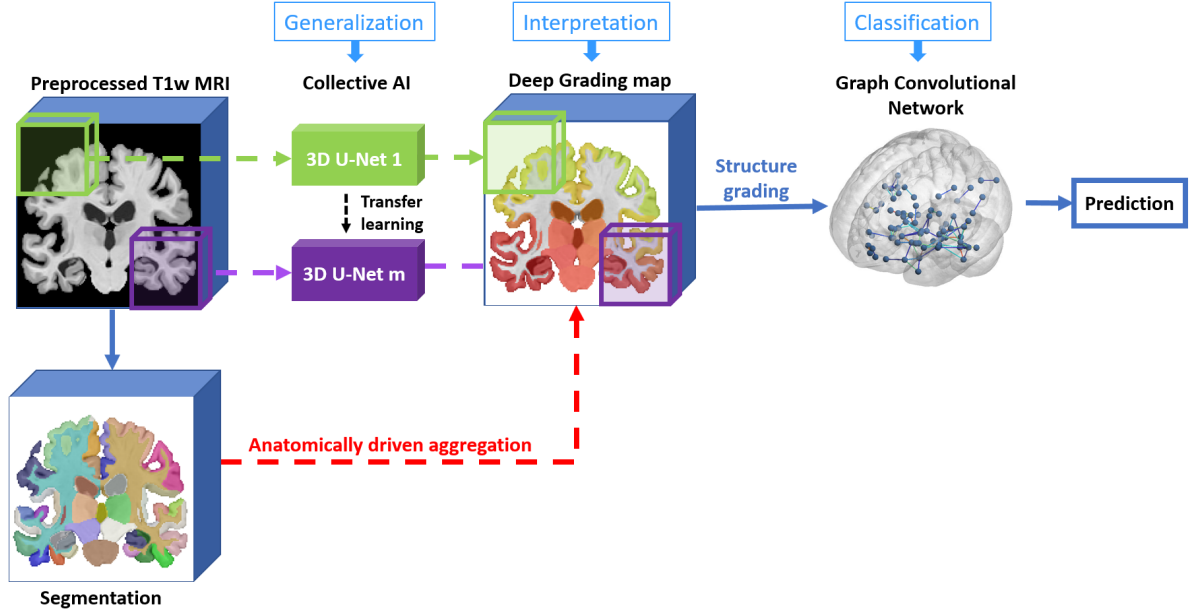


Figure 2.1: Overview of our processing pipeline. The MRI image, its segmentation and the deep grading map illustrated here are from an AD subject.

Concretely, a preprocessed T1-weighted MRI with the size of $181 \times 217 \times 181$ voxels was downsampled to $91 \times 109 \times 91$ voxels to reduce the computational cost. The downsampled image was divided into $k \times k \times k$ (*i.e.*, $k = 5$) overlapping patches of the same size (*i.e.*, $32 \times 48 \times 32$ voxels). We used $m = k \times k \times k$ (*i.e.*, $m = 125$) 3D U-Nets to grade these patches. The 125 grading patches were fused to reconstruct a global grading map of $91 \times 109 \times 91$ voxels. This map was upsampled using interpolation to have the same size as the original input. After that, the segmentation of the original input (obtained with AssemblyNet [50]) was used to compute the average grading score for each structure. In this way, we obtained a vector of s elements where s is the number of segmented structures (*i.e.*, $s = 133$). Finally, we created a fully connected graph with s nodes presenting the characteristic of s structures (*e.g.*, structure grading, structure volume, subject’s age) and used a graph convolutional neural network for the classification.

2.3.2 Deep grading

In AD diagnosis and prognosis, most of deep learning models only use CNN as a binary classification tool. In this study, we use CNN to produce 3D interpretable maps indicating the structural alterations caused by AD.

To capture these anatomical alterations, we extend the idea of several patch-based grading frameworks [51, 234, 48, 98]. The main objective is to provide a 3D grading map with a score between -1 and 1 at each voxel reflecting the disease severity. In [51], the authors proposed to grade the hippocampus. For each voxel of this structure, they defined a surrounding patch and used a locally adaptive search algorithm to find the corresponding patch in all of training images. Similarity scores were then computed between the testing patch and training patches. These scores were used to estimate the grade (*i.e.*, degree of similarity to one group or another) for the considered voxel. Then, an average grading value was computed for the structure. The subject was classified as AD or CN depending on the sign of the grading value. They found that grading feature is more powerful than the measure of structure volume in distinguishing AD and CN subjects. Tong *et al.* used a sparse coding process to select a small number of discriminative voxels over the whole brain [234]. They showed that grading feature was efficient for AD prognosis even when training with AD/CN subjects. Contrary to these previous methods based on handcrafted feature extraction, here we propose a novel deep grading framework based on a large ensemble of 3D U-Nets (*i.e.*, 125 U-Nets).

Concretely, each of our 125 U-Nets (with the architecture similar to [50]) takes a 3D sMRI patch (*e.g.*, $32 \times 48 \times 32$ voxels in the MNI space) and outputs a grading map with value in range $[-1, 1]$ for each voxel. Voxels with a higher value are considered closer to AD, while voxels with a lower value are considered closer to CN. For the ground-truth used during training, we assign the value 1 (resp. -1) to all voxels inside a patch extracted from an AD patient (resp. CN subject). All voxels outside of ICC are set to 0.

Once trained, the deep models are used to grade patches. These local outputs are gathered to reconstruct the final grading map (see Section 2.3.3). Using the structure segmentation, we represent each brain structure grading by its average grading score (see Figure 2.1). This anatomically driven aggregation allows better and meaningful visualization of the disease progression. In this way, during the classification step (see Section 2.3.4), each subject is encoded by an s -dimensional vector where s is the number of brain structures (*i.e.*, $s = 133$).

2.3.3 Collective AI

In medical analysis, high generalization capacity across domains and unseen tasks presents potential clinical value as real data is diverse and may come from any source. As recently shown in [34, 250], current deep learning methods for AD classification can well generalize to similar datasets but poorly perform on datasets having differences such as MRI protocols, age ranges, country of origin or inclusion criteria. In our testing datasets, different age range and MRI protocol were present in OASIS, different country of origin was present in AIBL and different inclusion criteria was present in MIRIAD. It should be noted that for OASIS, MCI and AD patients are mixed, so we used the ADNI inclusion criteria to separate AD patients and be able to assess the diagnosis of AD.

In this study, we propose to use an innovative collective artificial intelligence strategy to improve the generalization across domains and to unseen tasks. As recently shown for segmentation problems [50, 127], the use of a large number of compact networks capable of communicating offers a better capacity for generalization. For brain segmentation, this strategy showed strong generalization to previously unexplored domains [50] (*i.e.*, trained on healthy adults and tested on children and AD patients). For the problem of multiple sclerosis lesion segmentation, this strategy also demonstrated the consistency across different natures of training domains [127]. There are many other advantages of using the collective AI strategy. First, the use of a large number of compact networks is equivalent to a big neural network with more filters but the computation capacity required remains affordable. It should be noted that the same model taking the whole image at full resolution cannot be trained due to the limited memory of current GPUs. Second, the voting system based on a large number of specialized and diversified models helps the final grading decision to be more robust against domain shift and different tasks.

Concretely, after preprocessing and downsampling steps, we obtain $m = k \times k \times k$ patches X_1, \dots, X_m (*i.e.*, $m = 125$) with about 50% overlapping volume. During training, for each patch location, a specialized model is trained. Therefore, we train m 3D U-Nets to cover the whole image (see Figure 2.1). Moreover, each U-Net is initialized using transfer learning from its nearest neighbor U-Net, except the first one trained from scratch as proposed in [50]. As adjacent patches share common patterns, this communication allows grading models to share useful knowledge between them.

To obtain the final grading map, we propose a robust fusion strategy based on an average between overlapping patches, weighted by the accuracy obtained on the validation set. This weighted average for grading score fusion is computed as follows:

$$G_i = \frac{\sum_{x_i \in X_j} a_j * g_{ij}}{\sum_{x_i \in X_j} a_j}$$

where G_i is the grading score of the voxel x_i in the final grading map, g_{ij} is the grading score of the voxel x_i in the local grading patch X_j , and a_j is the balanced accuracy on validation of the patch j . This weighted vote enables to give more weight to the decision of accurate models during the reconstruction.

2.3.4 Feature classification

Most of current methods globally compared classes (*e.g.*, AD *vs.* CN) to perform classification. This kind of approach finds useful information from inter-subject similarities. For Alzheimer’s disease, the anatomical changes may occur in different brain areas and are different between subjects. These intra-subject variabilities may provide useful information for accurate AD detection. Consequently, it should be beneficial to combine these two characteristics for efficient classification. This can be done with the help of graph-based modeling. Indeed, following the idea of [98], we modeled the intra-subject variabilities using a graph representation to capture the relationships between brain regions. We defined an undirected graph $\mathcal{G} = (\mathbb{N}, \mathbb{E})$, where $\mathbb{N} = n_1, \dots, n_s$ is the set of nodes for the s brain structures and $\mathbb{E} = s \times s$ is the matrix of edge connections. In our approach, all nodes were connected with each other in a complete graph, where nodes embed brain features (*e.g.*, our proposed DG feature) and potentially other types of external features.

Indeed, besides the grading map, the volume of structures obtained from the segmentation could be helpful to distinguish AD patients from CN since AD yields to structure atrophy [234, 98]. In addition, the subject’s age is also an important factor since anatomical patterns in the brain of young AD patients could be similar to elder CN. Therefore, the combination of those features is expected to improve our classification performance. In our graph, each node could embed the structure grading score DG, structure volume V, and subject’s age A. All possible combinations are studied in Section 2.4.1. Different types of graph edges are compared in Section 2.4.1. Finally, we used a graph convolutional neural network (GCN) [133] as the way to pass messages between nodes and perform the final decision. A comparison between different classifiers is provided in Section 2.4.1 to explain our choice of GCN.

2.3.5 Implementation details

For each of the 125 patch locations, 80% of the training dataset (*i.e.*, ADNI1) was used for training a 3D U-Net and the remaining 20% for validation. To avoid bias resulting from dataset imbalance, the training/validation sets employed the same number of AD and CN. As the number of images in ADNI1 dataset was small, the training/validation

data was re-split for each patch location to exploit the maximum information possible. The model was trained with voxel-wise mean absolute error (MAE) loss and Adam optimizer with a learning rate of 0.001. All voxels equally contribute to the loss function during training. The training process is stopped after 20 epochs without improvement in validation loss. We employed several data augmentation and sampling strategies to alleviate the overfitting issue during training. To train a U-Net for the i -th patch, first, the corresponding cropping position of patch was randomly translated by $t \in \{-1, 0, 1\}$ voxel in 3 dimensions of the image. Second, we sampled a patch X_{i_1} (with the label Y_{i_1}) from AD population, another patch X_{i_2} (with the label Y_{i_2}) from CN population and applied Mixup technique [261] to create a new sample: $X_{new} = \alpha X_{i_1} + (1 - \alpha)X_{i_2}$, $Y_{new} = \alpha Y_{i_1} + (1 - \alpha)Y_{i_2}$ where $\alpha \sim \text{Beta}(0.3, 0.3)$. This sample was used as the only input during the training.

Once the DG feature was obtained, we represented each subject by a graph of 133 nodes. Each node represented a brain structure and embeds its characteristic (*e.g.*, DG, V, A). Our classifier was composed of three layers of GCN [133] with 32 channels, followed by a global mean average pooling layer and a fully connected layer with an output size of 1. The model was trained using the binary cross-entropy loss, Adam optimizer with a learning rate of 0.0003. No data augmentation was applied during training. The training process was stopped after 20 epochs without improvement in validation loss. During testing, we randomly added noise $X_{noise} \sim \mathcal{N}(0, 0.01)$ to the node features and computed the average of 3 predictions to get the global decision [248]. Experiments showed that it helps our GCN to be more stable. For training and evaluating steps, we used a NVIDIA TITAN X with 12GB of memory. The total training time for $m = 125$ U-Nets and the GCN model is about 23 hours. The total inference time of our method is about 1.63 seconds per preprocessed image.

2.4 Experimental results

2.4.1 Performance study

In this section, the 125 CNN grading models and the classifier were trained using AD and CN subjects of the ADNI1 dataset. Then, we assessed their generalization capacity to domain shift using AD and CN subjects from ADNI2, AIBL, OASIS and MIRIAD. The generalization capacity for unseen tasks was studied using pMCI, sMCI subjects (AD prognosis) from ADNI1 (same domain) and AIBL (out of domain). Due to the imbalanced nature of testing datasets, we used the balanced accuracy (BACC) and area under receiver operating characteristic curve (AUC) to measure the performance of different

classifiers. The global BACC/AUC for diagnosis and prognosis was measured with all available testing images for each task. Each experiment was repeated ten times (to reduce bias related to random nature of DL training) and the average results was provided as final results. All pair-wise comparisons were made using the Wilcoxon test by comparing the ten BACC/AUC values obtained over the 10 repetitions as recommended in [60]. The one-sided test was applied to confirm a superior performance. A confidence level of 5% is used so that $p_{value} < 0.05$ means the considered result is significantly better than a chosen baseline.

Features for classification

In this part, we study the different feature types used as input of the final classifier. The edges connecting the graph nodes are set to 1 in this comparison. The DG feature is denoted as DG_C (resp. DG_I) when obtained with the collective (resp. individual) AI strategy. The individual AI strategy refers to the use of a single U-Net to learn patterns from all patches of the input image. We also denote DG_{Cnw} for the no-weighted version of DG_C . The results of BACC performance are presented in Table 2.2. The result of AUC are in annexes (Table A.1).

Comparison of Grading vs. Volume

As discussed previously, brain atrophy is an important biomarker of Alzheimer’s disease. Many studies used structure volume for AD classification and achieved encouraging results [91, 142, 214]. So, we compare the proposed biomarker (grading, exp. 3) and the classical one (volume, exp. 4) to assess the efficiency of our new biomarker. The additional evaluation using the age feature (exp. 5) was performed to confirm that no age bias was present in the training/testing partitions.

The efficiency of DG_C (exp. 3) was clearly better than V (exp. 4). DG_C outperformed V in global diagnosis, global prognosis and all of the tests on an individual dataset (all $p_{value} < 0.05$). Thus, the proposed biomarker DG_C presents an important interest for AD classification.

Moreover, we trained a UMAP [179] with AD/CN subjects from ADNI1 and visualized the transformed test set in 2D space (see Figure 2.2). The transformed data was colored with respect to the diagnosis class. Two types of input were considered: grade (DG_C) and volume (V). The grading feature was visually better to separate AD and CN subjects than the volume feature. To confirm this assessment, we applied K-means with 2 clusters (we considered 1 cluster for CN/sMCI and 1 cluster for AD/pMCI) to this 2D data to assess the separability of the two clusters. The silhouette score [211] was used to

Table 2.2: Comparison of different types of features for classification for AD diagnosis and prognosis. All the edges are set to 1, the classifier used is GCN. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. All the methods were trained on the AD/CN subjects of the ADNI1 dataset. Value in bold: p of one-sided Wilcoxon test comparing with our baseline (in gray) is lower than 0.05, meaning a significantly superior performance is found compared to the baseline. A comparison using area under curve (AUC) is provided in annexes.

| No. | Features | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis (p/sMCI) |
|-----|--------------|----------------------|-------------------|--------------------|--------------------|-----------------------|------------------|--------------------------------|---------------------------------|
| | | ADNI2 $N = 330$ | AIBL $N = 279$ | OASIS $N = 756$ | MIRIAD $N = 69$ | ADNI1 $N = 300$ | AIBL $N = 32$ | All $N = 1434$ | All $N = 332$ |
| 1 | DG_I | 88.6 | 82.3 | 88.0 | 96.2 | 68.2 | 71.4 | 88.4 | 68.2 |
| 2 | DG_{Cnw} | 86.4 | 88.0 | 89.1 | 99.3 | 70.3 | 73.0 | 88.5 | 70.4 |
| 3 | DG_C | 87.2 | 88.5 | 88.9 | 99.8 | 70.6 | 75.4 | 89.0 | 71.0 |
| 4 | V | 67.4 | 64.0 | 72.8 | 70.6 | 56.1 | 61.2 | 69.8 | 56.5 |
| 5 | A | 50.5 | 52.7 | 46.1 | 42.2 | 49.8 | 50.3 | 46.5 | 50.0 |
| 6 | V, A | 63.2 | 59.8 | 58.5 | 54.5 | 52.9 | 55.7 | 57.6 | 53.0 |
| 7 | DG_C, V | 86.3 | 88.4 | 88.4 | 98.7 | 70.8 | 75.2 | 88.3 | 71.0 |
| 8 | DG_C, A | 87.5 | 92.1 | 88.8 | 99.0 | 73.8 | 74.5 | 89.5 | 73.7 |
| 9 | DG_C, V, A | 87.3 | 91.8 | 88.2 | 98.7 | 73.9 | 72.7 | 88.9 | 73.6 |

measure this separability. This score ranges from -1 to 1 . A higher value means clusters are more distinguishable. As a result, the silhouette score obtained with DG_C was 0.55 , better than 0.41 obtained with V .

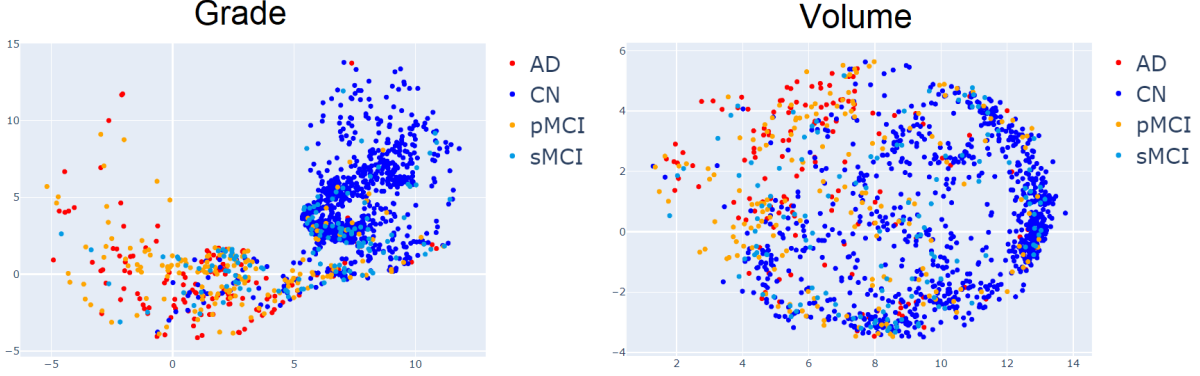


Figure 2.2: UMAP visualization of test set for the AD diagnosis and prognosis.

Comparison of Collective AI features vs. Individual AI features

We aimed at assessing the efficiency of the collective AI strategy. To do this, we compared the efficiency of DG_C and DG_I features (exp. 1, 3) (see Table 2.2). Experimental results showed that DG_C (exp. 3) is significantly better than DG_I (exp. 1) for both global diagnosis ($p_{value} = 0.007$) and global prognosis ($p_{value} = 0.001$). Consequently, collective AI strategy offered a significant improvement for unseen domain (AD diagnosis) and unseen task (AD prognosis). In terms of generalization, DG_C alleviated the drop in performance in AIBL dataset for AD diagnosis.

Efficiency of the weighted fusion strategy in Collective AI features

We validate the efficiency of the weighted fusion strategy in Collective AI by comparing this strategy with its no-weighted version (exp. 2, 3) (see Table 2.2). Experimental results showed that DG_C (exp. 3) is significantly better than DG_{Cnw} (exp. 2) for global diagnosis ($p_{value} = 0.019$) and similar for global prognosis ($p_{value} = 0.188$). Consequently, the weighted fusion strategy can improve the model performance in AD diagnosis while keeping a good performance in AD prognosis.

Combination of grading and additional features

Several works showed that complementary information about the subject could help to improve the performance of their classifier [234, 99]. In these studies, different cognitive scores were used such as MMSE, CDR-SB, RAVLT, FAQ, ADAS11, and ADAS13

cognitive tests. However, this information is not always available. Instead, we employed brain structure volumes and the subject’s age as additional features here.

Four experiments were made using multiple types of features in graph nodes (exp. 6, 7, 8, 9). The best performance of diagnosis and prognosis was obtained using DG_C, A (exp. 8) and it was significantly better than using only DG_C (exp. 3) for both global scores (all $p_{value} < 0.05$). Overall, using subject’s age in addition to DG_C produced the best results. Consequently, in the rest of the chapter, we use DG_C and the age feature as input for further analysis.

Comparison of different types of graph edges

In this part, we compare different types of graph edges. In general, when constructing a graph from neuroimage data, there are different ways to define the connection between nodes [26]. Huang et al. defined it by counting fiber tracts in Diffusion Tensor Imaging [105]. Li et al. computed this as the pairwise correlations of functional magnetic resonance imaging time series [150]. With sMRI data, Mahjoub et al. defined the connection between two ROI as the absolute difference between their averaged cortical attributes [168]. In this study, we propose different edge types as follows: Fully-one edge (all edges are set to 1), correlation-based edge (the edge connecting each pair of brain structures is defined as the Pearson’s correlation based on their grading scores), volume difference-based edge (the edge connecting each pair of brain structures is the absolute difference of their volumes). The results of the comparison are presented in Table 2.3. We observe that the edge based on structure volume difference leads to a better classification performance than other tested types of edge in all datasets and all tasks (almost all $p_{value} < 0.05$). Thus, we use the edge based on structure volume difference in the rest of the chapter.

Comparison of different classifiers

In this section, we study different solutions for the graph classification. We compare the use of GCN with other classifiers such as SVM, multi-layer perceptron, Transformer Graph [219], sample and aggregate graph (SAGE) [93], residual gated graph (Res-GatedGraph) [31], graph attention network (GAT) [244] and topology adaptive graph (TAG) [64]. Table 2.4 shows the results of this comparison. We can observe that GCN achieves the best performance most of the time (all $p_{value} < 0.05$ for global diagnosis and prognosis). Consequently, we chose GCN as a classifier in our framework.

Table 2.3: Comparison of different graph edge types for AD diagnosis and prognosis. The classifier used is GCN and the input features is DG_C and A. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. All the methods were trained on the AD/CN subjects of the ADNI1 dataset. A comparison using area under curve (AUC) is provided in annexes.

| Edge | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis (p/sMCI) |
|-------------------|----------------------|-----------|-----------|----------|-----------------------|----------|--------------------------------|---------------------------------|
| | ADNI2 | AIBL | OASIS | MIRIAD | ADNI1 | AIBL | All | All |
| | $N = 330$ | $N = 279$ | $N = 756$ | $N = 69$ | $N = 300$ | $N = 32$ | $N = 1434$ | $N = 332$ |
| Fully-one | 87.5 | 92.1 | 88.8 | 99.0 | 73.8 | 74.5 | 89.5 | 73.7 |
| Correlation | 87.5 | 91.8 | 88.4 | 98.6 | 73.4 | 74.1 | 89.2 | 73.3 |
| Volume difference | 87.6 | 92.4 | 89.1 | 99.6 | 73.9 | 75.6 | 89.6 | 73.9 |

Table 2.4: Comparison of different classifiers for AD diagnosis and prognosis. For graph-based approaches (*i.e.*, all the approaches except SVM and multi-layer perceptron), the edge based on structure volume difference is used and the input features is DG_C and A. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. All the methods were trained on the AD/CN subjects of the ADNI1 dataset. A comparison using area under curve (AUC) is provided in annexes.

| Classifier | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis |
|------------------------|----------------------|-------------|-------------|-------------|-----------------------|-------------|--------------------------------|---------------------|
| | ADNI2 | AIBL | OASIS | MIRIAD | ADNI1 | AIBL | All | All |
| | $N = 330$ | $N = 279$ | $N = 756$ | $N = 69$ | $N = 300$ | $N = 32$ | $N = 1434$ | $N = 332$ |
| SVM | 85.7 | 88.7 | 87.4 | 95.6 | 69.0 | 69.7 | 87.6 | 68.9 |
| Multi-layer perceptron | 82.5 | 87.4 | 83.4 | 88.0 | 66.4 | 61.7 | 84.6 | 65.8 |
| Transformer | 87.9 | 91.3 | 87.9 | 98.5 | 72.8 | 75.4 | 89.1 | 72.9 |
| SAGE | 87.2 | 91.8 | 88.1 | 98.3 | 73.4 | 73.3 | 88.9 | 73.2 |
| ResGatedGraph | 84.6 | 87.6 | 81.9 | 92.7 | 72.5 | 70.8 | 84.0 | 70.3 |
| GAT | 87.7 | 91.6 | 88.7 | 98.2 | 73.4 | 72.5 | 89.3 | 73.1 |
| TAG | 87.4 | 91.3 | 87.8 | 97.7 | 73.3 | 74.2 | 88.8 | 73.2 |
| GCN | 87.6 | 92.4 | 89.1 | 99.6 | 73.9 | 75.6 | 89.6 | 73.9 |

Comparison with state-of-the-art methods

Tables 2.5 and 2.6 summarize the current performance in BACC of state-of-the-art methods proposed for AD diagnosis and prognosis classification that have been validated on external datasets. A comparison of performance in AUC is provided in annexes (Tables A.4 and A.5). In this comparison we consider five categories of deep methods: patch-based strategy based on a single model (Patch-based CNN [250]), patch-based strategy based on multiple models (Landmark-based CNN [157], Hierarchical FCN [152]), ROI-based strategy based on a single model focused on hippocampus (ROI-based CNN [250]), subject-based considering the whole image based on a single model (Subject-based CNN [250], Efficient 3D [257] and AD^2A [90]) and a classical voxel-based model using a SVM (Voxel-based SVM [250]). Only methods evaluated across different datasets were selected here.

Comparison with methods under the same condition

For a fair comparison, we retrained and evaluated four methods whose code is available: Patch-based CNN, ROI-based CNN, Subject-based CNN and Voxel-based SVM [250] with our training/testing data. The results are reported in Table 2.5.

For AD diagnosis (*i.e.*, AD/CN), as ADNI2 and ADNI1 (training set) are very similar, we used the performance on ADNI2 as a reference to assess the capacity of generalization on other datasets (*i.e.*, AIBL, OASIS, MIRIAD). Based on that, we observed a major drop in performance in Patch-based CNN method for AIBL, OASIS and MIRIAD, ROI-based CNN method for AIBL (see Table 2.5). For AD prognosis (*i.e.*, pMCI/sMCI), we also observed a drop in performance between AIBL and ADNI1 (training domain) in Patch-based CNN method, ROI-based CNN method and Subject-based CNN method. Overall, our method shows a good generalization capacity against domain shift and to unseen tasks compared to other methods. Moreover, our method always achieves the best result in terms of performance for all datasets/tasks and outperforms the traditional method (*i.e.*, Voxel-based SVM) by a large margin.

Literature comparison

We also detail the results of four other methods without available implementation performing evaluation across different datasets. In this case, we present the results of the original papers in Table 2.6. Consequently, there are many different factors between methods: number of subjects in training/testing sets, selection criteria, etc. However, this could help to get an idea of the performance of current methods in the application of AD diagnosis/prognosis.

Table 2.5: Comparison of our method with state-of-the-art methods with available code that have been retrained on our training dataset and tested on our dataset. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. All the methods are trained on the AD/CN subject of the ADNI1 dataset, the same training/testing partition is used for evaluation. A comparison using area under curve (AUC) is provided in annexes.

| Method | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | |
|-------------------------|----------------------|-------------|-------------|-------------|-----------------------|-------------|
| | ADNI2 | AIBL | OASIS | MIRIAD | ADNI1 | AIBL |
| | $N = 330$ | $N = 279$ | $N = 756$ | $N = 69$ | $N = 300$ | $N = 32$ |
| Patch-based CNN [250] | 72.4 | 63.4 | 67.5 | 63.0 | 62.5 | 47.5 |
| ROI-based CNN [250] | 79.7 | 74.4 | 79.0 | 81.5 | 65.5 | 62.5 |
| Subject-based CNN [250] | 76.1 | 81.5 | 86.0 | 89.1 | 64.8 | 55.8 |
| Voxel-based SVM [250] | 83.3 | 88.2 | 87.4 | 93.5 | 67.2 | 70.0 |
| Our method | 87.6 | 92.4 | 89.1 | 99.6 | 73.9 | 75.6 |

Overall, our method has most of the time the best or the second best result. Furthermore, it should be noted that our model is trained using only 340 images (from ADNI1) without any domain adaptation technique but outperforms Efficient3D (trained on 2843 images) and AD^2A (with domain adaptation) in most of datasets/tasks.

2.4.2 Interpretation of deep grading maps

To highlight the interpretability capabilities offered by our DG feature, we computed the average DG map for each group: AD, pMCI, sMCI and CN (see Figure 2.3). First, we could note that the average grading increased between each stage of the disease. Second, we estimated the top 10 structures with highest absolute value of grading score over all the testing subjects. Nine of these structures were known to be specifically and early impacted by AD. These structures were: bilateral hippocampus [81], left amygdala and left inferior lateral ventricle [52], left parahippocampal gyrus [129], left posterior insula [76], left thalamus [124], left transverse temporal gyrus [159], left ventral diencephalon [141]. These results showed a high correlation with current physiopathological knowledge on AD [116].

Typical individual grading maps of each population (*i.e.*, CN, sMCI, pMCI, AD) were selected and are presented in Figure 2.4. First, we observed that older people had higher grade than younger people as expected. Second, for the same age range, the color of grading maps changed progressively depending to the disease severity. Third, CN/AD populations seemed to be more distinguishable from each other than sMCI/pMCI populations. We observed high similarity between older sMCI patients (80-90 years old) and

Table 2.6: Comparison of our method with state-of-the-art methods using published results. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. All the methods are trained on the AD/CN subject of the ADNI1 dataset. However, there are many different factors: number of subjects in training/testing sets, selection criteria, etc. A comparison using area under curve (AUC) is provided in annexes.

| Method | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | |
|--------------------------|----------------------|-------------|-------------|-------------|-----------------------|-------------|
| | ADNI2 | AIBL | OASIS | MIRIAD | ADNI1 | AIBL |
| Landmark-based CNN [157] | 90.8 | - | - | 92.4 | - | - |
| Hierarchical FCN [152] | 89.5 | - | - | - | 69.0 | - |
| AD^2A [90] | 88.3 | 87.8 | - | - | - | - |
| Efficient3D [257] | - | 90.7 | 91.9 | 95.7 | 70.1 | 65.2 |
| Our method | 87.6 | 92.4 | 89.1 | 99.6 | 73.9 | 75.6 |

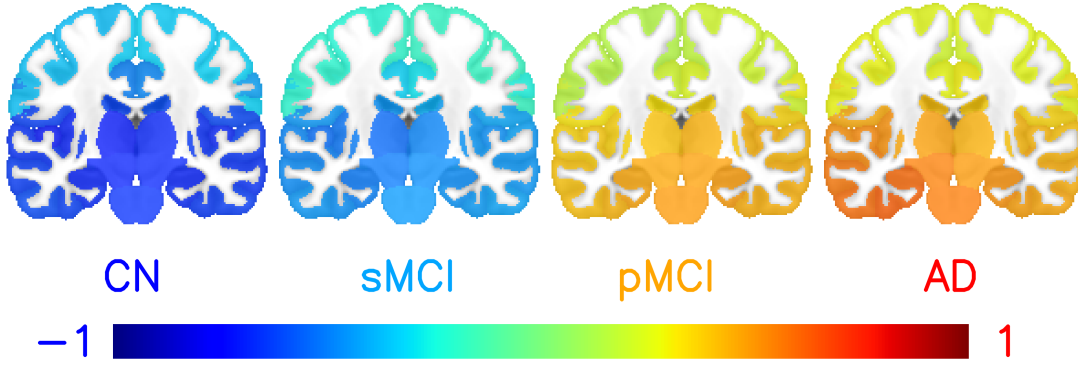


Figure 2.3: Average grading map per group of subjects (CN, sMCI, pMCI and AD).

younger pMCI patients (60-70 years old). This might be the reason why the performance of AD prognosis was lower than AD diagnosis and why the use of age improved the results of AD prognosis. Finally, we observed that the earliest brain alteration started from hippocampus and its surrounding regions (sMCI at 70-80 years old in Figure 2.4) and spanned over time to the whole brain (AD at 80-90 years-old in Figure 2.4). All of these findings demonstrated the potential capacity of deep grading maps to assist clinicians in practice.

2.4.3 Consistency study

Thibeau-sutre *et al.* have recently shown that for the same CNN architecture, different training data or even training runs can lead to different explanations [230]. They suggested that a good explanation method should not depend on training data or train-

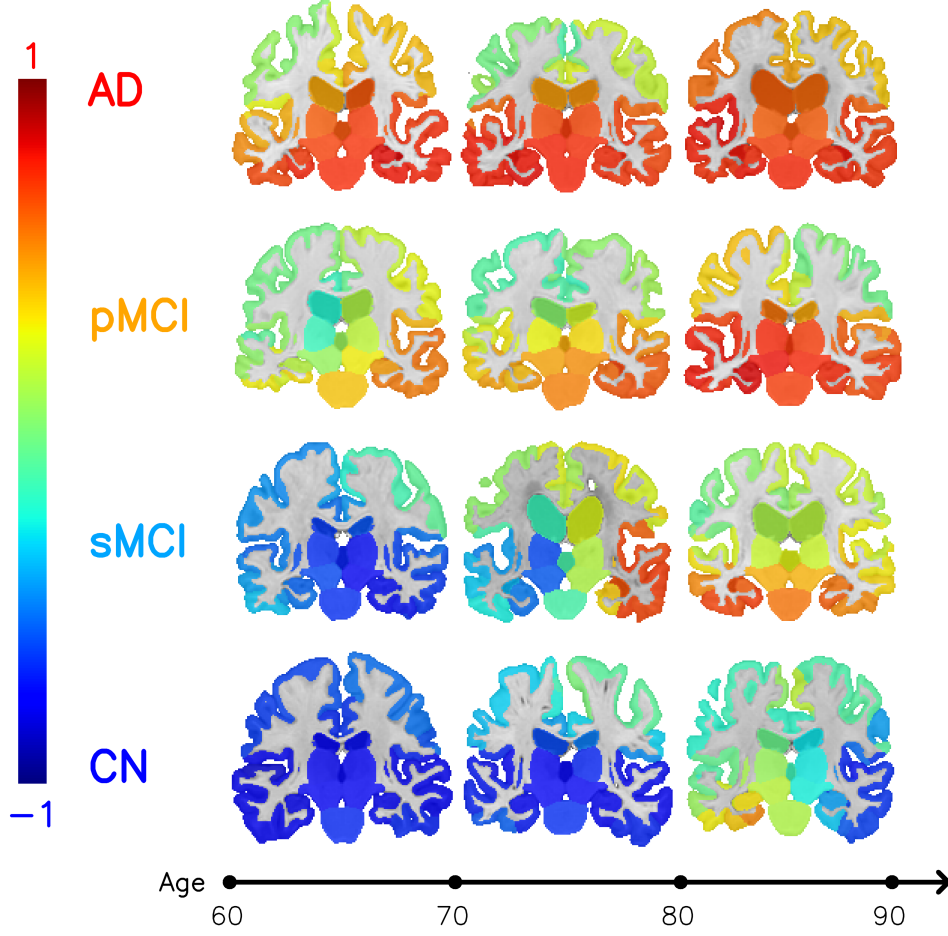


Figure 2.4: Typical grading maps (from individual subjects) for each state of Alzheimer’s disease with respect to age.

ing initialization. In this study, we analyzed these two aspects for our grading maps (dependency to data training and model initialization) by performing two experiments.

First, we trained two grading models (each one consisting of 125 U-Nets) on ADNI1 (model 1) and ADNI2 (model 2) datasets. For each model, we then calculated the DG_C vector from the grading map for all images in testing set (excluding ADNI1 and ADNI2). Finally, we measured the cosine similarity of two DG_C vectors obtained from each image. We obtained a median of 0.92 as similarity between two DG_C vectors from two models training on different datasets that demonstrate the good robustness to domain shift of our method.

For the second one, we trained the grading model twice using only ADNI1 as training set (models 1 & 3). Finally, we obtained a median of 0.95 as similarity between DG_C vectors from two retrained models on the same dataset that demonstrate the good robustness to training initialization of our method.

Figure 2.5 shows examples of individual grading maps for four considered populations

(*i.e.*, CN, sMCI, pMCI, AD). We can visually see the similarities between grading models trained on different datasets (*i.e.*, models 1 & 2) and between grading models trained several times on the same dataset (*i.e.*, models 1 & 3). Overall, the three models identify AD-related areas in a similar way. These experiments show the consistency of Deep Grading maps across different training runs and different training sets.

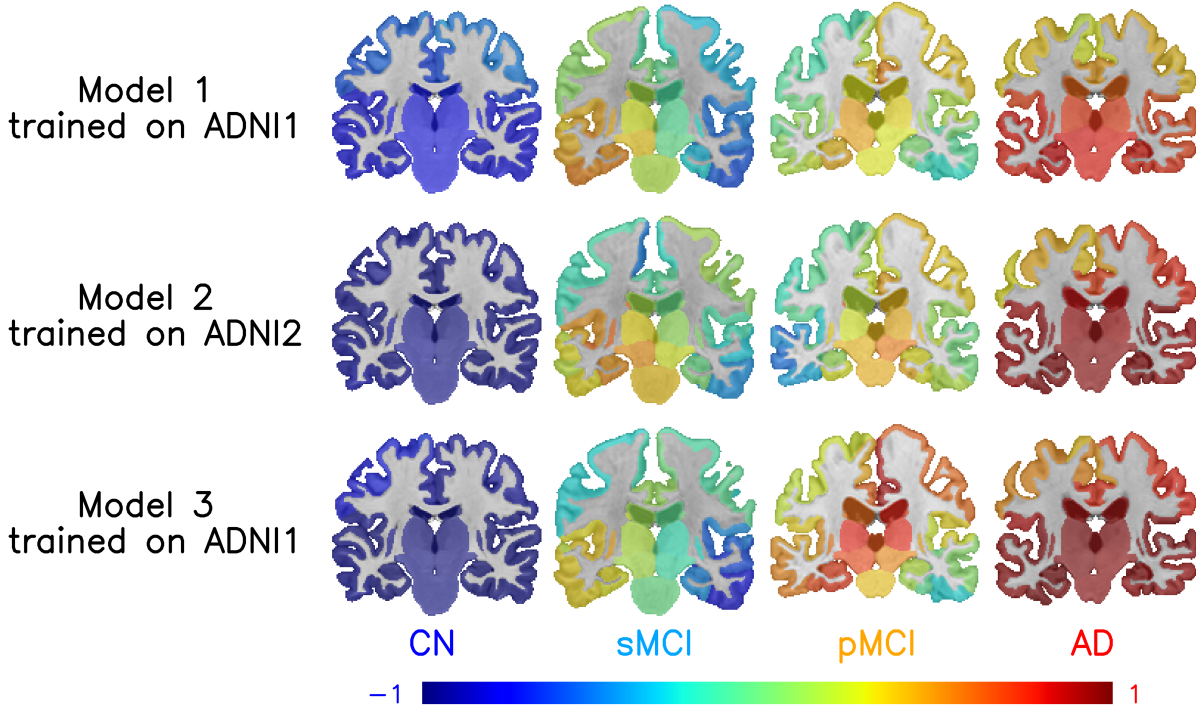


Figure 2.5: Consistency of grading maps between retrained grading models (models 1 & 3), and between grading models trained on different datasets (models 1 & 2).

2.5 Discussion

In this chapter, we proposed a novel deep grading framework dedicated to single-disease diagnosis, applying to Alzheimer’s disease diagnosis and prognosis. Our framework was designed to overcome three main limitations of current deep learning methods for AD classification: performance compared to conventional machine learning method (*i.e.*, SVM), generalization to unseen datasets/tasks and interpretability.

While many studies found that deep learning and SVM methods had a similar performance for AD classification problem [34, 250], the authors also suggested that better model design could improve the performance of DL methods. Indeed, Jo *et al.* indicate that hybrid methods using CNN features and a conventional classifier showed better accuracy than pure deep learning methods [121]. In this study, we combined CNN features

with a GCN classifier. The use of a GCN also allows to combine additional demographic information to further improve the model performance. As a result, our model showed a better performance with a large margin compared to traditional methods (*e.g.*, SVM). Furthermore, a careful design of edge connectivity in the subject’s graph may also boost the model performance. In this study, we propose to define the edge connection between two brain structures as the absolute difference of their volumes. This type of connectivity allows our model to make more accurate decisions. The analysis of this connectivity is provided in Appendix A.

This study is one of the few assessing deep model performance on multiple independent datasets (*i.e.*, ADNI2, AIBL, OASIS, MIRIAD) and unseen tasks (*i.e.*, AD prognosis) [34, 250, 257, 157, 152, 90]. For AD diagnosis, the result on ADNI2 dataset was 87.6% in BACC which was competitive with the current performance reported in the literature. On OASIS, we achieved the second place with 89.1% accuracy. On AIBL and MIRIAD, our model outperformed current state-of-the-art methods with respectively 92.4% and 99.6%. Besides several studies found a drop in performance when evaluating on independent datasets [34, 250]. This performance drop could come from differences in MRI protocols, age ranges, country of origin and inclusion criteria. Our results demonstrated the high generalization capacity of our method against datasets with such differences. Especially, the use of collective AI enabled a better generalization to unseen tasks (*i.e.*, AD prognosis) than other deep learning methods. Finally, the use of the weighted fusion strategy could improve even more the model performance (in AD diagnosis).

In terms of interpretability, our framework provides 3D grading maps capable of indicating regions impacted by AD. The most important structures highlighted by our grading map were correlated with knowledge about the disease in the literature. Furthermore, our experiments showed that grading features were more efficient than volume features for both AD diagnosis and prognosis which confirmed the finding of [51, 49]. When coupling the grading map with a GCN classifier, it yielded high performance across datasets. Hence, grading maps are not only an interpretable visualization but provide also discriminative features for AD classification. However, the use of GCN made our framework become not fully interpretable. The fully-interpretable framework can be done by replacing the final GCN by an SVM classifier or a simple threshold. Although, this implies a trade-off between interpretability and performance.

Compared to the explanation maps of explainable methods such as Class Activation Mapping [266] and Layer-wise relevance propagation [27], our grading map exhibits interesting properties. Indeed, our grading maps provide quantitative value reflecting the disease severity, while explanation maps give qualitative information about the relative importance of each feature during the decision-making process. For example, in an explanation map of an AD subject, we do not know if the non-highlighted regions (structures

unused by the models to take their decision) are healthy or just non-informative (redundant information with other structures, too noisy due to high inter-subject variability, etc.). Moreover, the explanation maps are generally normalized to the same range of values $[0, 1]$, making the comparison of two explanation maps only qualitative.

This study is among a few studies proposing an interpretable model for AD classification problem. Another approach for an interpretable model is to carefully design the graph neural network classifier and its input. Li *et al.* define individual graphs using features extracted from neuroimaging data and a graph neural network with ROI-aware convolutional layer and an appropriate loss function [150]. Similar to our result, this approach can provide both salient brain regions at the subject level and the community level. However, similar to explainable methods discussed above, it cannot provide quantitative information on the disease severity for a given region.

While the performance of our framework across external datasets and unseen tasks was quite high, there also exist some limitations. First, the ground truth used for grading was potentially not optimal due to a lack of consensus on structures relevant to Alzheimer’s disease. Indeed, there may be some structures that are not impacted by AD and the ground-truth of these structures should be zero. With our ground-truth annotation, small structures surrounding another one highly related to AD had a high chance to appear together in all patches. Thus, those structures would be also predicted as related to AD. Another direction should focus on an unsupervised learning manner to find only abnormalities caused by AD to improve the interpretability. Second, this study exploited only structural MRI while the performance could be improved using multi-modal inputs such as PET, functional MRI, diffusion MRI or perfusion MRI [33]. Better disease patterns are expected to be learned with this kind of input. However, a multi-modal input implies even larger differences between different datasets. Thus, a new generalization study should be considered to see if the gain in performance from better disease patterns can overcome the performance drop resulting from differences between different datasets.

Chapter 3

Multi-channel deep grading for differential diagnosis

| | | |
|-------|--|----|
| 3.1 | Introduction | 57 |
| 3.2 | Materials | 58 |
| 3.3 | Method description | 59 |
| 3.3.1 | Method overview | 59 |
| 3.3.2 | Multi-class Deep Grading-based classification | 59 |
| 3.3.3 | Atrophy-based classification | 61 |
| 3.3.4 | Implementation details | 61 |
| 3.4 | Experimental results | 62 |
| 3.4.1 | Ablation study for binary classification tasks | 63 |
| 3.4.2 | Performance for multi-disease classification | 64 |
| 3.4.3 | Comparison with state-of-the-art methods | 65 |
| 3.4.4 | Interpretation of deep grading map | 67 |
| 3.5 | Discussion | 69 |

The two methods presenting in this chapter is related to the following publications and softwares:

- [5] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Interpretable Differential Diagnosis for Alzheimer’s Disease and Frontotemporal Dementia”. In: *Medical Image Computing and Computer Assisted Intervention. MICCAI 2022*. Vol. 13431. 2022. DOI: [10.1007/978-3-031-16431-6_6](https://doi.org/10.1007/978-3-031-16431-6_6).
- [2] **Nguyen, Huy-Dung**, Michaël Clément, Vincent Planche, Boris Mansencal, and Pierrick Coupé. “Deep grading for MRI-based differential diagnosis of Alzheimer’s disease and Frontotemporal dementia”. In: *Artificial Intelligence in Medicine* 144 (2023), p. 102636. DOI: [10.1016/j.artmed.2023.102636](https://doi.org/10.1016/j.artmed.2023.102636).
- [8] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *AssemblyNet-AD-FTD*. Under deposit.

In this chapter, we will focus on multi-disease differential diagnosis problems, with the differential diagnosis of CN *vs.* AD *vs.* FTD as the chosen targeted application. This case study is selected because AD and FTD are among the most prevalent forms of dementia and they exhibit overlapping clinical symptoms and morphological changes in the brain. These similarities pose challenges for clinicians attempting to accurately distinguish between these diseases and establish appropriate treatment plans for patients.

Although the DG biomarker presented in the previous chapter is effective for single-disease classification, its inherent design makes it unsuitable for multi-disease problems. The DG design takes into account only one pathology, with its severity represented by a single score. In situations involving multiple diseases, it is necessary to jointly determine both the presence and severity of each disease. Utilizing a single scalar to describe multiple diseases in this context is not feasible. As a result, it is essential to extend the DG biomarker to address the challenges associated with multi-disease differential diagnosis. In this chapter, we will explore how to adapt the DG biomarker for multi-disease classification problems.

3.1 Introduction

AD and FTD, both classified as forms of dementia, demonstrate distinct age-related patterns of occurrence. Specifically, AD is more prevalent among individuals aged 65 and above, whereas FTD exhibits comparable incidence rates to AD within the 45 to 65 age range. They also have other differences such as AD patients have more problems with visuospatial abilities or praxies while FTD patients have more frequent and severe behavioral changes¹. However, there are also many overlapping symptoms, such as episodic memory loss, dysexecutive syndrome and/or language impairment [29]. Accurate differential diagnosis is essential for the management of patient’s daily life and for the implementation of dedicated clinical trials. However, the similar symptoms mentioned above make the diagnosis challenging, even though the two diseases have different clinical diagnostic criteria [206, 180]. Moreover, the prevalence of FTD is lower compared to AD (about 300-fold smaller) [66], limiting our knowledge of FTD pathology. Indeed, many studies have demonstrated that isolated cognitive tests cannot reliably distinguish FTD from AD populations [110, 258]. Furthermore, CN people may also exhibit some changes in behavior and memory as a result of the natural aging process. Consequently, a multi-class differential diagnostic tool that could distinguish between AD, FTD and CN would be extremely helpful in clinical practice. Indeed, such a tool can help clinicians in reviewing their hypotheses and thus, making more informed decisions.

Several studies have demonstrated that AD and FTD can be individually detected using structural magnetic resonance imaging (sMRI) [65, 184]. The areas of atrophy caused by the two diseases may differ [58]. For instance, AD seems to mainly affect the medial temporal area [114] while FTD affects different regions depending on its subtypes [164]. The behavioral variant frontotemporal dementia (bvFTD) is often associated with atrophy in the frontal and anterior temporal region. Patients with Progressive non-fluent aphasia (PNFA) have motor speech impairments, mainly controlled by the left inferior frontal lobe. The semantic variant (SV) mainly affects the left anterior temporal area [199]. Hence, using sMRI for disease classification and differential diagnosis should be beneficial. Indeed, some approaches have previously been proposed to address these problems using volumetric and shape measurements extracted from sMRI [65, 202]. However, most existing methods focus only on binary classification tasks (*i.e.*, AD *vs.* CN, FTD *vs.* CN and AD *vs.* FTD). While the multi-class diagnosis provides potential value in clinical practice, only a few studies consider this problem [33, 132, 167, 104]. Additionally, current approaches mainly use traditional machine learning techniques with handcrafted features that might not fully include all disease patterns. As a result, deep

¹<https://www.alz.org/alzheimers-dementia/what-is-dementia/types-of-dementia/fronto-temporal-dementia>

learning techniques have lately been explored. However, the outcomes of these methods are usually difficult to understand. This limitation hinders our understanding of these neurodegenerative diseases.

In this chapter, we propose a biomarker for multi-disease differential diagnosis (*i.e.*, CN *vs.* AD *vs.* FTD). We will further demonstrate its effectiveness for binary classification tasks (*i.e.*, CN *vs.* AD, CN *vs.* FTD, AD *vs.* FTD). In addition to providing an accurate diagnostic tool, our goal is to expand our knowledge about different dementia types. Our contributions in this chapter are two-fold. First, we extend the DG framework to generate 3D grading maps capable of identifying brain regions with specific disease-related patterns (AD-like or FTD-like patterns). To achieve this, we introduce a novel multi-channel grading framework designed to differentiate between three classes. As a result, the grading networks produce a 2-channel disease coordinate (DC) 3D map. This DC map can be converted into an interpretable grading map, which aids clinicians in obtaining a deeper comprehension of AD and FTD pathologies. Furthermore, the DC map can be combined with an MLP for classification. Second, we propose to enhance the MLP decision by incorporating an SVM that uses brain structure volumes to improve the model’s classification performance and generalization capacity. By integrating structural grading and atrophy, the proposed framework exhibits state-of-the-art performance in both disease detection and differential diagnosis.

3.2 Materials

The data used in this study includes 3319 MRIs selected at the baseline from multiple open access databases: ADNI2, NIFD and NACC. As the majority of MRIs with FTD pathology are acquired with 3 Tesla machines, only 3T MRIs are selected for each class. The purpose of this is to avoid possible bias due to the acquisition protocol of different databases [231]. We use ADNI2 (*i.e.*, 180 CN and 149 AD) and NIFD (*i.e.*, 136 CN and 150 FTD) to perform a 10-fold cross-validation. We apply the stratified split strategy to alleviate the bias due to the imbalanced nature of different available classes. The cross-validation result is denoted as in-domain performance. We additionally evaluate our framework on an external dataset (*i.e.*, NACC with visits conducted between September 2005 and November 2021) to assess the generalization capacity of the compared methods or out-of-domain performance. Table 3.1 summarizes the demographic of the subjects used in this study. We only use the three sub-types of FTD in NIFD dataset: bvFTD, PNFA and SV. The reason for this is that the other variant of FTD (*i.e.*, logopenic variant) is typically associated with AD neuropathological changes [95, 23]. Finally, only subjects with consistent diagnosis thorough their follow-up sessions are included in this study.

Table 3.1: Summary of participants used in our study. Data used for training are in bold, therefore MRIs from ADNI2 and NIFD are in-domain data while MRIs from NACC dataset are out-of-domain data.

| | Dataset | Statistic | CN | Dementia | |
|---------------|---------|----------------------|-----------------|----------------|----------------|
| | | | | AD | FTD |
| In-domain | ADNI2 | No. subjects | 180 | 149 | |
| | | Age (Mean \pm Std) | 73.4 ± 6.3 | 74.7 ± 8.1 | |
| | NIFD | No. subjects | 136 | | 150 |
| | | Age (Mean \pm Std) | 63.5 ± 7.4 | | 63.9 ± 7.1 |
| Out-of-domain | NACC | No. subjects | 2182 | 485 | 37 |
| | | Age (Mean \pm Std) | 68.2 ± 10.9 | 72.3 ± 9.6 | 64.1 ± 6.9 |

3.3 Method description

3.3.1 Method overview

Figure 3.1 provides an overview of our method. After the preprocessing pipeline, a T1w MRI is downsampled with a factor of 2 to the size of $91 \times 109 \times 91$ voxels. The resulting image is then used to extract k^3 (*i.e.*, $k = 5$) overlapping sub-volumes of the same size $32 \times 48 \times 32$ voxels and evenly distributed along the 3 image dimensions. We use $m = k^3$ (*i.e.*, $m = 125$) U-Nets to grade these m sub-volumes. The output of one U-Net has a size of $2 \times 32 \times 48 \times 32$ voxels (as the disease status is presented by a 2D point, see Section 3.3.2). The m outputs are then used to reconstruct a DC map of size $2 \times 91 \times 109 \times 91$ voxels. This 2-channels map is upsampled to the same spatial size as the original input. After that, we compute the averaged DC for each brain structure with the help of an AssemblyNet-based brain segmentation [50] (see Section 3.3.2). The structure DC can be either used as input of a MLP classifier for classification or transformed into a 3D grading map for visualization (see Figure 3.1). Moreover, the structure volumes are used as input for an SVM classifier. Finally, we ensemble the results of two classifiers to get the diagnosis prediction.

3.3.2 Multi-class Deep Grading-based classification

In medical imaging applications, it is more beneficial to provide the regions affected by diseases rather than just a classification result. To achieve this, in AD detection, several grading frameworks have been proposed to capture anatomical alterations caused by the disease [51, 234, 48, 98, 4] (see also Section 2.3.2 for more details about these

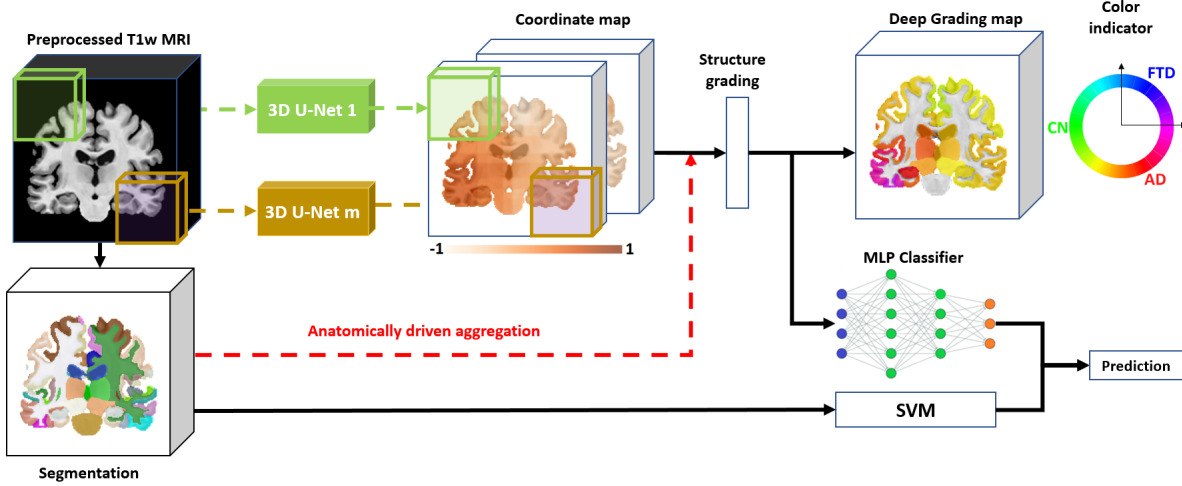


Figure 3.1: An overview of the proposed multi-channel grading method. The T1w image, its segmentation and the deep grading map are taken from an AD patient.

methods). The objective of such approaches is to compute a 3D grading map reflecting the disease severity at the voxel level. Here, we propose to extend the DG framework described in Chapter 2 to the problem of multi-disease diagnosis.

As current grading systems only consider one pathology, its severity may be described by a single score. When many diseases are taken into account, we need to jointly determine which disease is present and also its severity. In this case, using a single scalar is impossible. To this end, we propose to assign each available class to a point in a 2D plan. Concretely, on a circle with a radius of 1, we assign $(-1, 0)$ to CN, $(-\frac{\sqrt{3}}{2}, 0.5)$ to AD and $(\frac{\sqrt{3}}{2}, 0.5)$ to FTD (see the color indicator circle in Figure 3.1). All voxels outside of ICC are set to $(0, 0)$ as they are not related to any pathology. With this definition, a predicted point depicting the disease status can be every point on that circle. Thus, the grading map is not only able to show the severity of each disease but also the common patterns of AD and FTD. We denote this approach as MCDG meaning multi-channel deep grading.

Based on the new definition of ground truth, each of our $m = 125$ U-Nets takes a 3D sub-volume and outputs a DC map with 2 values for each voxel. For instance, when AD-like anatomical patterns are detected in a part of the brain, the produced values in this area should be close to $(\frac{\sqrt{3}}{2}, 0.5)$.

After that, we compute the averaged DC point for each brain structure. The obtained features are denoted as structure DC. By doing this, the grading map is encoded into a 2D matrix of size $2 \times s$ where s is the number of brain structures. Finally, we use a fully connected classifier to perform classification.

3.3.3 Atrophy-based classification

Besides the structure DC features, brain atrophy patterns are also important to identify AD and FTD patients. To exploit the atrophy features, we train an SVM to perform the same classification task using normalized brain structure volumes. The output of the SVM model is combined with the MLP model to make the final decision. The detail of training the SVM and the ensembling process is provided in Section 3.3.4.

3.3.4 Implementation details

In each iteration of 10-fold cross-validation, we used 10 data folds d_i where $i \in \{1, \dots, 10\}$ as follows. First, d_1, \dots, d_7 were used for training/validation of the 125 U-Nets. Then, d_1, \dots, d_7 were re-used for training the MLP (and SVM) classifier and d_8 for its validation. After that, we used d_9 for ensembling the MLP and the SVM model. Finally, the ensemble model was evaluated on d_{10} .

To train each 3D U-Net, the data (d_1, \dots, d_7) is split into 80%/20% for training/validation. The data was common for all of $m = 125$ U-Nets. However, each time we train a new U-Net, this data was combined and re-shuffled before splitting into training/validation to exploit the maximum information possible from our limited data. The loss used during training was voxel-wise mean square error (MSE) with Adam optimizer, batch size of 16 and a learning rate of 3e-4. The first U-Net was trained from scratch and was stopped after 400 epochs without improvement in validation loss. The following U-Nets took advantage of transfer learning from a neighborhood U-Net (see [50] for details) and thus, converted more quickly, their number of epochs for early stopping was set to 100.

To alleviate the overfitting phenomenon while training, we applied the following data augmentation schema: First, we randomly translated a sub-volume by $t \in \{-1, 0, 1\}$ voxel in its 3 axes. Second, we adapted Mixup [261] for MCDG. Concretely, given 2 pairs {input voxel intensity, target DC point}: $\{Intensity_1, (dc_{x_1}, dc_{y_1})\}$, $\{Intensity_2, (dc_{x_2}, dc_{y_2})\}$ taken from 2 patches with class DC target (DC_{x_1}, DC_{y_1}) and (DC_{x_2}, DC_{y_2}) ², the mixup

² $(DC_{xi}, DC_{yi}) \neq (dc_{xi}, dc_{yi}) = (0, 0)$ when the voxel is outside of ICC, see Section 3.3.2

with a coefficient $\alpha \sim \text{Beta}(0.3, 0.3)$ is calculated as follows:

$$\begin{cases} \text{Intensity}_{mixup} = \alpha \times \text{Intensity}_1 + (1 - \alpha) \times \text{Intensity}_2 \\ \phi_1 = \text{atan2}(DC_{y1}, DC_{x1}) \\ \phi_2 = \text{atan2}(DC_{y2}, DC_{x2}) \\ \phi_{mixup} = \alpha\phi_1 + (1 - \alpha)\phi_2 \\ dc_{x_{mixup}} = \cos \phi_{mixup} \times [\alpha(dc_{x1}^2 + dc_{y1}^2) + (1 - \alpha)(dc_{x2}^2 + dc_{y2}^2)] \\ dc_{y_{mixup}} = \sin \phi_{mixup} \times [\alpha(dc_{x1}^2 + dc_{y1}^2) + (1 - \alpha)(dc_{x2}^2 + dc_{y2}^2)] \end{cases}$$

When training the MLP classifier, we used cross-entropy loss with Adam optimizer, batch size of 8 and learning rate of 0.0003.

For the SVM classifier, we applied a grid search of three kernels (linear, polynomial, and radial basis function) and 500 values of C in $[10^{-5}, 10^5]$ on the validation set for tuning hyper-parameters. During training, due to the class imbalance nature of the dataset, we used balanced weights (available in scikit-learn library [194]) to compensate for the problem.

To ensemble the MLP and SVM classifier, we made their prediction on the d_9 . After that, we found a coefficient in $[0, 1]$ that maximizes the balanced accuracy of the linear combination of MLP and SVM probabilities. Finally, the ensemble model was evaluated on d_{10} .

3.4 Experimental results

In this section, the 125 U-Nets were used as a feature extractor for every classification task. After the 10-fold cross-validation, we obtained in total 10 models.

To estimate the model performance on in-domain data, we evaluated 10 ensemble models on their corresponding in-domain test fold. By doing this, each testing sample was evaluated by one model and has one final prediction. We then concatenated all the prediction of 10 folds and compute different metrics based on that prediction.

To estimate the model performance on out-of-domain data, we evaluated 10 ensemble models on the out-of-domain data and averaged the output of these 10 models to boost the model generalization. By doing this, each testing sample was evaluated by ten models and had one final prediction. We then computed different metrics based on that prediction.

3.4.1 Ablation study for binary classification tasks

Table 3.2 describes our ablation study for different binary classification tasks. This is done by evaluating 4 tasks: dementia diagnosis (*i.e.*, AD and FTD *vs.* CN), AD diagnosis (*i.e.*, AD *vs.* CN), FTD diagnosis (*i.e.*, FTD *vs.* CN) and 2-class differential diagnosis (*i.e.*, AD *vs.* FTD). When training our classifiers, the 10 folds are remaining the same but all subjects with irrelevant classes are removed for each classification task. The balanced accuracy is used to assess the model performance, other metrics are also provided in the Appendix B.

Table 3.2: Ablation study of our method for binary classification tasks. We use the balanced accuracy (BACC) to assess the performance. We perform 10-fold cross validation on ADNI+NIFD dataset to estimate the in-domain performance (exp. 1, 2, 3). Additionally, we evaluate on NACC dataset to estimate the out-of-domain performance (exp. 4, 5, 6) by averaging the outputs of 10 trained models. The results are presented in %. **Red**: best result, **Blue**: second result.

| No. | Evaluation | Features | Dementia diagnosis Dem. <i>vs.</i> CN | AD diagnosis AD <i>vs.</i> CN | FTD diagnosis FTD <i>vs.</i> CN | Differential diagnosis AD <i>vs.</i> FTD |
|-----|---------------|----------|---|-------------------------------------|---------------------------------------|--|
| | | | $N = 615$ | $N = 465$ | $N = 466$ | $N = 299$ |
| 1 | In-domain | Volumes | 85.3 | 82.3 | 86.6 | 81.3 |
| 2 | | Grades | 86.3 | 87.1 | 91.0 | 94.3 |
| 3 | | Ensemble | 87.5 | 87.5 | 90.7 | 91.0 |
| | | | $N = 1627$ | $N = 1605$ | $N = 1353$ | $N = 296$ |
| 4 | Out-of-domain | Volumes | 86.6 | 86.7 | 87.0 | 88.9 |
| 5 | | Grades | 86.1 | 83.2 | 88.6 | 84.0 |
| 6 | | Ensemble | 86.9 | 86.8 | 89.1 | 87.1 |

Based on the results, we observe higher balanced accuracy of grade features than volume features for in-domain evaluation (exp. 1 *vs.* 2; ADNI+NIFD datasets) in every binary classification task. However, when evaluating on out-of-domain data (NACC dataset), the volume features are better than grade features (exp. 4 *vs.* 5) in all tasks except FTD diagnosis. Since the ensembling of two models can improve the performance in most of the cases compared to a single model, both grade and volume features are crucial for our classifications. However, they might focus on different characteristics of data (*e.g.*, grade features are more sensitive with FTD and volume features are more sensitive with CN, see Section 3.4.2), making different rankings for in-domain and out-of-domain datasets.

3.4.2 Performance for multi-disease classification

Table 3.3 shows the results obtained for the 3-class differential diagnosis (*i.e.*, AD *vs.* CN *vs.* FTD). Different metrics are used to estimate the model performance: accuracy (ACC), balanced accuracy (BACC), area under curve (AUC) and sensitivity for each class. We observe that the volume features with the SVM classifier provide high CN sensitivity compared to grade features with the MLP classifier for both in-domain and out-of-domain evaluation. Besides, the grade features with MLP classifier provide high FTD sensitivity compared to volume features with SVM classifier for both in-domain and out-of-domain evaluation. These properties are important for the multi-class classification. Consequently, the combination of grade and volume consistently shows the best or second results in various metrics for both in-domain and out-of-domain evaluation. In the following, the results of our ensemble framework is used to compare with state-of-the-art methods.

Table 3.3: Performance of different models for the multi-disease classification CN *vs.* AD *vs.* FTD. We denote ACC for accuracy, BACC for balanced accuracy, AUC for area under curve and Sen. for sensitivity. We perform 10-fold cross validation on ADNI+NIFD dataset to estimate the in-domain performance (exp. 1, 2, 3). Additionally, we evaluate on NACC dataset to estimate the out-of-domain performance (exp. 4, 5, 6) by averaging the outputs of 10 trained models. The results are presented in %. The best and second performances are respectively in red and blue.

| No. | Evaluation | Features | ACC | BACC | AUC | CN Sen. | AD Sen. | FTD Sen. |
|-----|---------------|----------|------|------|------|---------|---------|----------|
| 1 | In-domain | Volumes | 81.3 | 77.2 | 91.5 | 92.4 | 68.5 | 70.7 |
| 2 | | Grades | 85.4 | 84.6 | 93.4 | 87.3 | 84.6 | 82.0 |
| 3 | | Ensemble | 86.0 | 84.7 | 93.8 | 89.6 | 83.2 | 81.3 |
| 4 | Out-of-domain | Volumes | 87.9 | 79.9 | 91.2 | 91.6 | 72.6 | 75.7 |
| 5 | | Grades | 82.7 | 79.2 | 88.8 | 85.2 | 71.3 | 81.1 |
| 6 | | Ensemble | 87.1 | 81.6 | 91.6 | 89.6 | 76.9 | 78.4 |

Moreover, we trained a UMAP [179] with training data (ADNI + NIFD) and visualized the transformed test set (NACC) in 2D space (see Figure 3.2). The transformed data was colored with respect to the diagnosis class. Two types of input were considered: grades (disease coordinate) and volumes (V). Upon observation, we noticed that the blue points (representing FTD) exhibited better clustering when using the grade feature. Conversely, when using the volume feature, the green points (representing CN) demonstrated improved grouping. This finding may provide an explanation as to why grade features exhibit higher sensitivity for FTD, while volume features demonstrate greater sensitivity for identifying CN cases. Finally, the ensemble features seems to better separate the data into 3 classes. Lastly, the ensemble features appear to exhibit the best separation among the three classes. Indeed, the silhouette score [211] of the ensemble is 0.66, greater than

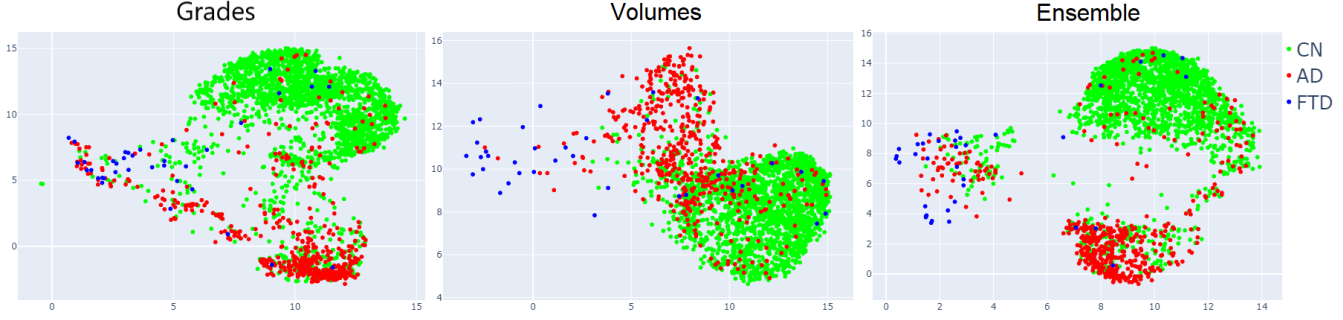


Figure 3.2: UMAP visualization of grades, volumes and the ensemble on out-of-domain data (NACC). The transformed data was colored with respect to the diagnosis class (ground truth).

the grade score of 0.54 and the volume score of 0.40.

3.4.3 Comparison with state-of-the-art methods

In this section, we compare our method with two other deep learning based methods. In the first method, Hu *et al.* used a ResNet-like architecture for classification based on the intensities of a whole MR image [104]. They then used a guided backpropagation based method to visualize the dominant regions of AD and FTD pathologies. In the second method, Ma *et al.* firstly extract structure volume and cortical thickness (Cth) features from an MR image [167]. They then trained a Generative Adversarial Network (GAN) using these features and added an additional class for the fake data. At the inference time, the probability of this class is discarded for the final decision.

We retrained the method of Hu *et al.* with the official publicly available code ³. In the case of the second method, we re-implement it based on the associated paper. For a fair comparison, we use the same 10 folds to train 10 models of each method. To train each model, 7 folds were used for training, 2 folds for validation. The remaining data fold was used to assess the in-domain performance. Finally, we applied the same data preprocessing pipeline used in our proposed method for training the state-of-the-art methods mentioned.

Table 3.4 shows the comparison of our method with state-of-the-art methods for different problems of binary classification. Balanced accuracy (BACC) is used to assess the model performance. Other metrics, such as accuracy and area under curve are provided in the annexes. Our method consistently achieves the best results across all tasks, both in-domain (exp. 1, 2, 3) and out-of-domain diagnosis (exp. 4, 5, 6). This indicates the superior performance and effectiveness of our approach. Furthermore, our method

³https://github.com/BigBug-NJU/FTD_AD_transfer

demonstrates robustness to domain shift, surpassing other methods. On average, the performance drop between out-of-domain and in-domain evaluations for our method is only 1.7%. In comparison, [167] exhibits an average drop of 2.1%, while [104] shows a substantial average drop of 9.3%. Overall, our method demonstrated high performance on different tasks and datasets and is more robust to external validation than other methods, highlighting its generalization capacity on unseen data and, thus, in clinical practice.

Table 3.4: Comparison of our method with current state-of-the-art methods for binary classification tasks. Our reported performances are the average of 10 repetitions and presented in %. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. We denote Dem. for dementia (AD and FTD), CNN for convolutional neural network, GAN for generative adversarial network and Cth for cortical thickness.

| No. | Evaluation | Method | Dementia diagnosis Dem. <i>vs.</i> CN | AD diagnosis AD <i>vs.</i> CN | FTD diagnosis FTD <i>vs.</i> CN | Differential diagnosis AD <i>vs.</i> FTD |
|-----|---------------|------------------------|---|-------------------------------------|---------------------------------------|--|
| | | | | | | |
| 1 | In-domain | Hu <i>et al.</i> [104] | 81.8 | 75.9 | 83.8 | 82.3 |
| 2 | | Ma <i>et al.</i> [167] | 85.1 | 85.3 | 85.7 | 77.9 |
| 3 | | Our method | 87.5 | 87.5 | 90.7 | 91.0 |
| 4 | Out-of-domain | Hu <i>et al.</i> [104] | 81.3 | 76.1 | 68.0 | 61.2 |
| 5 | | Ma <i>et al.</i> [167] | 77.9 | 86.6 | 80.8 | 80.5 |
| 6 | | Our method | 86.9 | 86.8 | 89.1 | 87.1 |

Table 3.5 presents the comparison of our method with the state-of-the-art methods under different metrics: accuracy (ACC), balanced accuracy (BACC), area under curve (AUC) and the sensitivity for each class (*i.e.*, CN, AD and FTD). Our method presents higher performance than other methods in global performance metrics (*i.e.*, ACC, BACC and AUC) for both in-domain and out-of-domain evaluation. Furthermore, our method presents similar performances in all ACC, BACC, AUC metrics, between in-domain and out-of-domain evaluations. This property is not observed in other methods [167, 104]. It shows the high generalization capacity of our framework. In terms of sensitivity, our method achieves most of the time first or second place for all classes (*i.e.*, CN, AD and FTD).

Overall, our framework exhibits high performance and generalization capacity across various tasks, including binary and multi-disease diagnosis. However, it is important to note that there is a trade-off associated. Our framework, consisting of 125 U-Nets and an MLP classifier, comprises 393 million parameters, requires 25.9 TFLOPs for computation, takes 110 hours for training and has an inference time of 1.63 seconds (mainly due to the patch extracting and image reconstructing times). In comparison, the method of [104] presents 46 million parameters, 1 TFLOPs, 6 hours for training and an inference time of 1.4×10^{-3} seconds, while the method of [167] presents 0.11 million parameters, 6.8×10^{-6}

Table 3.5: Comparison of our method with current state-of-the-art methods for 3-class differential diagnosis AD *vs.* FTD *vs.* CN. **Red**: best result, **Blue**: second best result. We denote ACC for accuracy, BACC for balanced accuracy, AUC for area under curve, Sen. for sensitivity, CNN for convolutional neural network, GAN for generative adversarial network and Cth for cortical thickness.

| No. | Evaluation | Method | ACC | BACC | AUC | CN Sen. | AD Sen. | FTD Sen. |
|-----|---------------|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1 | In-domain | Hu <i>et al.</i> [104] | 76.3 | 72.5 | 90.0 | 58.4 | 86.4 | 96.5 |
| 2 | | Ma <i>et al.</i> [167] | 77.1 | 75.9 | 86.4 | 80.4 | 81.2 | 66.0 |
| 3 | | Our method | 86.0 | 84.7 | 93.8 | 89.6 | 83.2 | 81.3 |
| 4 | Out-of-domain | Hu <i>et al.</i> [104] | 85.2 | 68.8 | 86.5 | 68.0 | 94.1 | 48.6 |
| 5 | | Ma <i>et al.</i> [167] | 69.1 | 74.6 | 87.5 | 66.1 | 82.1 | 75.7 |
| 6 | | Our method | 87.1 | 81.6 | 91.6 | 89.6 | 76.9 | 78.4 |

TFLOPs, 0.4 hours for training and an inference time of 0.4×10^{-3} seconds.

3.4.4 Interpretation of deep grading map

To assess the interpretability provided by the grading map, we compute the averaged DC points (133 points for 133 brain structures) over subjects from each class. The considered subjects are taken from in-domain dataset. The averaged DC maps are transformed into grading maps for visualization. Figure 3.3 shows sagittal and coronal views of these grading maps.

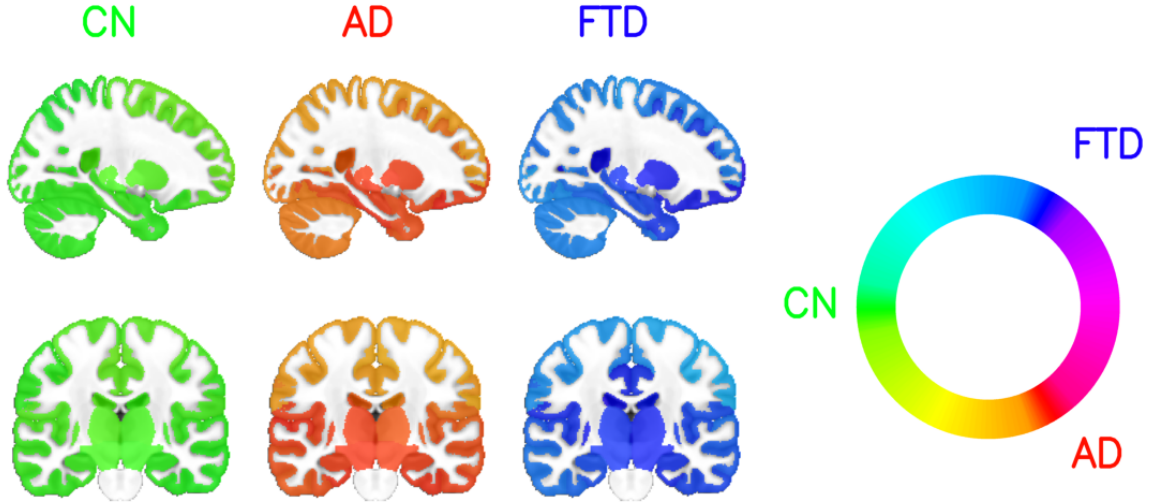


Figure 3.3: Average grading map per group of subjects in the MNI152 space with neurological orientation (with the right of the patient at the right).

First, we can observe that our framework produces average grading maps well-separated for each class. As expected for the group of healthy people (*i.e.*, CN), all

regions are detected as normal. For AD patients, the regions around the hippocampus are detected as AD-related patterns (red color). More generally, the temporal lobe is detected as strongly related to AD-like patterns in this population. The prevalence of AD in this region is widely documented [215]. For the FTD class, we observe that FTD-like anatomical patterns are detected in similar areas. These results indicated that our method found diseases-specific anatomical anomalies (dissimilar patterns between AD and FTD) in similar locations for AD and FTD. This experiment highlights the need of grading map based on the multi-channel disease’s coordinates. To further analyze our grading map, we compute the averaged map for each of its variants (*i.e.*, bvFTD, PNFA, SV) (see Figure 3.4).

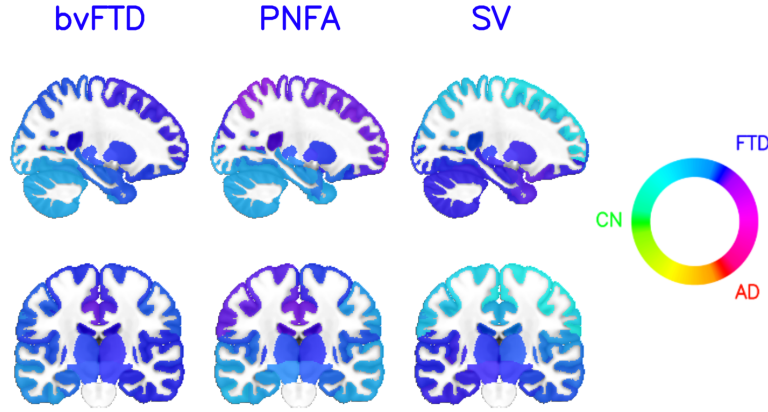


Figure 3.4: Average grading map per variant of FTD in the MNI152 space with neurological orientation (with the right of the patient at the right).

We observe that the three sub-types present different FTD-related patterns. In the bvFTD group, the grading map highlights the frontal and temporal areas which are shown to be related to this pathology [251]. In the PNFA group, the left frontal region [29] and especially the left inferior frontal gyrus [123] are highlighted which is typical of this syndrome. For the SV group, the left temporal pole is the most affected brain region. Indeed, this area presents typical atrophy in SV patients [123]. We remark with the 3 variants of FTD that the disease severity is asymmetric, which is in line with the finding of Boeve *et al.* [29].

Finally, we select typical deep grading maps of each class (*i.e.*, CN, AD and FTD) at different ages (see Figure 3.5). We observe that in older healthy people, some areas have similar deep grading patterns with FTD [42] and AD [233]. In AD and FTD patients, both diseases start at a specific region (around the hippocampus for AD and frontotemporal lobes for FTD) and tend to expand to the whole brain over time.

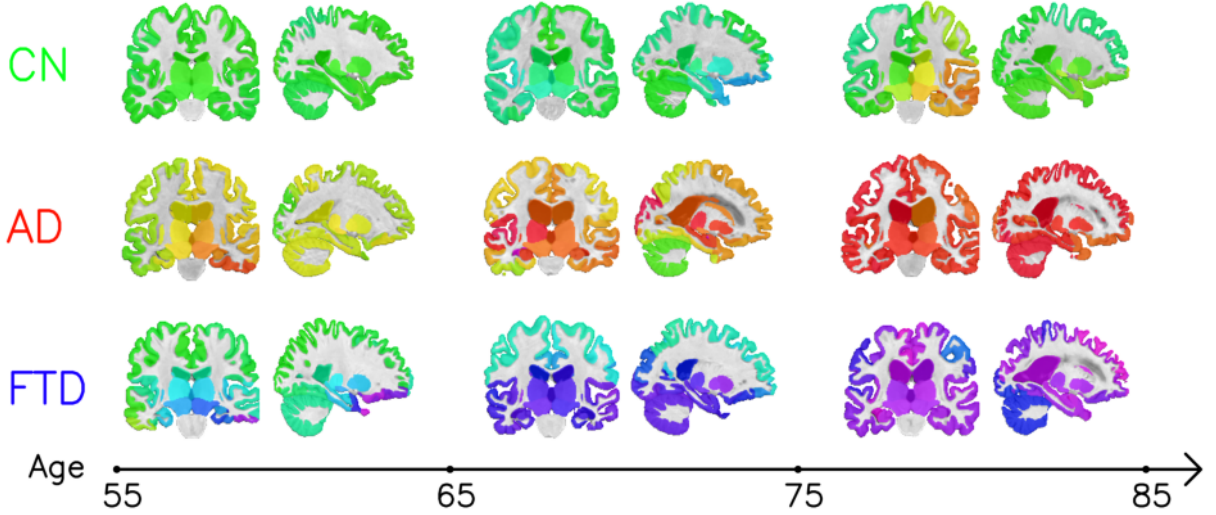


Figure 3.5: Individual grading maps of each group (CN, AD and FTD) of subjects with respect to age.

3.5 Discussion

In this chapter, we proposed a novel deep grading framework dedicated to multi-disease classification problems. Moreover, we aimed at expanding our knowledge on AD and FTD disease-related patterns. So, beyond the predicted class for each individual, we provided also the color map indicating regions with specific disease patterns. First, we observed that our method find similar affected areas of the two diseases but dissimilar patterns. Second, the regions highlighted in each group of people (*i.e.*, CN, AD, bvFTD, PNFA, SV) as well as the asymmetric characterization provided by our framework are coherent with current knowledge of these diseases in the literature. Finally, we further investigate the three variants of FTD to describe the variability of this disease as suggested by Hu *et al.* [104]. This is expected to help clinicians to deeper understand FTD and to make more accurate diagnoses.

In this study, we take advantage of two types of biomarkers: structure grading and structure atrophy. While structure grading features provided by several U-Nets might offer information about anatomical patterns similarity with each class (*i.e.*, CN, AD, FTD), structure atrophy offers information about the abnormality of each brain structure in terms of size. Table 3.3 demonstrates that the first biomarker can help to better detect FTD patients and the second one can accurately identify healthy people (*i.e.*, CN). As a result, our ensemble model improves the model performance not only in multi-disease tasks but also in many binary classification tasks (see Table 3.2).

This study is one among a few studies addressing the problem of multi-disease classification using sMRI data [33, 132, 167, 104]. We tried our best effort to make a fair

comparison with state-of-the-art methods. Compared to these approaches, our method shows promising performance on different classification tasks (*i.e.*, dementia *vs.* CN, AD *vs.* CN, FTD *vs.* CN, AD *vs.* FTD and CN *vs.* AD *vs.* FTD). Experimental results demonstrate that our method is not only good with in-domain dataset but the learned patterns are generalizable, expressed by the lower drop of performance when evaluating on an out-of-domain dataset compared to other state-of-the-art methods. This characterization is shown in both binary classification and multi-disease classification tasks. This is very important in clinical practice where data are heterogeneous.

It is noteworthy that that we utilized the same (preprocessed) data to train our method and the state-of-the-art methods we compared to in this study. This choice was made based on the observation that the performance achieved by these methods using these preprocessed data was better than that achieved using raw data. Therefore, our preprocessing pipeline played a crucial role in enhancing the in-domain performance of both our method and the state-of-the-art methods while also contributing to improved generalization capacity on out-of-domain data.

Besides, the training data is an important factor leading to a good classification model. For instance, we used data coming from two different datasets with different classes: ADNI contains CN and AD patients while NIFD contains CN and FTD patients. These two datasets are chosen for their popularity and a lack of datasets with sufficient subjects for each class: CN, AD and FTD. However, it may exist some dataset-side biases. To alleviate the problem, only 3 Tesla images are selected as in [104]. It is possible that some people are misdiagnosed in these databases, where biological biomarkers are not always available, making a noisy ground-truth. Future works should consider the outlier removal to further improve model reliability. Finally, this study relies only on the sMRI at the baseline with the goal to detect brain diseases as early as possible. However, the patient’s condition changes over time, it could be beneficial to use longitudinal data to make more accurate predictions and further track the progression of the disease.

Chapter 4

Brain structure ages for differential diagnosis

| | | |
|-------|--|----|
| 4.1 | Introduction | 73 |
| 4.2 | Materials | 76 |
| 4.2.1 | Chronological age prediction | 76 |
| 4.2.2 | Multiple pathologies classification | 76 |
| 4.3 | Methods | 77 |
| 4.3.1 | Method overview | 77 |
| 4.3.2 | Implementation details | 79 |
| 4.3.3 | Validation Framework | 81 |
| 4.4 | Experimental results | 81 |
| 4.4.1 | Chronological age estimation | 81 |
| 4.4.2 | Disease classification | 83 |
| 4.4.3 | Predicted brain age of different populations | 86 |
| 4.4.4 | Interpretation of brain structure age gap estimation | 86 |
| 4.5 | Discussion | 87 |
| 5.1 | Introduction | 91 |
| 5.2 | Materials | 92 |
| 5.3 | Method description | 93 |

| | | |
|-------|--|-----|
| 5.3.1 | Model overview | 93 |
| 5.3.2 | Data augmentation | 96 |
| 5.3.3 | Validation framework and ensembling | 96 |
| 5.3.4 | Implementation details | 96 |
| 5.4 | Experimental results | 97 |
| 5.4.1 | Ablation study | 97 |
| 5.4.2 | Comparison with state-of-the-art methods | 98 |
| 5.4.3 | Visualization of deformable patch location | 99 |
| 5.5 | Comparison between the proposed Transformer and CNNs | 100 |
| 5.5.1 | Performance comparison | 100 |
| 5.5.2 | Comparison of Interpretability and Explainability | 103 |
| 5.6 | Discussion | 105 |
| 6.1 | Conclusion | 109 |
| 6.2 | Future works and perspectives | 111 |

The method presenting in this chapter is related to the following publications and softwares:

- [3] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “Brain Structure Ages - A new biomarker for multi-disease classification”. In: *HAL preprint* (2023). Under revision in *Human Brain Mapping*. eprint: [hal-04080401](#).
 - [9] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *BrainDiseaseDiagnosis*. Under deposit.
 - [10] Pierrick Coupé, José Vicente Manjón, **Huy-Dung Nguyen**, Boris Mansencal and Michaël Clément. *BrainStructureAges*. Under deposit.
-

In the previous chapter, the MCDG biomarker can effectively tackle the challenges of multi-disease differential diagnosis while maintaining a high level of interpretability. However, when dealing with more than two diseases, where the total number of classes, denoted as n , is three or greater, each class is represented as a vector in an $n-1$ dimensional space. Consequently, as n increases, the readability of the visualization decreases. In this chapter, we will present an alternative approach to the MCDG framework for multi-disease differential diagnosis, ensuring interpretability even in scenarios involving an arbitrary number of diseases. To demonstrate the feasibility of this approach, we have chosen the proof-of-concept differential diagnosis of CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ.

4.1 Introduction

In the medical field, chronological age is widely used as an indicator to describe people. It depicts a reference curve that healthy organs should follow. The deviation from that reference may be associated with different factors such as the interaction of genes, environment, lifestyle and diseases [78]. To measure this deviation, the concept of biological age (BA) has been created. It is an estimation of individual’s age based on various advanced strategies [38, 203, 188] and is expected to be able to take into account all the factors mentioned above. Consequently, an accelerated (or delayed) aging process results in a higher (or lower) value of BA with respect to the chronological age.

The analysis of BA can be associated with a whole-body system or a specific organ. On the one hand, the whole-body evaluation approaches typically use non-imaging data (*e.g.*, DNA methylation patterns [39], protein [112]) and are unable to account for the variations in aging between individual organs [15]. Such global information might be difficult to use in clinical practice. On the other hand, imaging studies of BA dedicated to a particular organ may provide important details about that organ’s condition, and the brain is one of the most commonly studied organs. Brain structure changes are demonstrated to be mutually caused by the natural aging process and neurodegenerative diseases [197, 108, 119, 232, 52, 54]. Cole *et al.* demonstrated that biological brain age can enable the development of treatment plans and a better understanding of disease processes [46]. The authors emphasized that the difference between the predicted brain age and the chronological age is a valuable bio-marker since it shows a correlation with aging as well as with diseases. This difference is denoted as BrainAGE for Brain Age Gap Estimation. Since its introduction, this new bio-marker has been widely used in many studies to analyze various diseases [78]. Generally, a model is trained with brain images from a healthy population and then used to estimate the age of patients with diseases.

In BrainAGE, sMRI is the most used modality (about 88% of studies [183]). It has been shown that reasonable prediction error can be achieved using this modality. Moreover, sMRI is commonly available in medical environments [183]. Initially, sMRI was used with some traditional machine learning algorithms such as relevance vector regression [80], support vector regression [153] and Gaussian process regression [44] to perform BrainAGE. The prediction error of these methods ranges from 4.29 to 5.02 years for the mean absolute error (MAE) metric. Since the success of deep learning in many natural image processing applications, it has also become a useful technique in various medical imaging studies. Recent studies show the capacity of deep learning algorithms in the brain age estimation task based on sMRI with an MAE ranging from 1.96 to 4.16 years [15, 28, 43, 125, 25]. These promising results suggest using deep learning to estimate brain age for further analysis.

These deep learning based methods adapt famous CNN architectures to estimate the brain age. When employing a VGG-like architecture, Ueda *et al.* demonstrated that using 3D CNN can lead to better accuracy than 2D CNN for age prediction [240]. In another work, Cole *et al.* also used a VGG-like architecture and found that the grey matter extracted from 3D sMRI is better than white matter and raw image for age prediction [46]. Using a similar architecture, Bermudez *et al.* suggested to additionally take advantage of brain structure volume to improve the model performance [25]. Bintsi *et al.* employed ResNet architecture to predict age on several sub-volumes of brain image [28]. The final prediction was aggregated using a linear regression model. Armanious *et al.* proposed to use the inception module with squeeze-and-excitation module to accurately predict

healthy brain age [15]. Bashyam *et al.* customized the inception-resnetv2 to build their model and trained it on 11729 healthy subjects.

After training a brain age prediction model, the next step is to apply it to a population of interest to compare healthy and diseased groups (*i.e.*, analysis at population level). For example, Franke *et al.* analyzed the brain maturation during childhood and adolescence [79]. By applying a trained model on subjects being born before the 28th and after the 29th week of gestation, they found that the BrainAGE of the first group was significantly lower than the second group, showing a delayed structural brain maturation of the first group. Applying the same technique, Koutsouleris *et al.* demonstrated an accelerated aging of 5.5 years in schizophrenia and 4.0 years in major depression patients compared to normal aging [138]. In another study dedicated to AD, the BrainAGE was estimated about +10 years in AD patients, implying accelerated aging of this population [80].

Although the BrainAGE can provide a description of a specific population, its application in individual diagnosis is still limited. Only a few works suggested performing disease detection or differential diagnosis using BrainAGE at subject level. For instance, the BrainAGE was used as a biomarker to perform differential diagnosis between mild cognitive impairment and AD in [83] and to diagnose AD (*i.e.*, AD patients *vs.* healthy controls) in [77, 243]. More recently, Cheng *et al.* used deep learning to accurately predict brain age and they use BrainAGE as the only feature for various binary diagnosis tasks (*i.e.*, diseased subjects *vs.* healthy subjects) [41]. Although encouraging results were obtained, these works performed only binary classification tasks but not multi-class classification. The reason for this may be due to the coarse description of brain’s state provided by the global BrainAGE. Indeed, BrainAGE can only describe the aging process of the whole brain but does not provide any details about brain structures’ state. Therefore, it is difficult to use BrainAGE for involved tasks such as the differential diagnosis of multiple pathologies.

In this study, we propose to extend the notion of the global brain age to local brain structure ages. Our main hypothesis is that the aging process is heterogeneous over the brain and specifically, different brain structures may present different ages. Consequently, we first estimate the brain age at the voxel level. This results in a 3D aging map of voxelwise brain ages. By averaging predicted brain ages by brain structure, we obtain the Brain Structure Ages, denoted as BSA. This local BSA is expected to provide more information about the subject’s condition than a global age prediction of a whole subject’s brain. As shown later, this novel biomarker can be used as input of an MLP to accurately estimate the subject’s age. During validation, our framework showed competitive results compared to state-of-the-art methods. Furthermore, the difference between BSA and the subject’s chronological age, denoted as BSAGE for Brain Structure Age Gap Estimation, can be also used with an SVM for multi-class classification (*i.e.*, CN *vs.* AD *vs.* FTD *vs.* MS

vs. PD vs. SZ). In our experiments, we demonstrated the important gain of using BSAGE compared to BrainAGE for the multi-disease classification task. Finally, by projecting the BSAGE on a brain atlas, we can visually observe the brain regions affected by different diseases.

4.2 Materials

The data used in this study comprise 39255 images from various datasets: ABIDE, ADNI, AIBL, ICBM, IXI, NDAR, OASIS, C-MIND, UKBioBank, SRPBS, COBRE, CamCAN, PPMI, NIFD, OFSEP, NACC, DLBS, MIRIAD and BrainGluSchi (see Section 1.1.1 for more details). All the T1 weighted images at the baseline were used.

4.2.1 Chronological age prediction

Among available data, 32718 images were used to study the accuracy of our chronological age predictor. First, eight datasets including 2887 images (*i.e.*, ABIDE I, ADNI, AIBL, ICBM, C-MIND, IXI, NDAR, OASIS1) were used in training/validation. Second, two external datasets (*i.e.*, out-of-domain) were used for testing. Concretely, CN subjects of ABIDE II (*i.e.*, 580 images) were used to estimate the model accuracy on a young population and CN from UKBioBank (*i.e.*, 29251 images) were used to estimate the model accuracy on an older population (see Table 4.1). For ABIDE, we ensured that no subject in phase I was presented in phase II.

4.2.2 Multiple pathologies classification

Besides, we assessed the classification performance using BSAGE on 6537 images composed of 6 classes (*i.e.*, CN, AD, FTD, MS, PD and SZ). Eight datasets including 1992 images (ADNI, AIBL, SRPBS, COBRE, CamCAN, PPMI phase 1, NIFD and OFSEP centers 1-2) were used to perform a 10-fold cross validation (in-domain validation) (see Table 4.2). Then, we constructed an out-of-domain dataset including 4545 images using seven cohorts (*i.e.*, NACC, DLBS, MIRIAD, OASIS3, BrainGluShi, PPMI phase 2 and OFSEP-other-centers) to assess the generalization capacity of such models. For the OFSEP, we used the acquisition sites to split this global dataset into two non-overlapping domains. For PPMI, we ensured that no subject in phase I was presented in phase II.

Table 4.1: On top, summary of participants used for training age predictor. On bottom, description of the external datasets used for testing.

| Usage | Dataset | Male/Female | Age (Mean \pm Std) |
|--------------------------|-----------|-------------|----------------------|
| Age prediction training | ABIDE I | 408/84 | 17.5 \pm 7.8 |
| | ADNI | 201/203 | 74.8 \pm 5.8 |
| | AIBL | 112/120 | 72.3 \pm 6.7 |
| | ICBM | 112/182 | 33.7 \pm 14.3 |
| | C-MIND | 107/129 | 8.4 \pm 4.3 |
| | IXI | 242/307 | 48.8 \pm 16.5 |
| | NDAR | 208/174 | 12.4 \pm 6.0 |
| | OASIS1 | 111/187 | 45.3 \pm 23.8 |
| Young population testing | ABIDE II | 403/177 | 14.8 \pm 9.3 |
| Older population testing | UKBioBank | 14917/14334 | 64.2 \pm 7.9 |
| Total | | 16821/15897 | 14.8 \pm 9.3 |

4.3 Methods

4.3.1 Method overview

Figure 4.1 provides an overview of our method. First, we estimate the brain ages map at voxel level from a preprocessed T1 image using a large number of U-Nets. Then, this 3D map is used with a segmentation mask to compute the BSA features (Section 4.3.1). Finally, the BSA features can be employed to estimate the chronological age using a MLP model or combined with brain structure volumes to perform multi-disease classification using an SVM classifier (Section 4.3.1).

Brain structure age estimation

In order to produce the 3D aging map, we extracted $m = k^3$ overlapping 3D sub-volumes of the same size for each T1w MRI. Next, we trained m U-Nets to predict age at voxel level with these m 3D sub-volumes. The goal of this training strategy is dual. First, as the size of a sub-volume is relatively small compared to the original image, it can be trained with a lighter weight model and thus, require only a low computation capacity. Second, we limit the receptive field of each model to a local brain region in order to force

Table 4.2: Number of participants (Male/Female) used for multi-class classification.

| Usage | Dataset | CN | AD | FTD | MS | PD | SZ |
|--|---------------------------|---------|---------|--------|----------|---------|--------|
| 10-fold cross validation training | ADNI | | 181/150 | | | | |
| | AIBL | | 18/28 | | | | |
| | SRPBS | 88/60 | | | | | 84/58 |
| | COBRE | 11/7 | | | | | 54/14 |
| | CamCAN | 75/85 | | | | | |
| | PPMI phase 1 | 35/13 | | | | 228/131 | |
| | NIFD | 15/15 | | 87/56 | | | |
| | OFSEP centers 1-2 | | | | 161/338 | | |
| Out-of -domain testing | NACC | 47/104 | 318/419 | 22/23 | | | |
| | DLBS | 117/196 | | | | | |
| | MIRIAD | 12/11 | 19/27 | | | | |
| | OASIS3 | 270/385 | 46/46 | | | | |
| | Brain- GluSchi | 61/25 | | | | | 71/11 |
| | PPMI phase 2 | | | | | 74/58 | |
| | OFSEP other centers | | | | 585/1598 | | |
| | Total | 731/901 | 582/670 | 109/79 | 746/1936 | 302/189 | 209/83 |

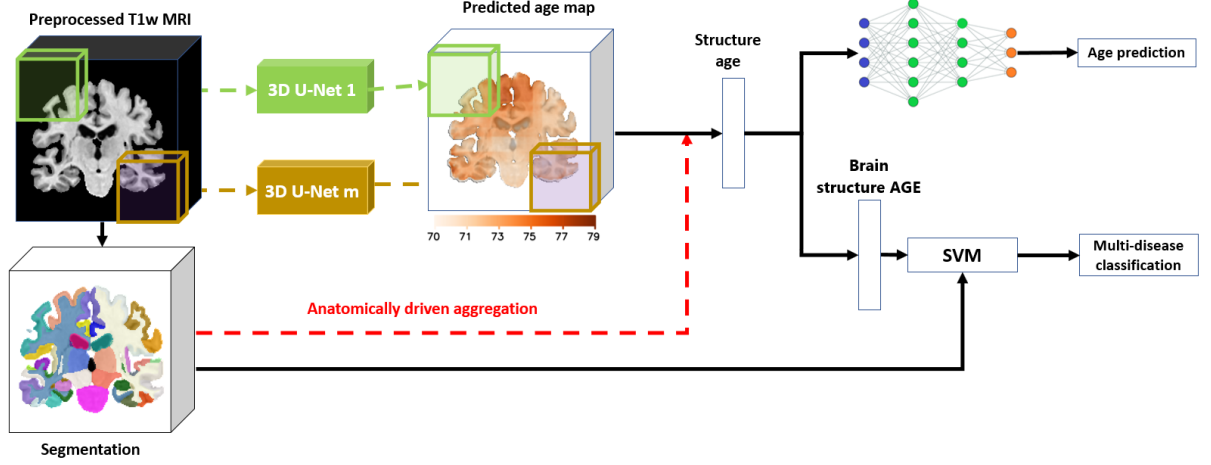


Figure 4.1: An overview of the proposed chronological age prediction and multi-disease classification. The T1w image, its segmentation and the age map are taken from a 71 years old healthy person.

it to locally describe the brain age. The outputs were then used to reconstruct a 3D brain age map. Finally, the BSA was computed with the help of an AssemblyNet-based brain segmentation [50]. In practice, we estimated the mean value of voxel-wise age estimation for each structure segmentation.

Application to chronological age prediction and multi-disease classification

To demonstrate different use cases of the BSA, we performed two experiments using this biomarker: chronological age prediction which can help to briefly describe a population and multi-disease classification which can guide clinicians to focus on certain pathologies.

To predict the chronological age of healthy people, we employed a classical MLP and used the predicted BSA as its input. For the multi-disease classification, we first computed the BSAGE (*i.e.*, the difference between BSA and the subject’s chronological age) and then used it as input of an SVM classifier to address the 6-class problem CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ. Moreover, structure volume is used as additional feature of BSAGE for SVM-based classification.

4.3.2 Implementation details

First, a preprocessed T1w MRI in the MNI space of size $181 \times 217 \times 181$ voxels at $1mm^3$ was downsampled with a factor of 2 to the size of $91 \times 109 \times 91$ voxels. After that, we extract k^3 (*i.e.*, $k = 5$) overlapping sub-volumes of the same size $32 \times 48 \times 32$ voxels and

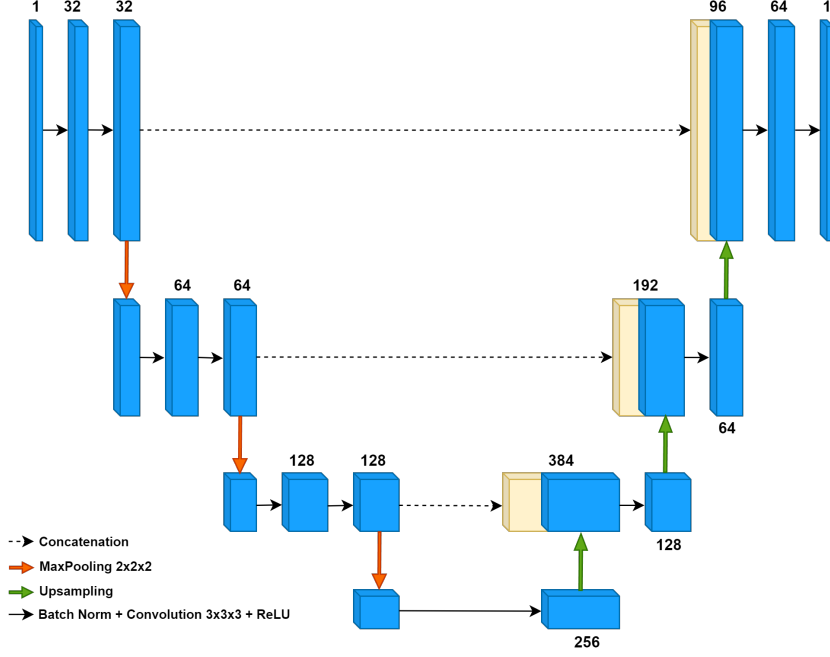


Figure 4.2: Architecture of an unit U-Net used for voxel-level age prediction. The number above each block is the number of channel.

evenly distributed along the 3 image’s dimensions from the downscale image. We trained $m = k^3$ (*i.e.*, $m = 125$) U-Nets to predict age at voxel level with these m sub-volumes. Figure 4.2 shows the architecture of our unit U-Net used for voxel-level age prediction. The m outputs were then used to reconstruct a 3D age map of size $91 \times 109 \times 91$ voxels. Of note, the predicted brain age located at overlapping voxel positions of more than 1 sub-volume was averaged. The reconstructed image was upsampled using trilinear interpolation to the same spatial size as the original input. This 3D map was used to compute BSA and then BSAGE features.

To train the 125 U-Nets, we use the MAE as loss function and SGD optimizer. The batch size is set to 8 and the training was terminated after 20 epochs without any improvement on validation loss. The first U-Net was trained from scratch and other U-Nets were trained with transfer learning from their adjacent U-Net (see [50] for more details). The training data is split into training/validation sets with a ratio of 80%/20% (see Table 4.1). In addition, when a new U-Net was trained, the training and validation data were gathered and re-split to exploit the maximum information from available data. Finally, we employed different data augmentation techniques to alleviate the overfitting problem. Concretely, we randomly shifted a patch by $t \in \{-1, 0, 1\}$ voxel in each dimension (denoted as random shift technique) and then applied mixup data augmentation [261].

Before using BSA features, we applied an age correction technique for each of their elements. We followed a simple method of Smith *et al.* [223] to eliminate bias in each

structure brain age. Concretely, we denoted the real age as A_r (an $N_{subjects} \times 1$ vector), the brain age as A_b , a brain structure age as A_s (an $N_{subjects} \times 1$ vector) and the bias term δ . So, we predicted A_b from A_s : $A_b = A_r + \delta = A_s\beta$. This is equivalent to $A_r = A_s\beta - \delta$. This regression can be solved with $\beta = (A_s^T A_s)^{-1} A_s^T A_r$. Finally, $A_b = A_s(A_s^T A_s)^{-1} A_s^T A_r$.

The MLP used to estimate chronological age is composed of 4 layers with respectively $s \times 4$, $s \times 2$, s and 1 neuron. To train the MLP, we used the MAE as loss function and Adam optimizer. The batch size is set to 8 and the training was terminated after 50 epochs without any improvement on validation loss.

When training the SVM for multi-disease classification, three kernels were used to select the best model through our cross-validation: linear, polynomial and radial basis function. One hundred values of C in the log-space $[-1.5; 0.5]$ were used in the hyper-parameter search. We performed a grid-search for the kernel and the hyper-parameter C.

4.3.3 Validation Framework

For the chronological age prediction, we compute the BSA features of U-Nets' training subjects. This data is used to train the MLP-based regression (10-fold cross-validation). This results in 10 MLP models. At testing time, the outputs of the 10 MLP models were averaged to make the final prediction. We used two separate out-of-domain datasets to assess our method accuracy (see Table 4.1).

For the multi-disease classification, we also performed a 10-fold cross-validation to train the SVM classifier (see Table 4.2). We denote the classification performance on this dataset as in-domain performance. In addition, we assessed the generalization capacity on an out-of-domain dataset (see Table 4.2). Similar to the chronological age prediction problem, the outputs of 10 models were averaged to get the final prediction on these external dataset.

4.4 Experimental results

4.4.1 Chronological age estimation

Ablation study

In this part, we aim at studying different factors influencing the model performance: Data amount, different augmentation strategies (*e.g.*, random shift, mixup [261]) and an

age correction technique (see Section 4.3.2). Table 4.3 shows the comparison results.

First, we can observe that increasing the data amount (exp. 1, 2, 3) consistently improve the model accuracy on both young and old population in all metrics (*i.e.*, MAE and R^2). Second, applying different data augmentation techniques (*i.e.*, random shift, mixup in exp. 4, 5) is important in the application of age prediction. This is in line with the finding of [195]. Finally, many studies have shown the advantages of using age correction techniques. In our case, the implemented technique only slightly improves the result (exp. 6). However, this technique can enhance the discriminative capacity of BSA for better disease classification (see Section 4.4.2). Overall, each factor contributes to our model accuracy. In the rest of the paper, BSA is computed using 100% data, random shift, mixup and structural age correction techniques unless otherwise specified.

Table 4.3: Ablation study for the chronological age estimation. **Red**: best result, **Blue**: second best result. Text or symbols in black: Changes compared to the previous experiment. Text or symbols in gray: No change compared to the previous experiment. The model performance is estimated by different metrics: Mean absolute error (MAE) and the coefficient of determination (R^2).

| No. | Data amount | Random Shift | Mixup | Structure Age Correction | Young population | | Old population | |
|-----|-------------|--------------|-------|--------------------------|------------------|---------|----------------|---------|
| | | | | | MAE ▼ | R^2 ▲ | MAE ▼ | R^2 ▲ |
| 1 | 50% | ✗ | ✗ | ✗ | 4.65 | 0.30 | 8.32 | -0.89 |
| 2 | 75% | ✗ | ✗ | ✗ | 3.44 | 0.64 | 7.72 | -0.62 |
| 3 | 100% | ✗ | ✗ | ✗ | 2.38 | 0.85 | 4.60 | 0.45 |
| 4 | 100% | ✓ | ✗ | ✗ | 2.11 | 0.89 | 3.98 | 0.58 |
| 5 | 100% | ✓ | ✓ | ✗ | 1.92 | 0.91 | 3.89 | 0.60 |
| 6 | 100% | ✓ | ✓ | ✓ | 1.91 | 0.91 | 3.87 | 0.61 |

Comparison with state-of-the-art methods

In this part, we compare our method with different state-of-the-art methods. For each method below, we used the code available ^{1 2} and retrained the model using the same data split as in our training process. The first method by Jonsson *et al.* uses a ResNet-like architecture and demonstrated promising results in age prediction [125]. More recently, Peng *et al.* presented a lightweight architecture named Simple Fully Convolutional Net-

¹https://github.com/ha-ha-ha-han/UKBiobank_deep_pretrain

²<https://github.com/benniatli/BrainAgePredictionResNet>

work (SFCN) for this problem [195]. They considered age prediction as a classification problem. To introduce a relationship between close classes, they used a soft label during training. The soft label is a probability distribution centered around the ground-truth age. In another work, Leonardsen *et al.* reused the SFCN backbone and demonstrated that the soft label can lead to better accuracy with in-domain data but the regression version presents a better generalization capacity on out-of-domain data [144]. Table 4.4 shows the results of the comparison. For the young population, we can remark that our method presents a very low MAE (1.91 years) and very high R^2 (0.91) compared to other state-of-the-art methods. For the older population, all methods present a drop in performance. In this case, our method shows $MAE = 3.87$ years and $R^2 = 0.61$, presenting the best prediction error over all methods.

Table 4.4: Comparison with state-of-the-art methods for the age estimation task. **Red**: best result, **Blue**: second best result. The model performance is estimated by different metric: Mean absolute error (MAE) and the coefficient of determination (R^2). The results are the average accuracy of 10-fold cross validation. The age for each population is under the form: mean \pm std.

| No. | Method | Young population Age: 14.8 \pm 9.3 | | Older population Age: 64.2 \pm 7.9 | |
|-----|-----------------------|---|------------------------|---|------------------------|
| | | MAE \blacktriangledown | R^2 \blacktriangle | MAE \blacktriangledown | R^2 \blacktriangle |
| 1 | ResNet-like [125] | 2.86 | 0.71 | 4.14 | 0.54 |
| 2 | SFCN soft label [195] | 2.78 | 0.71 | 5.12 | 0.32 |
| 3 | SFCN regression [144] | 2.87 | 0.69 | 4.88 | 0.40 |
| 4 | Our method | 1.91 | 0.91 | 3.87 | 0.61 |

4.4.2 Disease classification

Ablation study for binary classification tasks

In this part, we aim at assessing the BSAGE (*i.e.*, the difference between BSA and the chronological subject’s age) feature in the context of specific disease detection (binary classification). To do it, we compare this feature with the brain structure volume feature (denoted as V). We denote $BSAGE_{nc}$ as the BSAGE without age correction (see Section 4.3.2). Finally, we propose to take advantage of both BSAGE and structure volume biomarker to improve the discriminative capacity of our model.

Table 4.5 shows the results of the comparison between different features for different classification problems. The balanced accuracy (BACC) is presented. Other metrics are

Table 4.5: Ablation study for binary classification tasks. **Red**: best result, **Blue**: second best result. The balanced accuracy (BACC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. We denote BSAGE_{nc}, BSAGE and V for respectively BSAGE with no age correction, BSAGE with age correction and structure volume.

| | No. | Features | AD <i>vs.</i> CN | FTD <i>vs.</i> CN | MS <i>vs.</i> CN | PD <i>vs.</i> CN | SZ <i>vs.</i> CN |
|---------------|-----|---------------------|------------------|-------------------|------------------|------------------|------------------|
| In-domain | | | $N = 781$ | $N = 547$ | $N = 903$ | $N = 763$ | $N = 614$ |
| | 1 | BSAGE _{nc} | 76.3 | 71.4 | 70.2 | 71.8 | 63.9 |
| | 2 | BSAGE | 88.2 | 86.3 | 83.7 | 73.5 | 77.3 |
| | 3 | V | 89.1 | 89.4 | 79.4 | 64.8 | 78.2 |
| | 4 | BSAGE + V | 91.8 | 91.3 | 84.6 | 65.7 | 81.0 |
| Out-of-domain | | | $N = 2103$ | $N = 1273$ | $N = 3411$ | $N = 1360$ | $N = 1310$ |
| | 5 | BSAGE _{nc} | 62.3 | 63.6 | 79.3 | 63.3 | 73.1 |
| | 6 | BSAGE | 78.5 | 90.6 | 84.3 | 52.8 | 69.0 |
| | 7 | V | 86.3 | 90.1 | 71.1 | 58.3 | 76.6 |
| | 8 | BSAGE + V | 86.0 | 91.0 | 83.0 | 59.8 | 83.2 |

provided in the appendix. First, we can remark that BSAGE (exp. 2, 6) is better than the non corrected version BSAGE_{nc} (exp. 1, 5) in most classification problems with a large margin (*i.e.*, AD, FTD and MS detection). Only in case of PD detection, the version without age correction is better than the corrected version in both in-domain and out-of-domain dataset. Second, we observe that BSAGE (exp. 2, 6) is better than structure volume (exp. 3, 7) in MS detection while the structure volume is better in AD detection and SZ detection. In other cases (*i.e.*, FTD and PD detection), one feature is better than the other one on in-domain data and worse on out-of-domain data. From this observation, both the apparent brain structure ages and the structure volumes demonstrate discriminative power for different disease detection tasks. Thus, it should be beneficial to combine them for a better discriminative capacity. As a result, the combination of BSAGE and structure volume (exp. 4, 8) shows most of the time the best or the second best performance.

Multi-disease classification

Table 4.6 shows the results for the multi-disease classification problem. We estimated the balanced accuracy (BACC), accuracy (ACC) and area under curve (AUC) of our model. We performed classification using the true age (exp. 1, 6) and the predicted subject’s age (exp. 2, 7) to confirm that estimating brain age at structure level provides better results than using a global age estimation with real or estimated values. Moreover, these baseline methods enable to estimate the biases present between populations in terms

of age. Indeed, there was some bias in age distribution between diseases since for instance the SZ patients were young while compared to the AD patients. Thanks to this analysis, we can observe that BSAGE (exp. 3, 8) and V feature (exp. 4, 9) presents a far higher performance than the true age and the predicted subject’s age. This suggests that the structure-related information is valuable in classification context. Besides, although the BSAGE (exp. 3, 8) presents lower performance than the V feature (exp. 4, 9), the two biomarkers can be mutually used to achieve better classification performance. Indeed, their combination (exp. 5, 10) shows the best performance for all proposed metrics.

Table 4.6: Multi-disease classification results for CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ. **Red**: best result, **Blue**: second best result. The results are the average accuracy of 10 repetitions and presented in percentage. We denote BSAGE and V for BSAGE with correction and structure volume.

| | No. | Features | BACC | ACC | AUC |
|---------------|-----|---------------|-------------|-------------|-------------|
| In-domain | 1 | True age | 36.8 | 46.4 | 76.8 |
| | 2 | Predicted age | 32.8 | 39.9 | 74.9 |
| | 3 | BSAGE | 58.5 | 61.7 | 88.0 |
| | 4 | V | 64.5 | 65.1 | 90.2 |
| | 5 | BSAGE + V | 68.7 | 69.6 | 93.2 |
| Out-of-domain | 6 | True age | 34.0 | 51.4 | 74.3 |
| | 7 | Predicted age | 31.8 | 42.8 | 72.8 |
| | 8 | BSAGE | 44.7 | 57.0 | 82.6 |
| | 9 | V | 58.7 | 59.4 | 86.8 |
| | 10 | BSAGE + V | 63.3 | 66.1 | 90.6 |

Moreover, we trained a UMAP [179] with in-domain data and visualized the transformed out-of-domain data in 2D space (see Figure 4.3). The transformed data was colored with respect to the diagnosis class. Two types of input were considered: BSAGE and volume (V). Upon observation, we noticed that in case of BSAGE, the MS class (yellow points) is well separated from AD (red points) and CN (green points) classes. In case of volume features, the AD (red points) class is well separated from CN (green points) and MS (yellow points). Moreover, the FTD (blue points) class is better clustered with this type of features. In the combination of BSAGE and volume features, the CN (green points), AD (red points), MS (yellow points) and FTD (blue points) classes are well separated. Furthermore, SZ (pink points) is better clustered with the combination of features. This suggests that the combination of BSAGE and volume features can take advantage of the strengths of each feature type to better separate the data into different classes. However, the PD (brown points) class is not well separated from the other classes. This may be due to the fact that the PD class is difficult to be detected using T1w sMRI data.

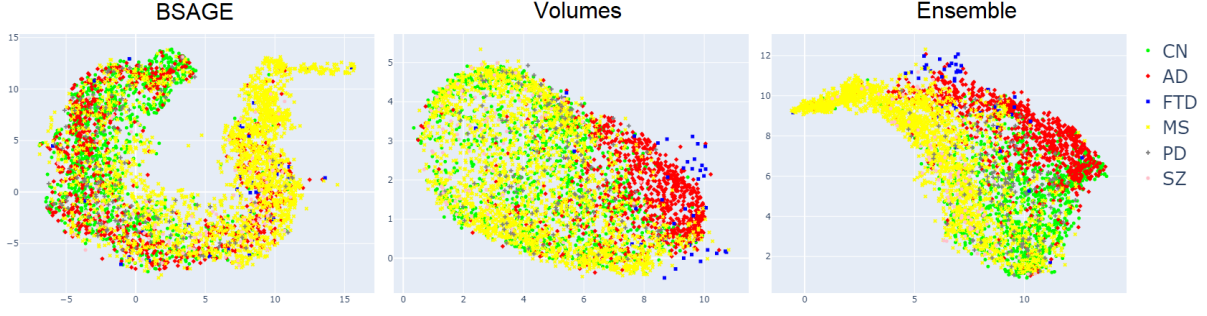


Figure 4.3: UMAP visualization of BSAGE, volume and the combination

Finally, the combination of appear to exhibit the best separation among the three types of features. Indeed, the silhouette score [211] of the combination is 0.66, greater than the disease coordinate score of 0.54 and the volume score of 0.40.

4.4.3 Predicted brain age of different populations

In this section, we compare the predicted brain age between different populations (*i.e.*, CN, AD, FTD, MS, PD and SZ). Figure 4.4 summarizes the distribution of predicted brain age for six considered populations. The median and mean predicted brain ages of CN, AD, FTD, MS, PD and SZ are respectively (-1.2, -1.1), (3.0, 3.3), (9.7, 10.4), (10.7, 11.4), (1.5, 1.3) and (5.2, 6.0). First, we observe that the CN class has the mean and median closest to 0 as expected. Second, the BrainAGE of all patient groups is significantly higher than the cognitively normal group ($p < 0.0001$ with T-test). Third, PD pathology seems to be closest to healthy people. Indeed, although T1 weighted MRI presents high contrast of grey/white matter, poor contrast may be found in structures related to PD (*e.g.*, subthalamic nuclei) [185]. This may explain the proximity of this class with CN class and the poor performance in PD detection (see Table 4.5). Fourth, the FTD group presents a more advanced aging process than AD group which is in line with the finding of Lee *et al.* [143]. Finally, we found the same magnitude of BSAGE for MS (10.7 years) as Cole *et al.* (about 10.8 years) [45], for AD (3.0 years) as Sendi *et al.* (2.1 years) [217] and for SZ (5.2 years) as Koutsouleris *et al.* (5.5 years) [138].

4.4.4 Interpretation of brain structure age gap estimation

In this section, we propose to visualize the variation of the age gap between brain structures. The presented results in Figure 4.5 correspond to the average BSAGE value for each structure on different populations of our out-of-domain datasets. We use the same color bar for all populations to compare the impact of each disease to the aging process.

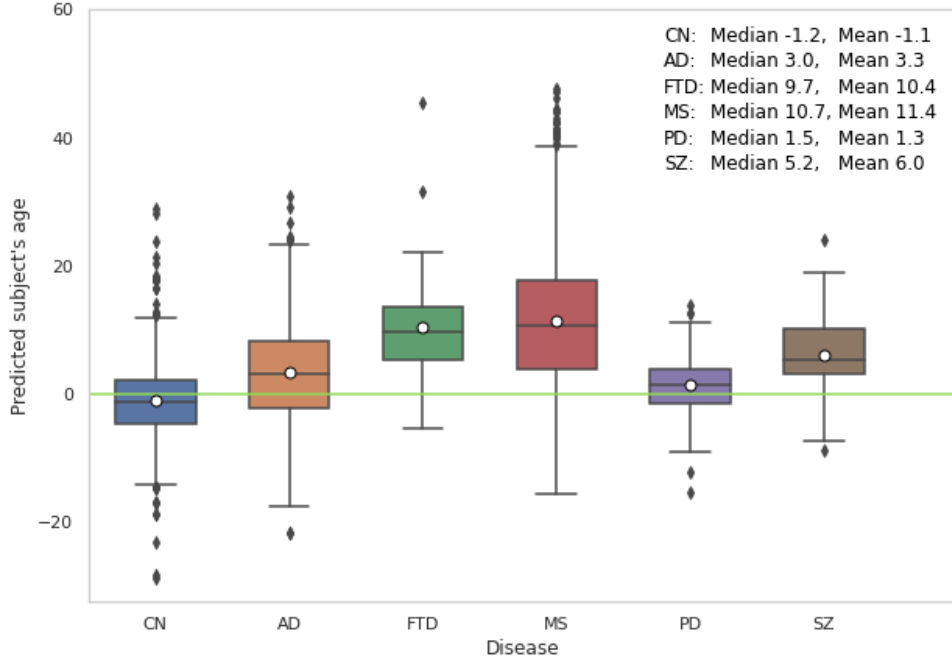


Figure 4.4: Predicted brain age of different populations in out-of-domain data. The white point presents the position of the mean value.

For the AD group, the region surrounding the hippocampus is highlighted as the most accelerated aging area. This region is well-known to be related to AD [81, 186, 120, 111]. For the FTD group, the accelerated aging pattern is mainly located in the temporal and frontal lobes which is in line with current literature [251, 29]. For the MS group, the area with the highest accelerated aging pattern is similar to the finding of Cortese *et al.* (*i.e.*, thalamus and global cortical grey matter) [47]. For the PD group, all regions seem to be close to healthy people as discussed in Section 4.4.3. Finally, for the SZ group, the prefrontal and medial temporal lobe regions are highlighted which is coherent with several studies [128, 59].

4.5 Discussion

In this work, we proposed an approach to estimate brain age at structure level. We showed that this feature can be used for different purposes. First, it can be directly used to accurately estimate the chronological age. Second, this can be used to compute the BSAGE (*i.e.*, the difference between brain structure ages and the chronological age). This biomarker presents discriminative patterns which are useful for the multi-disease classification problem (*i.e.*, CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ).

For the problem of chronological age estimation, we observed that the model ac-

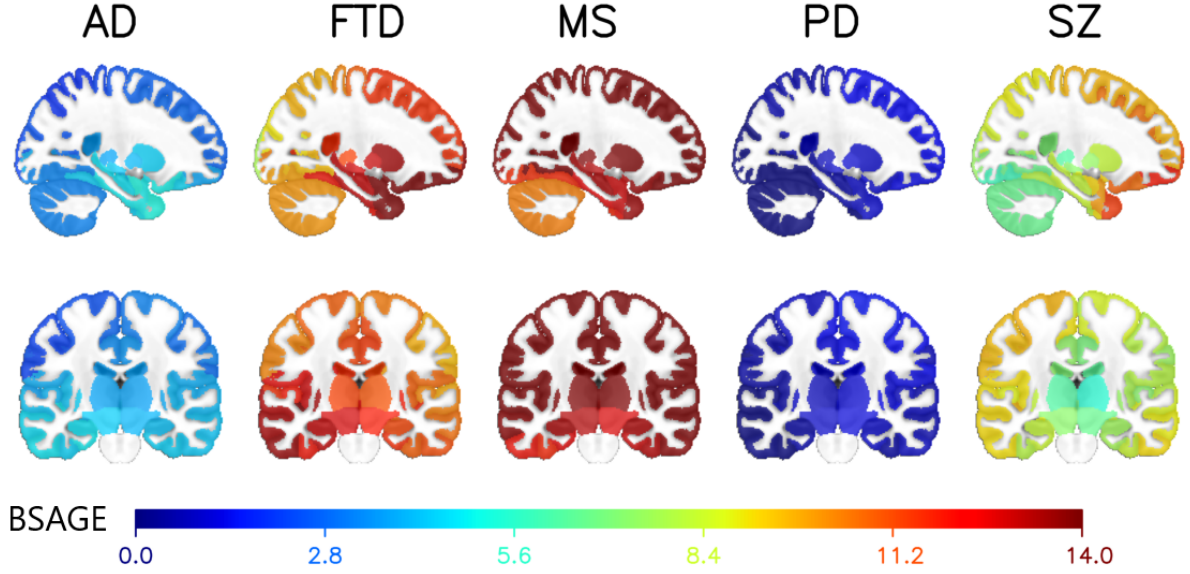


Figure 4.5: BSAGE of different populations in out-of-domain data.

curacy was heavily influenced by different factors: data amount and data augmentation techniques. In our experiments, the model accuracy was consistently improved when the data amount was increased and we did not observe saturation. This suggests that training our framework on more data could yield higher accuracy. In this study, due to the limited training data, we instead applied several data augmentation techniques to improve the generalization capacity of our model. Experimental results showed that the random shift and mixup techniques have a huge impact on our model performance (see Table 4.3). Moreover, the resulting BSA can enable a more accurate subject’s age prediction. While our framework was training over a large range of ages (*i.e.*, 0 to 95 years old), it achieved higher accuracy when predicting on young population (*i.e.*, ABIDE dataset $MAE = 1.91$ years) than on older population (*i.e.*, UKBioBank dataset $MAE = 3.87$ years). Our approach outperformed other state-of-the-art methods with 1.17 years MAE lower on the young population and 0.95 year MAE lower on the old population. When evaluating our approach on different disease populations (*i.e.*, AD, FTD, MS, PD, SZ), our findings were in line with current knowledge in the literature (see Section 4.4.3 for more details). Finally, we observed that the predicted age distributions are different between diseases, suggesting a discriminative power of our BSA feature.

While most papers used the global BrainAGE to show that a disease can present an accelerated or delayed aging process on a population [79, 138, 80], only a few approaches have proposed to use it for classification [83, 77, 243, 41]. Moreover, these studies dedicated to classification had a common limitation. Indeed, it might exist a range of global BrainAGE values that is presented in different populations. In our case, all populations had global BrainAGE values in range $[-5, 10]$ (see Figure 4.4). When the number of classes

is increased, this limitation becomes more challenging. This might explain why existing BrainAGE-based algorithms only address classification problems with a low number of class (*e.g.*, binary classification). This raises the need for other features better describing the brain aging process for classification. Thus, we propose to extend the notion of BrainAGE to BSAGE. This local feature offers a richer representation of the brain aging process than global BrainAGE estimation. Consequently, this information is important for improving the multi-disease classification as demonstrated in Section 4.4.2.

In addition to improve classification performance, BSAGE can be projected into a brain segmentation for visualization purpose. Principal remarks for each population were discussed in Section 4.4.4. Overall, it gives some insights about the specific structures impacted by each disease. The main patterns of each disease highlighted by our color maps are coherent with current literature (as discussed in Section 4.4.4). This presents an important clinical value of our framework in a real medical context.

Finally, it still exists some limitations in this study. For example, some diseases such as Parkinson’s Disease cannot be easily detected using T1 weighted MRI. Future works should focus on the multi-modal input to either accurately estimate brain age or produce a more discriminative BSAGE feature. In addition, we use the same CNN architecture to analyze different brain locations. This can be not optimal due to the fact that different brain locations may have a specific set of patterns. An auto-search algorithm to select an optimal architecture for each brain region would be beneficial for further analysis.

Chapter 5

Transformer for differential diagnosis

The method presenting in this chapter is related to the publication:

- [6] **Nguyen, Huy-Dung**, Michaël Clément, Boris Mansencal, and Pierrick Coupé. “3D Transformer based on deformable patch location for differential diagnosis between Alzheimer’s disease and Frontotemporal dementia”. In: *The 14th International Workshop on Machine Learning in Medical Imaging. MLMI 2023, Held in Conjunction with MICCAI 2023*. Vol. 14349. 2023, pp. 53–63. DOI: [10.1007/978-3-031-45676-3_6](https://doi.org/10.1007/978-3-031-45676-3_6).
-

In previous chapters, various biomarkers for single-disease and multi-disease diagnosis have been presented. These biomarkers were designed as two-stage frameworks. Following the completion of these frameworks, consideration was given to the possibility of a simpler, end-to-end framework that could achieve similar capabilities to the two-stage approaches. Although the two-stage frameworks generally outperformed state-of-the-art end-to-end methods, which often relied on CNN architectures, recent advancements in deep learning have highlighted the potential of Transformers to deliver competitive or even superior results compared to CNN methods. Consequently, this chapter delves into exploring the potential of Transformers for multi-disease diagnosis.

5.1 Introduction

CNN has been the dominant approach in computer vision for a long time, achieving state-of-the-art performance in various tasks such as object detection, image segmentation, and classification. In recent years, transformer-based models appear to be a promising alternative to CNN-based models in computer vision tasks. Despite the potential of transformer-based models, their utilization in disease diagnosis, particularly in tasks like differential diagnosis, remains relatively limited. This limitation arises from the computational demands and data requirements associated with these models. Medical applications necessitate the analysis of complex and high-dimensional data, which poses challenges for transformer-based models due to their heavy computational requirements and the need for substantial amounts of labeled medical data for effective training. This highlights the need for further research to optimize these models for medical applications.

To address the computation challenges, one possible approach is to consider classification as a 2D problem. Lyu *et al.* and Jang *et al.* employed 2D features extracted from MRI using a vision transformer (ViT) for AD classification [166, 118]. However, it is worth noting that such 2D approaches may not fully exploit the spatial information available in the data, potentially impacting their performance. Regarding 3D methods, for AD diagnosis, there have been efforts to adapt transformers for AD diagnosis. Li *et al.* downsampled the input image prior to feeding it into their transformer [147], while Zhang *et al.* reduced the feature map dimension by using a large patch size for embedding [262]. However, these strategies may result in the loss of fine-grained details in local regions. For natural image classification, other techniques have been explored to reduce computation. Local attention [160] and deformable attention [256] are two such approaches. Both methods aim to decrease the size of the attention matrix by reducing the number of query, key, and value points. The deformable attention mechanism, in particular, allows for the visualization of key points, aiding in better interpretation.

Transformer-based models typically demand a large amount of labeled data to achieve optimal performance [63]. However, the availability of labeled sMRI data in medical imaging is usually limited, hindering the effective training of transformer-based models. To address this data scarcity issue, data augmentation techniques have proven to be valuable in enhancing model generalization and performance in various tasks, particularly in natural image classification [237]. However, the specific effectiveness of data augmentation in the context of medical imaging and its impact on improving the performance of transformer-based models has not yet extensively investigated. Therefore, exploring and evaluating the potential benefits of data augmentation for transformer-based models in medical imaging is important for further research.

In this study, we firstly present a novel approach for the problem of multi-class differential diagnosis (*i.e.*, AD *vs.* FTD *vs.* CN) by proposing a 3D transformer-based architecture integrated with a deformable patch location (DPL) module. To reduce computation, we employ local attention instead of global attention in the backbone of our architecture. Our DPL module is inspired from the deformable attention [256], however, unlike the original model, deformable points in DPL are determined for each sub-volume of the image rather than being shared across the entire image. Secondly, we employ a data augmentation scheme during the training of transformer-based models to address the challenge of limited data availability. Our approach combines various commonly used data augmentation techniques to enhance the performance of 3D transformer-based classification models using sMRI data. While the individual techniques we employ are not novel, the specific combination we utilize, tailored for 3D transformer-based models, has not been extensively explored in the literature. Moreover, this augmentation scheme enables multi-scale predictions and contributes to improving the overall performance of our model. Finally, to further leverage the limited training data, we propose the fusion of our transformer-based method with a support vector machine (SVM) utilizing structural volumes. As a result, our framework shows competitive results compared to state-of-the-art methods for multi-class differential diagnosis.

5.2 Materials

In this study, we used the same datasets described as in Chapter 3 (see 3.2 for more details) to facilitate the comparison between this transformer-based method and the multi-channel deep grading method. To recap, the number of subjects used in this study is described in Table 5.1.

Table 5.1: Summary of participants used in our study. Data used for training are in bold, therefore MRIs from ADNI2 and NIFD are in-domain data while MRIs from NACC dataset are out-of-domain data.

| | Dataset | Statistic | CN | Dementia | |
|---------------|---------|----------------------|-----------------|----------------|----------------|
| | | | | AD | FTD |
| In-domain | ADNI2 | No. subjects | 180 | 149 | |
| | | Age (Mean \pm Std) | 73.4 \pm 6.3 | 74.7 \pm 8.1 | |
| | NIFD | No. subjects | 136 | | 150 |
| | | Age (Mean \pm Std) | 63.5 \pm 7.4 | | 63.9 \pm 7.1 |
| Out-of-domain | NACC | No. subjects | 2182 | 485 | 37 |
| | | Age (Mean \pm Std) | 68.2 \pm 10.9 | 72.3 \pm 9.6 | 64.1 \pm 6.9 |

5.3 Method description

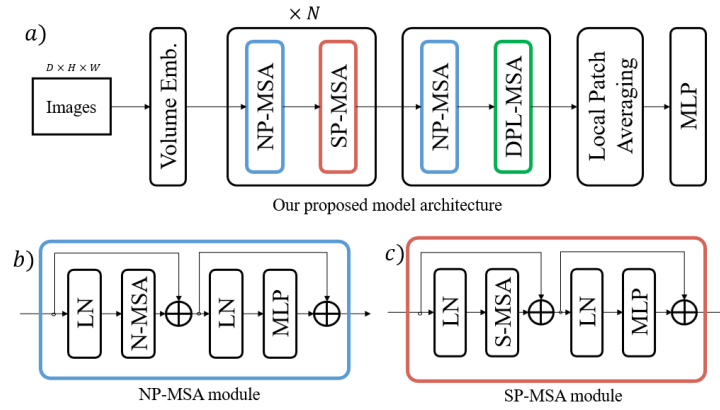


Figure 5.1: The architecture of our transformer model

5.3.1 Model overview

Figure 5.1 provides an overview of our proposed model. The model consists of four main components: a volume embedding (VE), N blocks of normal patch multi-head self-attention (NP-MSA) followed by shifted patch multi-head self-attention (SP-MSA), a block comprising NP-MSA and a deformable patch location multi-head self-attention module (DPL-MSA), a local patch averaging layer and a multi-layer perceptron (MLP).

The VE module serves to encode an input MRI into a 3D volume of token vectors. Subsequently, the tokens are processed by N blocks of NP-MSA and SP-MSA, acting as feature extractors. Additionally, the block comprising NP-MSA and DPL-MSA predicts deformable patch locations and applies attention to these patches. Unlike standard

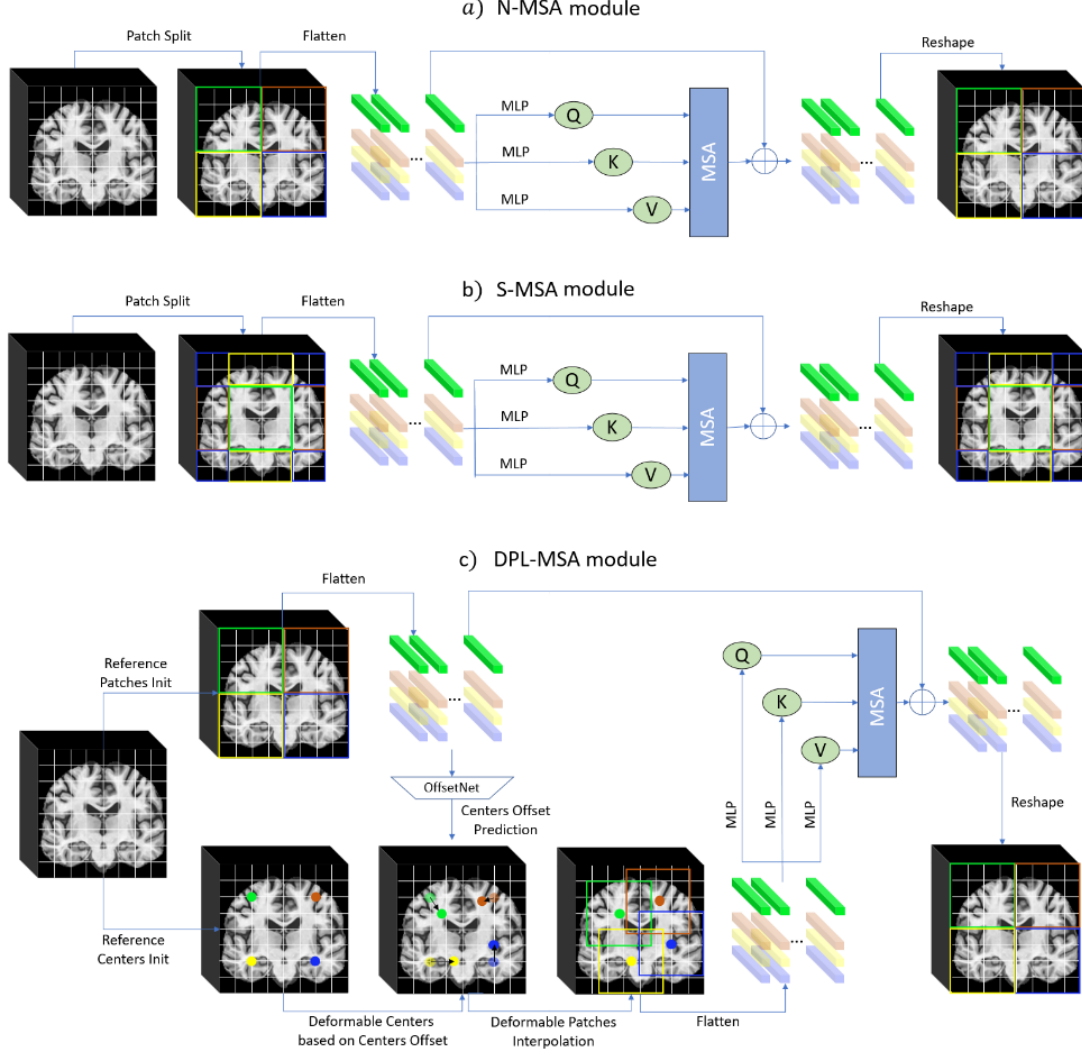


Figure 5.2: Details of different modules in our transformer model

transformer-based approaches that perform a global average of all patches together [160], our method incorporates local averaging of patches within the same area or sub-volume. This involves splitting the brain feature map into 27 sub-volumes, evenly distributed along the three axes (resulting in a $3 \times 3 \times 3$ sub-volumes). The reason behind this approach is that different brain regions may be impacted by a disease differently, thus should be weighted differently in the model decision. Finally, the output is flattened and fed into an MLP for classification purposes.

Volume embedding

We begin with a preprocessed image of size $144 \times 168 \times 144$ (with a voxel size of $1mm^3$), as shown in Figure 5.1a. The volume embedding (VE) module utilizes a CNN architecture (similar to [235]) to convert the input image into token vectors. These token

vectors have an embedding dimension of 96, resulting in a 3D feature map with 96 channels and a size of $36 \times 42 \times 36$.

Feature extractor

The resulting 3D feature map is then passed through three (NP-MSA + SP-MSA) blocks. The details of each block are presented in 5.1b,c. Our implementation of these blocks is based on [160]. The local attention size is set to $6 \times 7 \times 6$. The illustrations of each block can be found in Figure 5.1b and c. Our implementation of these blocks is based on the approach proposed by [160]. The local attention size is set to $6 \times 7 \times 6$. The key difference between the NP-MSA and SP-MSA modules based on how the 3D feature map of size $36 \times 42 \times 36$ is divided into multiple local patches, each measuring $6 \times 7 \times 6$ (see Figure 5.2a and b).

DPL block

Using the output from the feature extractor, we first update the feature map using an NP-MSA module, resulting in a 96-channel 3D feature map of size $36 \times 42 \times 36$. This updated feature map is then passed to the DPL-MSA module (refer to Figure 5.2c for detailed information on this module).

In the DPL-MSA module, we divide the feature map into $6 \times 6 \times 6$ reference patches of size $(p_x, p_y, p_z) = (6 \times 7 \times 6)$, with their respective centers denoted as $(x_{ct}^i, y_{ct}^i, z_{ct}^i)$. The coordinates of these centers are normalized within the range of $[0, 1]$. Each reference patch serves as the input to an offset network, which predicts the offset logits $(\delta_x^i, \delta_y^i, \delta_z^i)$. The deformable patch center $(x_{Dct}^i, y_{Dct}^i, z_{Dct}^i)$ is calculated as follows:

$$x_{Dct}^i = x_{ct}^i + \tanh(\delta_x^i)/(2 \times p_x) \text{ (similarly for } y_{Dct}^i \text{ and } z_{Dct}^i).$$

Based on the deformable patch centers, we interpolate the feature map to obtain the corresponding deformable patches of size $p_x \times p_y \times p_z$. These deformable patches are then fed into an NP-MSA module. Finally, a shortcut connection is added from the reference patches to the output of the NP-MSA module (see Figure 5.2c).

Local patch averaging

The 96-channel 3D brain feature map obtained, with dimensions of $36 \times 42 \times 36$, is treated as a set of $3 \times 3 \times 3$ areas, each of size $12 \times 14 \times 12$ voxels. These areas are evenly distributed along the three dimensions. We compute the average of each area, resulting in a 96-channel mean token of size $1 \times 1 \times 1$. Then, the resulting tokens from each area

are concatenated and the concatenated representation is fed into a multi-layer perceptron (MLP) for the purpose of classification.

5.3.2 Data augmentation

In this part, we describe our combination of data augmentation techniques. First, we used mixup which had been known to reduce the over-fitting phenomenon in several applications [261]. Second, we applied some affine transformations to the image including rotation and scale. This technique is commonly used in medical imaging applications [109, 189]. Finally, we randomly (with a probability p) cropped the image at a random position (see Section 5.3.4) and resized it to the input resolution. The last technique was similar to the "Random resized crop" data augmentation in 2D imaging applications [236]. In natural image classification, this technique allowed to train a network at lower resolution and then efficiently finetune it to a higher input resolution [236]. In our case, this helped to prevent over-fitting and also enabled to evaluate images at different scales. Consequently, during inference, we performed multi-scale prediction to improve model accuracy.

5.3.3 Validation framework and ensembling

When evaluating our models, we made two predictions for each image: one for the whole image and one for a crop of that image. The cropping position was one of the nine cropping positions: a center crop and eight crops at corners. The crop position that produced the lowest loss on the validation set was selected for a model. Finally, we averaged the two obtained results.

To further exploit the limited amount of training data, we combined (*i.e.*, average) the transformer prediction with SVM prediction based on brain structures volumes (see Section 5.3.4).

5.3.4 Implementation details

The offset network consisted of 3 layers: 3D convolution with 24 channels, kernel = (6, 7, 6), GELU activation [94] and another 3D convolution with 3 channels, kernel = 1. For data augmentation, rotation range was $\pm 0.05rad$ and scale range was [0.9, 1.1], the crop size was (132, 154, 132), the probability $p = 0.7$. The model was trained for 300 epochs using AdamW optimizer [163], cosine learning rate scheduler (start at $3e-4$ and end at $5e-5$). To train the SVM models, we used a grid search of three kernels (linear, polynomial, and gaussian) and 50 values of the hyper-parameter C in $[10^{-2}, 10^2]$ on the valida-

tion for tuning hyper-parameters. The SVM models used the same train/validation/test (70%/20%/10%) splits of in-domain data during cross-validation than our deep learning models.

5.4 Experimental results

In this study, we first performed a 10-fold cross-validation on in-domain dataset. This resulted in 20 models (10 Transformers and 10 SVM models). We concatenated the prediction of 10 test folds to compute the global in-domain performance. For out-of-domain validation, we averaged all the predictions to estimate the model performance on the external dataset. We used 3 metrics to assess the model performance: ACC, BACC and AUC.

5.4.1 Ablation study

Performance study First, we studied the impact of each contribution on our model performance for the differential diagnosis CN *vs.* AD *vs.* FTD. These factors could be organized into 4 groups: Input type (2D/3D), architecture (local patch averaging, non linear volume embedding), validation framework (multi-scale prediction) and ensemble (combination with SVM). The used data augmentation schema was described in 5.3.2. Table 5.2 showed the results of the comparison.

First, we implemented a basic 2D transformer-based architecture (exp. 1) and its 3D version (exp. 2) to see if the spatial information from 3D input is valuable. We observed that the 3D version was better than the 2D version in all metrics. Second, using local patch averaging (exp. 3) improved our model performance, confirming the effectiveness of assigning different weights to different brain areas. Third, the nonlinear volume embedding (exp. 4) could also improve the performance of transformer, which was inline with [235]. Then, the DPL module demonstrated an improvement in performance across almost all metrics (exp. 5). Finally, the multi-scale prediction (exp. 6) and ensembling (exp. 7) increased even more our model performance in both in-domain and out-of-domain data.

Data augmentation study Table 5.3 shows the contribution of each data augmentation technique to our model performance for the differential diagnosis CN *vs.* AD *vs.* FTD. The ensembling with SVM was removed for analysis and the multi-scale evaluation was applied only when multi-crop was used. First, without any data augmentation, the obtained result (exp. 1) was lower than in other experiments. Second, combining different augmentations (exp. 2, 3, 4) progressively improved the model’s generalization. This

Table 5.2: Ablation study of the model performance for the differential diagnosis CN *vs.* AD *vs.* FTD. Results obtained using the data augmentation described in 5.3.2. Gray text, symbols: that option is the same as in the previous experiment. Red, Blue: best, second result.

| No. | 2D/3D | Local patch averaging | Nonlinear VF | DPL module | Multi-scale prediction | Combination with SVM | In-domain | | | Out-of-domain | | |
|-----|-------|-----------------------|--------------|------------|------------------------|----------------------|-----------|------|------|---------------|------|------|
| | | | | | | | ACC | BACC | AUC | ACC | BACC | AUC |
| 1 | 2D | ✗ | ✗ | ✗ | ✗ | ✗ | 68.8 | 64.1 | 81.1 | 77.4 | 63.3 | 78.4 |
| 2 | 3D | ✗ | ✗ | ✗ | ✗ | ✗ | 78.4 | 74.7 | 90.1 | 81.5 | 75.2 | 87.8 |
| 3 | 3D | ✓ | ✗ | ✗ | ✗ | ✗ | 82.9 | 79.5 | 92.7 | 85.4 | 78.2 | 89.3 |
| 4 | 3D | ✓ | ✓ | ✗ | ✗ | ✗ | 83.6 | 80.3 | 92.5 | 86.6 | 79.7 | 89.9 |
| 5 | 3D | ✓ | ✓ | ✓ | ✗ | ✗ | 83.4 | 80.7 | 93.4 | 87.1 | 80.1 | 90.5 |
| 6 | 3D | ✓ | ✓ | ✓ | ✓ | ✗ | 85.2 | 82.5 | 94.1 | 87.7 | 80.7 | 91.0 |
| 7 | 3D | ✓ | ✓ | ✓ | ✓ | ✓ | 86.2 | 83.4 | 94.5 | 89.3 | 82.8 | 91.6 |

showed the effectiveness of our data augmentation for medical imaging applications.

5.4.2 Comparison with state-of-the-art methods

In this section, we compare our results with current state-of-the-art methods for the multi-class diagnosis AD *vs.* FTD *vs.* CN. Hu *et al.* proposed an CNN-based architecture inspired by Resnet which processes the whole 3D MRI for classification [104]. Ma *et al.* used a MLP with cortical thickness (Cth) and brain structure volumes extracted from a 3D MRI [167]. They also used a generative adversarial network to generate new data to prevent over-fitting. For a fair comparison, we reimplemented these methods and trained them under the same training setting as our method and on the same data. Table 5.4 shows the results of the comparison.

Overall, our method presented all the time the best performance in all metrics (*i.e.*, ACC, BACC and AUC) and for both in-domain and out-of-domain data. Moreover, our method was the only method based on the transformer mechanism. This suggested that transformer-based methods can obtain competitive results compared to CNN-based networks even with a limited amount of data.

Table 5.3: Ablation study of the data augmentation on our model performance for the differential diagnosis CN *vs.* AD *vs.* FTD. Gray symbols: that option is the same as in the previous experiment. **Red**, **Blue**: best, second result.

| No. | Mixup | Rand. affine | Multi crops | In-domain | | | Out-of-domain | | |
|-----|-------|-----------------|----------------|-------------|-------------|-------------|---------------|-------------|-------------|
| | | | | ACC | BACC | AUC | ACC | BACC | AUC |
| 1 | ✗ | ✗ | ✗ | 74.6 | 69.0 | 87.8 | 84.3 | 73.3 | 87.3 |
| 2 | ✓ | ✗ | ✗ | 77.6 | 72.0 | 88.4 | 84.8 | 76.0 | 87.4 |
| 3 | ✓ | ✓ | ✗ | 82.1 | 78.9 | 91.5 | 86.2 | 78.6 | 90.0 |
| 4 | ✓ | ✓ | ✓ | 85.2 | 82.5 | 94.1 | 87.7 | 80.7 | 91.0 |

Table 5.4: Comparison with state-of-the-art methods for the differential diagnosis CN *vs.* AD *vs.* FTD. **Red**, **Blue**: best, second result.

| Method | In-domain | | | Out-of-domain | | |
|------------------------------|-------------|-------------|-------------|---------------|-------------|-------------|
| | ACC | BACC | AUC | ACC | BACC | AUC |
| CNN on intensities [104] | 76.3 | 72.5 | 90.0 | 85.2 | 68.8 | 86.5 |
| MLP on Cth and volumes [167] | 77.1 | 75.9 | 86.4 | 69.1 | 74.6 | 87.5 |
| Our method | 86.2 | 83.4 | 94.5 | 89.3 | 82.8 | 91.6 |

5.4.3 Visualization of deformable patch location

Figure 5.3 shows the deformable patch locations of our models for AD and FTD patients. For each group of patients, the positions of patches were the averaged positions of our ten models. For better visualization, we applied a GradCAM to estimate an importance score (in range $[0, 1]$) for each patch. Only patches with an importance score higher than 0.3 were displayed. The obtained results were coherent with the current knowledge about these diseases. Indeed, for AD patients, the structures that obtained higher score were the left hippocampus [215], bilateral entorhinal cortex, bilateral ventricle [52] and parietal lobe [222]. In FTD patients, the frontal pole [29], superior frontal gyrus [30] and left temporal cortex [251] were highlighted.

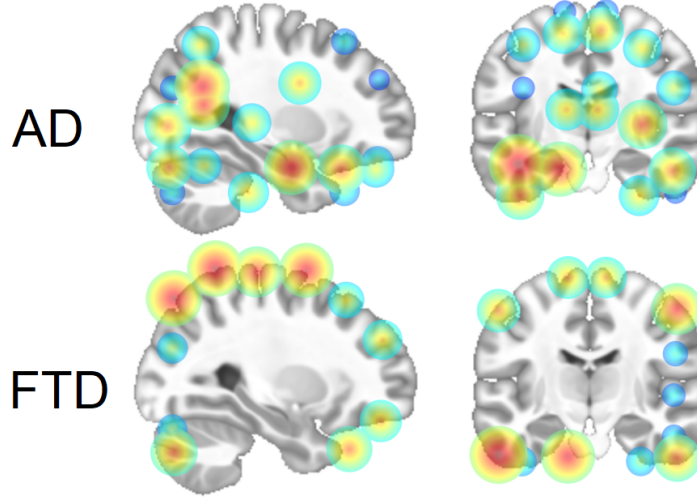


Figure 5.3: Visualization of deformable patch locations. The importance of each patch was estimated with GradCAM. Warmer color, larger radius mean higher importance score.

5.5 Comparison between the proposed Transformer and CNNs

In this section, in addition to the comparison between the proposed Transformer and CNNs for the differential diagnosis *CN vs. AD vs. FTD*, we also perform the comparison on AD diagnosis, prognosis, chronological age prediction, and for the problem of classification *CN vs. AD vs. FTD vs. MS vs. PD vs. SZ* to have an overview about our CNNs and Transformer approaches. In each comparison, the same datasets with same data split settings were used for both CNNs and Transformer models.

5.5.1 Performance comparison

Differential diagnosis *CN vs. AD vs. FTD*

Table 5.5 presents the performance comparison between our proposed CNN (multi-channel deep grading approach, see Chapter 3) and Transformer models on both in-domain and out-of-domain datasets. The information about the used data for this comparison is described in Section 3.2. The results indicate that the CNN and Transformer models demonstrate similar performance, with a slightly superior performance observed for the Transformer model. Thus, the Transformer model is the better choice in terms of generalization for the differential diagnosis of *CN vs. AD vs. FTD*.

AD diagnosis and prognosis

Table 5.6 shows the performance of the CNN model (deep grading approach, see Chapter 2) and the Transformer model on external datasets and various tasks, including AD

Table 5.5: Performance comparison between our proposed transformer and CNN (Multi-channel deep grading approach, see Chapter 3) for the classification CN *vs.* AD *vs.* FTD. The used datasets and data split setting is the same. The information about the used data for this comparison is described in Section 3.2. Red, Blue: best, second result.

| Method | In-domain | | | Out-of-domain | | |
|----------------------------|-----------|------|------|---------------|------|------|
| | ACC | BACC | AUC | ACC | BACC | AUC |
| Multi-channel deep grading | 86.0 | 84.7 | 93.8 | 87.1 | 81.6 | 91.6 |
| Transformer | 86.2 | 83.4 | 94.5 | 89.3 | 82.8 | 91.6 |

diagnosis and prognosis. The information about the used data for this comparison is described in Section 2.2. Upon examining the results, we can see that the CNN and Transformer models have similar performance for global AD diagnosis, indicating comparable generalization capacity on external datasets. However, the CNN model outperforms the Transformer model in global AD prognosis, with a margin of 73.9% versus 67.0%. As a result, the CNN method demonstrates superior generalization capacity on tasks that have not been seen before.

Table 5.6: Comparison of the CNN (Deep grading approach, see Chapter 2) and the Transformer method for AD diagnosis and prognosis. The used datasets and data split setting is the same. The information about the used data for this comparison is described in Section 2.2. **Red**: best result, **Blue**: second best result. The BACC is used to assess the model performance.

| Method | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis |
|--------------|----------------------|-------------------|--------------------|--------------------|-----------------------|------------------|--------------------------------|---------------------|
| | ADNI2 $N = 330$ | AIBL $N = 279$ | OASIS $N = 756$ | MIRIAD $N = 69$ | ADNI1 $N = 300$ | AIBL $N = 32$ | All $N = 1434$ | All $N = 332$ |
| Deep grading | 87.6 | 92.4 | 89.1 | 99.6 | 73.9 | 75.6 | 89.6 | 73.9 |
| Transformer | 89.3 | 90.9 | 90.6 | 97.8 | 66.6 | 68.6 | 89.5 | 67.0 |

Differential diagnosis CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ

Table 5.7 presents the classification results for our CNN-based model using BSAGE features (see Chapter 4) and Transformer approach in the context of multi-disease diagnosis, including CN, AD, FTD, MS, PD, and SZ. The information about the used data for this comparison is described in Section 4.2. The results show that Transformer features outperform BSAGE features, with a 25.9% improvement in in-domain data and a 10.9% improvement in out-of-domain data. However, it is noteworthy that the performance of the Transformer model drops significantly by 28.9% when transitioning from in-domain to out-of-domain data, suggesting a potential bias learned during the training process. One possible explanation could be the use of training data comprising two cohorts with different label lists, leading to the learned bias. Contrarily, the BSAGE features exhibit a smaller performance drop of 13.8% between out-of-domain and in-domain data, possibly due to a more diverse training data associated with this model and the different training objective. Finally, when combined with volume features, the BSAGE features demonstrate superior performance. Consequently, for this multi-disease diagnosis task, the recommended choice is the combination of BSAGE and volume features (referred to as BSAGE + V features).

Overall, our transformer method demonstrate better performance than the deep grading methods on in-domain data. However, on out-of-domain data or unseen task, the better performance depends on the specific application.

Table 5.7: Comparison of our BSAGE (brain structure ages gap estimation, see Chapter 4) and Transformer features for multi-disease classification CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ. The used datasets and data split setting is the same. The information about the used data for this comparison is described in Section 4.2. **Red**: best result, **Blue**: second best result. The results are presented in percentage. We denote V for structure volume.

| | No. | Features | BACC | ACC | AUC |
|---------------|-----|-----------------|-------------|-------------|-------------|
| In-domain | 1 | V | 64.5 | 65.1 | 90.2 |
| | 2 | BSAGE | 58.5 | 61.7 | 88.0 |
| | 3 | Transformer | 84.4 | 85.2 | 97.5 |
| | 4 | BSAGE + V | 68.7 | 69.6 | 93.2 |
| | 5 | Transformer + V | 84.9 | 86.0 | 97.2 |
| Out-of-domain | 6 | V | 58.7 | 59.4 | 86.8 |
| | 7 | BSAGE | 44.7 | 57.0 | 82.6 |
| | 8 | Transformer | 55.5 | 62.0 | 86.8 |
| | 9 | BSAGE + V | 63.3 | 66.1 | 90.6 |
| | 10 | Transformer + V | 59.8 | 65.7 | 89.2 |

5.5.2 Comparison of Interpretability and Explainability

Differential diagnosis CN *vs.* AD *vs.* FTD

Firstly, we compare the interpretable map generated by our multi-channel deep grading method with the explainable map produced by our transformer method (Figure 5.4). Upon observation, we find that both methods generally highlight the same brain structures. In AD patients, the left hippocampus and bilateral entorhinal cortex are consistently emphasized. In FTD patients, the frontal pole and left temporal cortex are prominently featured. Although, there are some minor differences between the two visualizations, such as the transformer method showing more attention to the superior frontal gyrus compared to the multi-channel deep grading method. Thus, both methods provide similar interpretations and explanations about AD and FTD.

Secondly, we computed the explainable map for three subtypes of FTD and compared them with the interpretable map generated by the multi-channel deep grading method (Figure 5.5). While the interpretable map generated by the multi-channel deep grading method successfully highlights the differences between FTD subtypes and is coherent with existing literature, the transformer method tends to produce similar explanation maps for all three subtypes. In this regard, the multi-channel deep grading method outperforms the transformer method.

AD diagnosis and prognosis

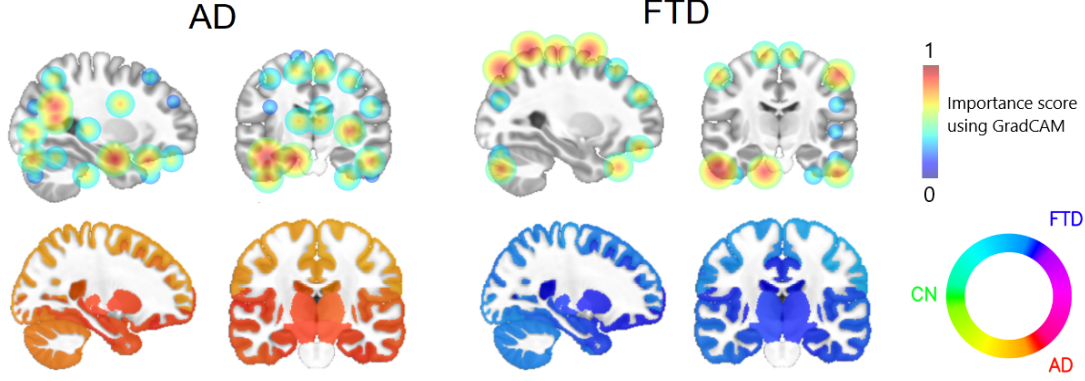


Figure 5.4: Visualization of deformable patch locations for different AD and FTD (top) and interpretable map using deep grading features (bottom).

Firstly, we examine the explainable map generated by the transformer method for AD patients. Comparing it to the deep grading map produced by our CNN method (Figure 5.6), we observe that both methods highlight the region around the bilateral hippocampus, which aligns with existing literature. However, once again, the transformer method tends to focus more on the upper regions of the brain, while the CNN method does not exhibit the same level of emphasis in those areas.

Secondly, we examine the explainable maps for sMCI and pMCI patients (Figure 5.7). We can observe that the explanation map for the pMCI group is indeed closer to the AD group, as expected. This finding indicates that the transformer method, similar to our grading map, is capable of effectively distinguishing different stages of disease progression. However, it is important to note that the progression of the disease is only visually observed in the grading map, whereas the transformer method can only show important regions leading to its decision.

Differential diagnosis CN *vs.* AD *vs.* FTD *vs.* MS *vs.* PD *vs.* SZ

We compare the explainable map generated by the transformer method for AD, FTD, MS, PD and SZ patients with the interpretable map using our BSAGE features (Figure 5.8). We can observe that both methods can effectively show the important regions for each disease. For example, hippocampus, entorhinal cortex are highlighted for patients with AD, FTD in both methods. However, there are some differences between the two methods. For example, the transformer method tends to show more emphasis on the hippocampus than other brain structure for MS and SZ patients, while our BSAGE method highlights the upper regions of the brain for MS patients and the frontal pole for SZ patients. In terms of differences between diseases, as an explainable methods, the transformer can only show the important regions leading to its decision, while our BSAGE method can show the age-related impact on brain.

Overall, both transformer and deep grading-based methods offer valuable insights

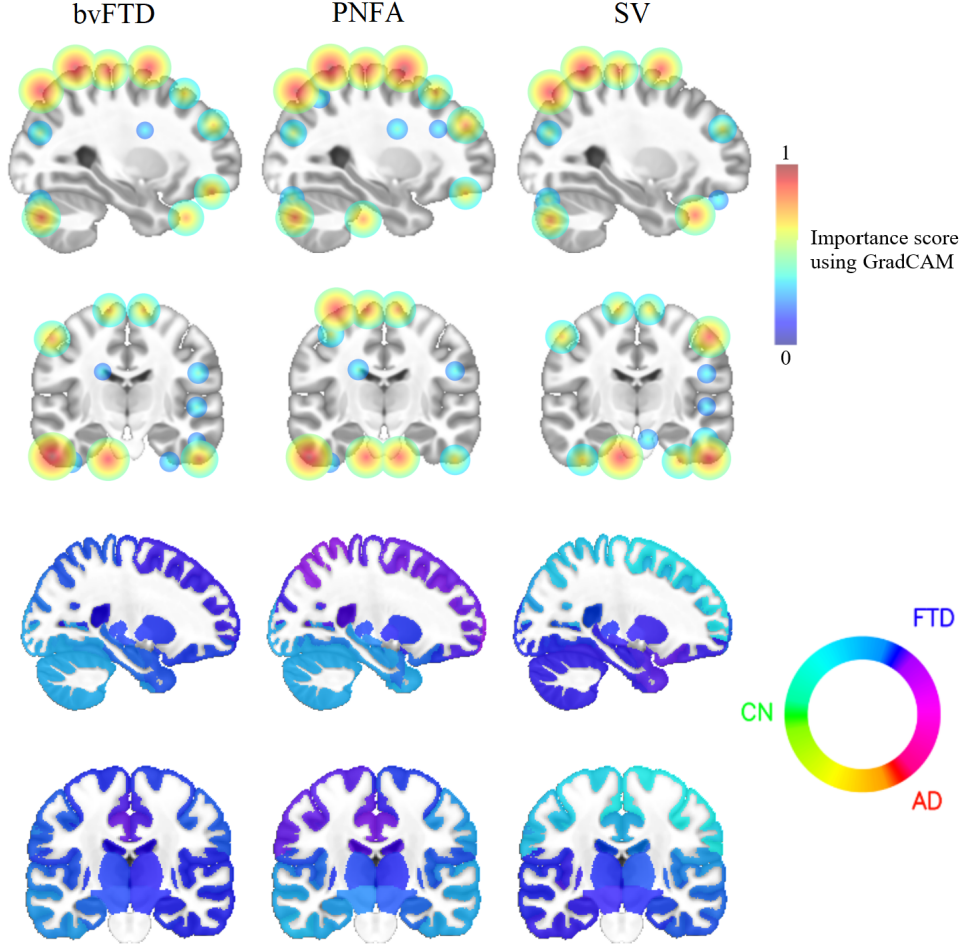


Figure 5.5: Visualization of deformable patch locations for different subtypes of FTD (top) and interpretable map using deep grading features (bottom).

of the disease, with the deep grading method demonstrating better distinction between disease’s subtypes compared to the transformer method.

5.6 Discussion

In this study, we have introduced a novel approach for multi-class differential diagnosis, specifically targeting the distinction between AD, FTD, and CN patients. Our proposed method combines a 3D transformer-based architecture with a deformable patch location (DPL) module. The experimental results demonstrate the effectiveness of our approach, indicating that an appropriate architecture, coupled with an effective data augmentation scheme, can enable transformer-based models to achieve competitive performance comparable to CNN models even with a limited amount of data. Notably, the performance of our transformer method surpasses existing state-of-the-art methods and

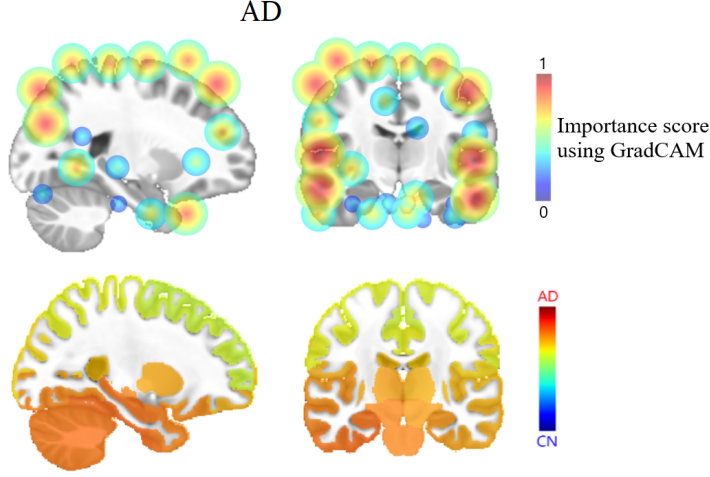


Figure 5.6: Explainable map generated for AD patients by our transformer method (top) and interpretable map using deep grading features (bottom).

slightly outperforms our deep grading method discussed in Chapter 2.

To address the challenge of limited data availability, we have designed a tailored data augmentation scheme for 3D transformer-based models. While the individual data augmentation techniques we employ are not novel, their specific combination within the context of 3D transformers remains unexplored in the literature. Leveraging techniques such as mixup, random affine transformations, and multi-crops, our augmentation scheme enhances the generalization and robustness of the model. Moreover, it enables multi-scale predictions, which significantly contributes to the overall performance improvement.

Furthermore, we propose a fusion strategy that combines our transformer-based method with a support vector machine (SVM) utilizing structural volumes. This fusion approach yields competitive results when compared to state-of-the-art methods for multi-class differential diagnosis. The combination of the transformer-based architecture and the SVM framework effectively leverages the limited training data, enhancing the discriminative capacity of our model.

One limitation of our approach is the computational complexity and resource requirements associated with the 3D transformer-based architecture. While we have employed local attention to reduce the computational problem compared to global attention, the large size of the attention matrix can still pose challenges, particularly in terms of GPU memory usage. For instance, in our experiments, a batch size of 2 requires at least one 24GB of VRAM to accommodate the attention matrix. This can limit the practicality of the model, especially in resource-constrained settings or when working with larger datasets.

Another limitation of our approach is the challenge of long-range attention when us-

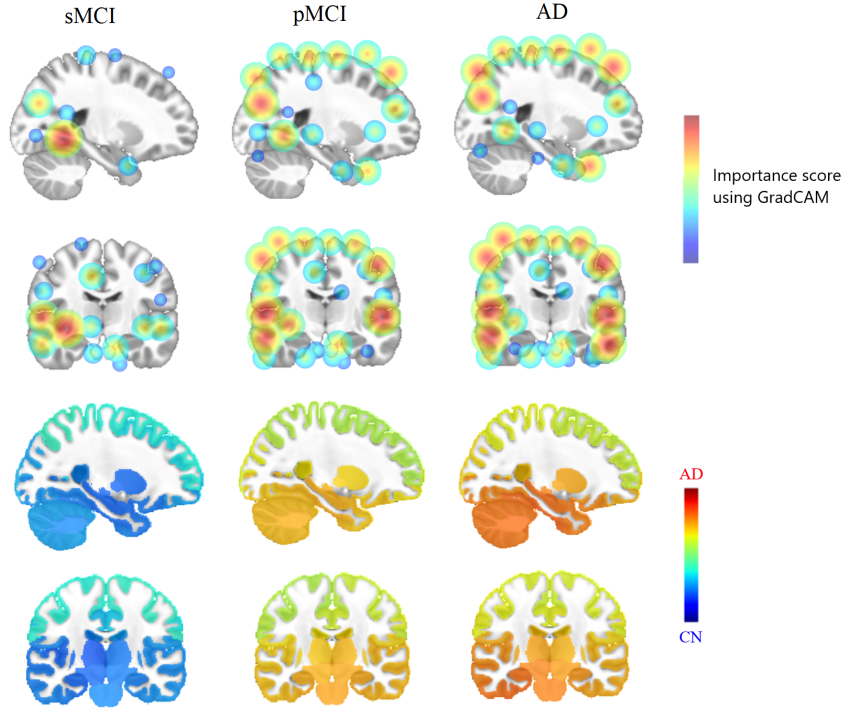


Figure 5.7: Explainable map generated for sMCI, pMCI and AD patients by our transformer method (top) and interpretable map using deep grading features (bottom).

ing local attention. Local attention is implemented to reduce computational requirements by focusing attention on a smaller region. However, this can lead to limited capturing of long-range dependencies between distant regions in the brain. While techniques such as reducing the feature map size at certain layers have been proposed to address this issue [160], implementing them significantly impacts the performance of our approach. This limitation may limit the model’s ability to capture complex long-range relationships. Future research could explore alternative strategies to effectively incorporate both local and long-range dependencies in the attention mechanism.

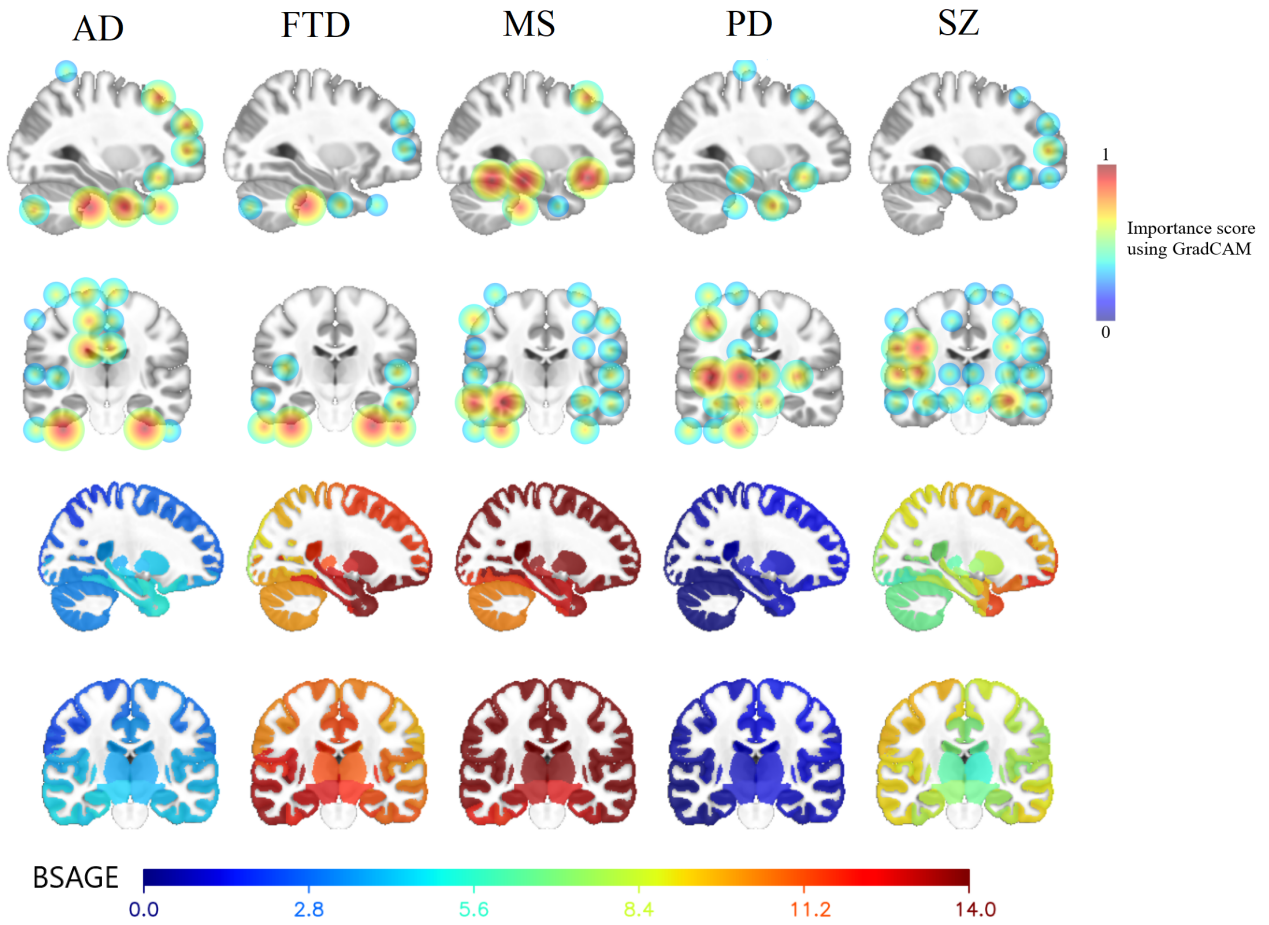


Figure 5.8: Explainable map generated for AD, FTD, MS, PD and SZ patients by our transformer method (top) and interpretable map using BSAGE features (bottom).

Chapter 6

Conclusion and perspectives

6.1 Conclusion

In this PhD, we aimed to advance the field of neurodegenerative disease diagnosis to facilitate early detection and improve the quality of life of patients. To achieve this goal, we focused on a novel generation of methods characterized by four key aspects: performance, generalization, interpretability, and applicability. The designed biomarkers had to exhibit robust performance to accurately identify neurodegenerative diseases. Moreover, these biomarkers had to demonstrate a high generalization capacity, enabling their effective application on unseen data. This ability is crucial to ensure the reliability and practicality of our diagnostic methods across diverse patient populations and data sources. Additionally, we prioritize creating biomarkers with high interpretability/explainability to gain insights to disease-related regions and severity levels, thus aiding clinicians and researchers in understanding the underlying pathology. Finally, the biomarkers had to present a broad clinical applicability, including single-disease diagnosis, multi-disease diagnosis. To address these challenges, we proposed a series of deep learning-based biomarkers for different clinical scenarios in the context of neurodegenerative disease diagnosis.

In the first study, we introduced the deep grading approach as a solution to address the challenges associated with single-disease diagnosis, with a particular focus on Alzheimer’s disease. The proposed deep grading approach proved to be highly effective in accurately diagnosing AD based on medical imaging data (*i.e.*, sMRI). One of the key strengths of the deep grading approach is its generalization capacity. We evaluated the approach on external datasets and observed that it demonstrated good generalization capacity, even when applied to unseen tasks such as AD prognosis. This highlights the robustness of the proposed approach. Another advantage of the deep grading approach is its interpretability. The proposed biomarker can offer insights into the underlying disease-

related regions and severity levels. By visualizing and quantifying the grading map, the deep grading approach allows clinicians to gain a deeper understanding of the disease and make informed decisions. As a result of this work, we have produced a conference paper [4], a journal article [1], and a software [7].

In the second study, we extended the deep grading biomarker to a multi-channel deep grading approach for the challenging task of multi-disease differential diagnosis (*i.e.*, CN *vs.* AD *vs.* FTD). Our approach yielded superior performance compared to state-of-the-art methods, demonstrating its effectiveness in both in-domain and out-of-domain data. Importantly, our method exhibited a smaller performance drop between out-of-domain and in-domain results, indicating its high generalization capacity. In terms of interpretability, the proposed biomarker revealed distinct patterns for each disease. Furthermore, our approach successfully differentiated between subtypes of FTD, providing valuable insights that contribute to a better understanding of the complexity of FTD. This work has led to the development of a conference paper [5], a journal article [2], and a software solution [8].

In the third study, we aimed to address the challenge of reduced readability in the grading map when dealing with a larger number of diseases. To tackle this issue, we proposed a novel approach based on structural brain age estimation. On one hand, the proposed biomarker demonstrated remarkable performance in estimating the chronological age of healthy subjects, surpassing current state-of-the-art methods. On the other hand, our biomarker can be used with traditional structure volume biomarkers to improve the classification performance in multi-disease differential diagnosis. Importantly, our biomarker provided a high level of interpretability thanks to the visualization of brain age gap estimation. This revealed distinct brain regions with abnormalities, which can be used to aid in the diagnosis process. Finally, this interpretability capability remains consistent regardless of the number of diseases considered. The outcomes of this research effort include a journal paper in revision [3], and two softwares [9, 10].

In the fourth study, our aim was to explore the potential of the recent transformer architecture in the field of medical imaging, specifically in the context of multi-disease diagnosis. We focused on evaluating the performance, generalization, and explainability of transformer-based models in comparison to our proposed CNN approaches. Through our investigation, we discovered that a carefully designed transformer architecture, combined with effective data augmentation techniques, can achieve competitive results comparable to our CNN models even with a limited amount of data. The transformer-based models demonstrated advantages such as compactness and faster inference times thanks to their end-to-end design. However, it is important to note that we observed a limitation in the explainability aspect of the transformer-based models. The generated explainable maps seemed to be less informative compared to the interpretable maps produced by our CNN methods. Furthermore, the transformer-based methods presented a bigger drop

in performance between in-domain and out-of-domain data compared to our CNN approaches, which may be considered as a lower generalization capacity. Nevertheless, the transformer architecture is a relatively novel approach in the field of medical imaging and requires further research to fully optimize its capacity. The contribution related to this work is a conference paper [6].

In summary, we have proposed and evaluated a series of deep learning-based biomarkers to address various clinical scenarios related to neurodegenerative disease diagnosis. On one hand, our biomarkers have demonstrated superior performance and generalization capacity compared to existing state-of-the-art methods. On the other hand, our biomarkers have shown a high level of interpretability/explainability. This aspect is crucial for the future integration of AI assistants in clinical practice. Finally, our biomarkers have demonstrated the potential to be used in a wide range of clinical applications, presenting a valuable toolkit to aid clinicians in different diagnostic scenarios.

6.2 Future works and perspectives

Although our proposed approaches demonstrated superior performance compared to existing state-of-the-art methods, there also exist some limitations. Notably, for multi-disease diagnosis, we observed a limitation in the generalization capacity of our methods on out-of-domain data, particularly when compared to the traditional structure volume biomarker. To alleviate this limitation, we decided to combine both biomarkers, aiming to enhance the model’s generalization ability. There are several potential reasons for this observation. One possibility is that the quantity of available data for multi-disease diagnosis is insufficient for effectively training deep learning models. To address this, conducting an extensive data augmentation study, such as leveraging a novel data augmentation pipeline utilizing generative adversarial networks (GAN), could be a promising approach. Another potential reason for the limited generalization of our deep learning biomarker in multi-disease diagnosis is the heterogeneity of the training data, which is derived from the combination of several distinct datasets with different list of labels. This may introduce dataset biases that decrease the generalization of the framework. To overcome this issue, it is important to explore techniques for removing dataset-side effects, such as some methods proposed in the literature [151, 131], or consider data harmonization approaches [62] to improve the generalization capacity of our approaches.

In all the proposed methods, we solely focused on utilizing the image acquired at the baseline. Our objective was to develop diagnostic methods capable of accurately identifying diseases at the earliest possible stage. By exclusively considering the baseline image, we aimed to capture the initial disease patterns and facilitate early detection. However,

it is important to note that longitudinal data can help to better validate the presence of neurodegenerative diseases [250] and their development [213]. This understanding is crucial for selecting appropriate treatment planning and intervention strategies. Future research may explore the integration of longitudinal data to further enhance the accuracy and provide a more comprehensive understanding of disease progression.

In addition to longitudinal data, the integration of multi-modal information presents potential for enhancing diagnostic accuracy. In our study, we focused solely on using the T1w contrast due to its widespread availability and usage. However, it is worth noting that certain diseases, such as Parkinson’s disease, may exhibit better recognition using T2w contrast. Furthermore, the combination of imaging data with other types of data can be a promising direction for further improving the classification performance. In this PhD, we demonstrated that incorporating the subject’s age can enhance the performance of our biomarkers, as seen in the case of our deep grading biomarker. Beyond age, various other types of data, such as genetic information, clinical results, or cognitive scores, can also be valuable in the context of disease diagnosis. One challenge of using multi-modalities input is how to effectively combine different types of input such as imaging data with tabular data. However, successfully addressing this challenge can lead to an improvement of generalization capacity [254]. Another challenge when dealing with multi-modal input is that not all modalities may be available for each subject. Effectively handling missing modalities, for example, through synthesizing missing data [155], is important and holds significant value in clinical practice. Finally, when working with multi-modal input, the interpretability of the framework may be reduced. Therefore, further research should focus on enhancing the interpretability of such frameworks, as it is a vital aspect for the future integration of AI assistants in disease diagnosis.

Appendix A

Appendix for Deep Grading

A.1 Performance measures using the AUC metric

This section presents the various performance measures of the paper using the AUC metric.

Table A.1: Comparison of different types of features for classification. All the edges are set to 1, the classifier used is GCN. **Red**: best result, **Blue**: second best result. The Area Under the ROC Curve (AUC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. All the methods were trained on the AD/CN subjects of the ADNI1 dataset. Value in bold: p of one-sided Wilcoxon test comparing with baseline (in gray) is lower than 0.05, meaning a significantly superior performance is found compared to the baseline.

| No. | Features | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis (p/sMCI) |
|-----|--------------|----------------------|-------------------|--------------------|--------------------|-----------------------|------------------|--------------------------------|---------------------------------|
| | | ADNI2 $N = 330$ | AIBL $N = 279$ | OASIS $N = 756$ | MIRIAD $N = 69$ | ADNI1 $N = 300$ | AIBL $N = 32$ | All $N = 1434$ | All $N = 332$ |
| 1 | DG_I | 97.3 | 94.8 | 93.1 | 99.5 | 74.8 | 74.6 | 95.6 | 74.3 |
| 2 | DG_{Cnw} | 96.8 | 96.5 | 95.3 | 100.0 | 76.7 | 77.1 | 96.2 | 76.6 |
| 3 | DG_C | 96.5 | 96.4 | 95.3 | 100.0 | 76.6 | 77.1 | 96.2 | 76.6 |
| 4 | V | 71.2 | 75.7 | 79.7 | 78.3 | 58.1 | 62.1 | 76.3 | 58.7 |
| 5 | A | 54.2 | 55.1 | 38.3 | 48.8 | 49.9 | 44.7 | 44.7 | 49.5 |
| 6 | V, A | 68.0 | 68.5 | 57.7 | 64.9 | 53.6 | 56.4 | 60.9 | 53.8 |
| 7 | DG_C, V | 95.8 | 96.9 | 94.5 | 99.9 | 76.5 | 77.6 | 95.7 | 76.5 |
| 8 | DG_C, A | 96.6 | 97.5 | 93.3 | 100.0 | 77.4 | 76.5 | 95.2 | 77.0 |
| 9 | DG_C, V, A | 95.9 | 97.5 | 92.8 | 99.9 | 77.4 | 76.9 | 94.8 | 77.0 |

Table A.2: Comparison of different graph edge types. The classifier used is GCN and the input features is DG_C and A. **Red**: best result, **Blue**: second best result. The Area Under the ROC Curve (AUC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. All the methods were trained on the AD/CN subjects of the ADNI1 dataset.

| Edge | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis (p/sMCI) |
|-------------------|----------------------|-------------------|--------------------|--------------------|-----------------------|------------------|--------------------------------|---------------------------------|
| | ADNI2 $N = 330$ | AIBL $N = 279$ | OASIS $N = 756$ | MIRIAD $N = 69$ | ADNI1 $N = 300$ | AIBL $N = 32$ | All $N = 1434$ | All $N = 332$ |
| Fully-one | 96.6 | 97.5 | 93.3 | 100.0 | 77.4 | 76.5 | 95.2 | 77.0 |
| Correlation | 96.8 | 97.4 | 93.0 | 100.0 | 77.3 | 76.9 | 94.1 | 76.8 |
| Volume difference | 96.8 | 97.5 | 93.4 | 100.0 | 77.3 | 76.6 | 94.4 | 76.9 |

Table A.3: Comparison of different classifiers. For graph-based approaches (*i.e.*, all the approaches except SVM and multi-layer perceptron), the edge based on structure volume difference is used and the input features is DG_C and A. **Red**: best result, **Blue**: second best result. The Area Under the ROC Curve (AUC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. All the methods were trained on the AD/CN subjects of the ADNI1 dataset.

| Classifier | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | | Global Diagnosis (AD/CN) | Global Prognosis |
|------------------------|----------------------|-------------|-------------|--------------|-----------------------|-------------|--------------------------------|---------------------|
| | ADNI2 | AIBL | OASIS | MIRIAD | ADNI1 | AIBL | All | All |
| | $N = 330$ | $N = 279$ | $N = 756$ | $N = 69$ | $N = 300$ | $N = 32$ | $N = 1434$ | $N = 332$ |
| SVM | 94.9 | 95.5 | 93.8 | 99.9 | 76.1 | 77.1 | 93.7 | 76.1 |
| Multi-layer perceptron | 90.4 | 92.8 | 91.0 | 99.9 | 73.0 | 74.9 | 89.7 | 72.8 |
| Transformer | 96.4 | 96.6 | 93.6 | 99.9 | 77.1 | 75.3 | 94.6 | 76.6 |
| SAGE | 96.7 | 97.4 | 93.0 | 99.9 | 77.1 | 76.0 | 94.1 | 76.6 |
| ResGatedGraph | 84.6 | 87.6 | 81.9 | 92.7 | 70.5 | 71.0 | 84.0 | 70.4 |
| GAT | 96.6 | 97.2 | 92.7 | 100.0 | 77.5 | 76.9 | 93.6 | 77.0 |
| TAG | 96.6 | 97.0 | 92.9 | 99.9 | 77.1 | 76.8 | 94.0 | 76.7 |
| GCN | 96.8 | 97.5 | 93.4 | 100.0 | 77.3 | 76.6 | 94.4 | 76.9 |

Table A.4: Comparison of our method with state-of-the-art methods that have been retrained on our training dataset using the available code and tested on our dataset. **Red**: best result, **Blue**: second best result. The Area Under the ROC Curve (AUC) is used to assess the model performance. All the methods are trained on the AD/CN subject of the ADNI1 dataset, the same training/testing partition is used for evaluation.

| Method | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | |
|-------------------------|----------------------|-------------------|--------------------|--------------------|-----------------------|------------------|
| | ADNI2 $N = 330$ | AIBL $N = 279$ | OASIS $N = 756$ | MIRIAD $N = 69$ | ADNI1 $N = 300$ | AIBL $N = 32$ |
| Patch-based CNN [250] | 79.3 | 86.8 | 87.8 | 88.6 | 65.5 | 52.5 |
| ROI-based CNN [250] | 90.8 | 90.8 | 92.7 | 97.4 | 69.6 | 75.0 |
| Subject-based CNN [250] | 85.4 | 90.4 | 92.4 | 98.8 | 70.0 | 59.6 |
| Voxel-based SVM [250] | 93.8 | 93.6 | 93.6 | 99.4 | 74.3 | 75.0 |
| Our method | 96.6 | 97.5 | 93.3 | 100.0 | 77.4 | 76.5 |

Table A.5: Comparison of our method with state-of-the-art methods using published results. **Red**: best result, **Blue**: second best result. The Area Under the ROC Curve (AUC) is used to assess the model performance. All the methods are trained on the AD/CN subject of the ADNI1 dataset. However, there are many different factors: number of subjects in training/testing sets, selection criteria, etc.

| Method | Diagnosis (AD/CN) | | | | Prognosis (p/sMCI) | |
|--------------------------|----------------------|-------------|-------------|--------------|-----------------------|-------------|
| | ADNI2 | AIBL | OASIS | MIRIAD | ADNI1 | AIBL |
| Landmark-based CNN [157] | 95.9 | - | - | 97.2 | - | - |
| Hierarchical FCN [152] | 95.1 | - | - | - | 78.1 | - |
| AD^2A [90] | 93.4 | 92.5 | - | - | - | - |
| Efficient3D [257] | - | - | - | - | - | - |
| Our method | 96.6 | 97.5 | 93.3 | 100.0 | 77.4 | 76.5 |

A.2 Cross-brain regions connectivity analysis

To analyze the cross-brain regions connectivity, we compute two averaged adjacency matrices (*i.e.*, edge weights) respectively for all AD patients and all CN subjects using the absolute difference of volumes. After that, we compute the absolute difference of these two matrices (see Figure A.1). This results in a matrix of size 133×133 , we then select 25 highest values (top 0.14% highest values). These values correspond to 25 pairs of structures. Among these pairs of structures, we observe some structures that have been presented in Section 2.4.2, such as bilateral hippocampus, left amygdala, left parahippocampal gyrus and left ventral diencephalon. These structures have been shown to be

related to AD [52, 81, 129, 159]. In AD patients, these structures may present more atrophy volumes than other structures. In CN people, the atrophy volumes of these structures (due to the normal aging process) may be close to other structures. Thus, the absolute difference volumes should be a discriminative feature for AD classification. And in our case, using the absolute difference volumes as the edge weights allows an improvement in performance.

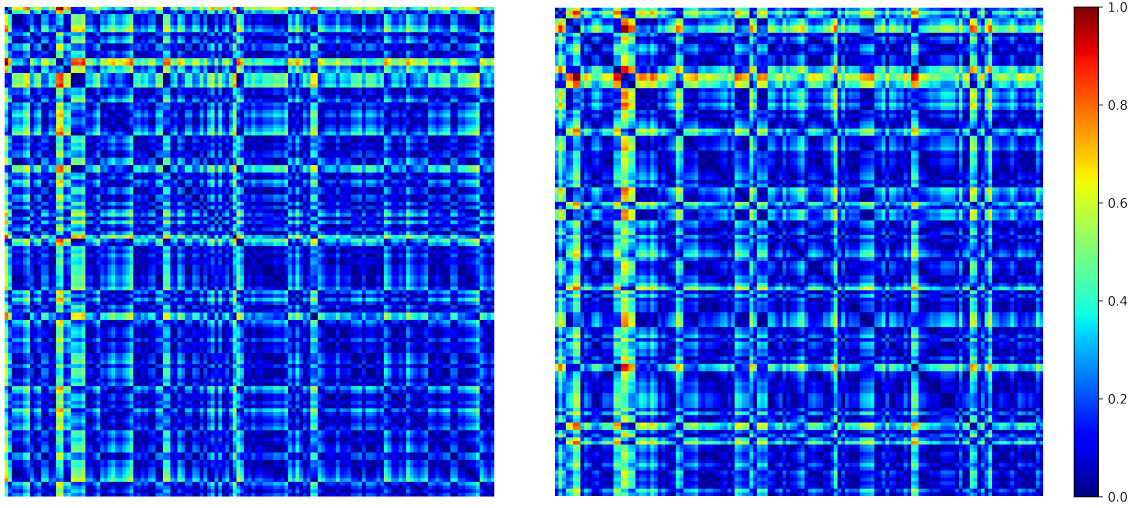


Figure A.1: Averaged adjacency matrices of AD population (left) and CN population (right). All the values are normalized to $[0, 1]$ for visualization.

Appendix B

Appendix for multi-channel deep grading

Our data splitting procedure

For each cross-validation iteration, we used seven folds as training/validation data for our 125 U-Nets in the first stage. In the second stage, we reused this data as training data for our MLP and SVM classifiers. We took one more data fold as validation data for these classifiers. Once the MLP and the SVM were trained, one more data fold was used to find the coefficient to ensemble the two classifiers. Finally, we obtained an ensemble model of MLP and SVM and one remaining unused test fold.

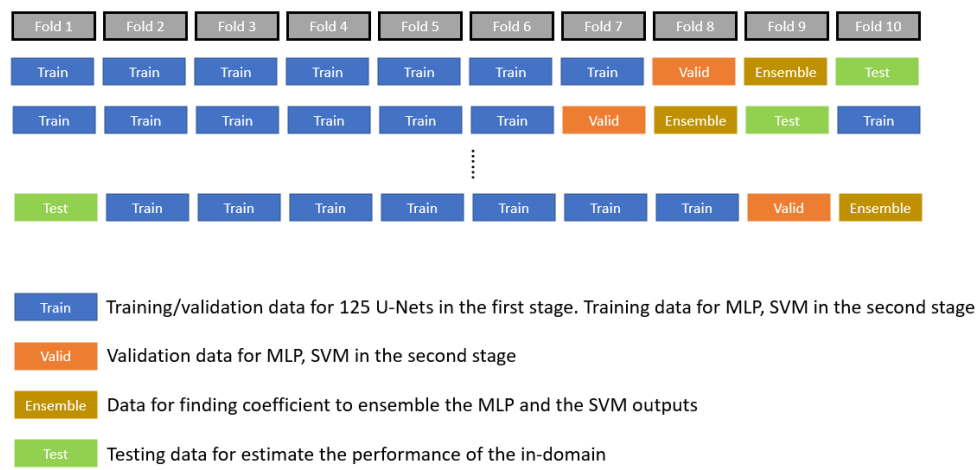


Figure B.1: Our data split procedure

Ablation study using different metrics

Table B.1: Ablation study of our method for binary classification tasks. We use the accuracy (ACC) to assess the performance. We perform 10-fold cross validation on ADNI+NIFD dataset to estimate the in-domain performance (exp. 1, 2, 3). Additionally, we evaluate on NACC dataset to estimate the out-of-domain performance (exp. 4, 5, 6) by averaging the outputs of 10 trained models. The results are presented in %. **Red**: best result, **Blue**: second best result.

| No. | Evaluation | Features | Dementia diagnosis Dem. <i>vs.</i> CN | AD diagnosis AD <i>vs.</i> CN | FTD diagnosis FTD <i>vs.</i> CN | Differential diagnosis AD <i>vs.</i> FTD |
|-----|---------------|----------|---|-------------------------------------|---------------------------------------|--|
| | | | $N = 615$ | $N = 465$ | $N = 466$ | $N = 299$ |
| 1 | In-domain | Volumes | 85.4 | 84.7 | 90.1 | 80.6 |
| 2 | | Grades | 86.3 | 87.5 | 93.1 | 94.6 |
| 3 | | Ensemble | 87.6 | 89.9 | 93.7 | 93.3 |
| | | | $N = 1627$ | $N = 1605$ | $N = 1353$ | $N = 296$ |
| 4 | Out-of-domain | Volumes | 89.7 | 92.2 | 96.6 | 86.1 |
| 5 | | Grades | 87.4 | 88.0 | 96.5 | 80.7 |
| 6 | | Ensemble | 89.5 | 91.2 | 98.2 | 85.1 |

Table B.2: Ablation study of our method for binary classification tasks. We use the area under curve (AUC) to assess the performance. We perform 10-fold cross validation on ADNI+NIFD dataset to estimate the in-domain performance (exp. 1, 2, 3). Additionally, we evaluate on NACC dataset to estimate the out-of-domain performance (exp. 4, 5, 6) by averaging the outputs of 10 trained models. The results are presented in %. **Red**: best result, **Blue**: second best result.

| No. | Evaluation | Features | Dementia diagnosis Dem. <i>vs.</i> CN | AD diagnosis AD <i>vs.</i> CN | FTD diagnosis FTD <i>vs.</i> CN | Differential diagnosis AD <i>vs.</i> FTD |
|-----|---------------|----------|---|-------------------------------------|---------------------------------------|--|
| | | | $N = 615$ | $N = 465$ | $N = 466$ | $N = 299$ |
| 1 | In-domain | Volumes | 92.3 | 91.3 | 93.7 | 87.6 |
| 2 | | Grades | 92.1 | 93.1 | 96.2 | 99.2 |
| 3 | | Ensemble | 93.5 | 93.9 | 95.3 | 96.6 |
| | | | $N = 1627$ | $N = 1605$ | $N = 1353$ | $N = 296$ |
| 4 | Out-of-domain | Volumes | 95.6 | 95.5 | 98.6 | 94.1 |
| 5 | | Grades | 94.2 | 93.4 | 92.3 | 87.3 |
| 6 | | Ensemble | 96.0 | 95.9 | 99.2 | 92.7 |

Comparison with current state-of-the-art methods using different metrics

Table B.3: Comparison of our method with current state-of-the-art methods for binary classification tasks. Our reported performances are the average of 10 repetitions and presented in %. **Red**: best result, **Blue**: second best result. The accuracy (ACC) is used to assess the model performance.

| No. | Evaluation | Method | Dementia diagnosis | AD diagnosis | FTD diagnosis | Differential diagnosis |
|-----|---------------|------------------------|-----------------------|------------------|-------------------|---------------------------|
| | | | Dem. <i>vs.</i> CN | AD <i>vs.</i> CN | FTD <i>vs.</i> CN | AD <i>vs.</i> FTD |
| 1 | In-domain | Hu <i>et al.</i> [104] | 82.0 | 81.7 | 87.3 | 82.3 |
| 2 | | Ma <i>et al.</i> [167] | 90.2 | 91.3 | 93.5 | 90.0 |
| 3 | | Our method | 87.5 | 89.0 | 93.3 | 91.0 |
| 4 | Out-of-domain | Hu <i>et al.</i> [104] | 86.8 | 86.8 | 97.3 | 85.8 |
| 5 | | Ma <i>et al.</i> [167] | 73.0 | 83.9 | 74.8 | 74.9 |
| 6 | | Our method | 87.5 | 90.0 | 96.8 | 80.7 |

Table B.4: Comparison of our method with current state-of-the-art methods for binary classification tasks. Our reported performances are the average of 10 repetitions and presented in %. **Red**: best result, **Blue**: second best result. The area under curve (AUC) is used to assess the model performance.

| No. | Evaluation | Method | Dementia diagnosis | AD diagnosis | FTD diagnosis | Differential diagnosis |
|-----|---------------|------------------------|-----------------------|------------------|-------------------|---------------------------|
| | | | Dem. <i>vs.</i> CN | AD <i>vs.</i> CN | FTD <i>vs.</i> CN | AD <i>vs.</i> FTD |
| 1 | In-domain | Hu <i>et al.</i> [104] | 88.5 | 86.1 | 89.3 | 90.2 |
| 2 | | Ma <i>et al.</i> [167] | 84.3 | 94.8 | 96.6 | 85.3 |
| 3 | | Our method | 93.5 | 93.7 | 95.0 | 95.0 |
| 4 | Out-of-domain | Hu <i>et al.</i> [104] | 88.4 | 88.9 | 86.2 | 75.3 |
| 5 | | Ma <i>et al.</i> [167] | 90.6 | 89.7 | 85.7 | 85.4 |
| 6 | | Our method | 93.8 | 94.4 | 90.3 | 95.7 |

Appendix C

Appendix for brain structure ages

Table C.1: Ablation study for binary classification tasks. **Red**: best result, **Blue**: second best result. The ACC is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. We denote BSAGE_{nc}, BSAGE and V for BSAGE with no age correction, BSAGE with age correction and structure volume.

| | No. | Features | AD <i>vs.</i> CN | FTD <i>vs.</i> CN | MS <i>vs.</i> CN | PD <i>vs.</i> CN | SZ <i>vs.</i> CN |
|---------------|-----|---------------------|------------------|-------------------|------------------|------------------|------------------|
| In-domain | | | $N = 781$ | $N = 547$ | $N = 903$ | $N = 763$ | $N = 614$ |
| | 1 | BSAGE _{nc} | 75.8 | 76.4 | 70.9 | 71.3 | 68.9 |
| | 2 | BSAGE | 77.1 | 89.8 | 83.5 | 73.8 | 81.3 |
| | 3 | V | 89.1 | 93.1 | 79.8 | 65.0 | 81.8 |
| | 4 | BSAGE + V | 91.8 | 93.8 | 84.6 | 66.1 | 83.9 |
| Out-of-domain | | | $N = 2103$ | $N = 1273$ | $N = 3411$ | $N = 1360$ | $N = 1310$ |
| | 5 | BSAGE _{nc} | 59.4 | 69.0 | 78.3 | 42.9 | 90.1 |
| | 6 | BSAGE | 57.6 | 94.3 | 83.1 | 58.2 | 93.1 |
| | 7 | V | 86.2 | 95.4 | 73.4 | 51.6 | 91.3 |
| | 8 | BSAGE + V | 86.3 | 95.0 | 83.5 | 54.9 | 94.0 |

Table C.2: Ablation study for binary classification tasks. **Red**: best result, **Blue**: second best result. The area under curve (AUC) is used to assess the model performance. The results are the average accuracy of 10 repetitions and presented in percentage. We denote BSAGE_{nc}, BSAGE and V for BSAGE with no age correction, BSAGE with age correction and structure volume.

| | No. | Features | AD <i>vs.</i> CN | FTD <i>vs.</i> CN | MS <i>vs.</i> CN | PD <i>vs.</i> CN | SZ <i>vs.</i> CN |
|---------------|-----|---------------------|------------------|-------------------|------------------|------------------|------------------|
| In-domain | | | $N = 781$ | $N = 547$ | $N = 903$ | $N = 763$ | $N = 614$ |
| | 1 | BSAGE _{nc} | 80.8 | 82.3 | 78.0 | 75.7 | 78.8 |
| | 2 | BSAGE | 94.8 | 94.6 | 91.5 | 79.0 | 88.0 |
| | 3 | V | 95.6 | 95.2 | 87.1 | 71.6 | 88.3 |
| | 4 | BSAGE + V | 96.6 | 97.2 | 93.0 | 72.2 | 91.4 |
| Out-of-domain | | | $N = 2103$ | $N = 1273$ | $N = 3411$ | $N = 1360$ | $N = 1310$ |
| | 5 | BSAGE _{nc} | 70.1 | 69.5 | 87.3 | 64.9 | 89.8 |
| | 6 | BSAGE | 85.2 | 94.0 | 91.0 | 53.2 | 84.7 |
| | 7 | V | 94.0 | 93.9 | 78.9 | 61.8 | 88.9 |
| | 8 | BSAGE + V | 93.5 | 94.6 | 91.2 | 63.3 | 94.2 |

Bibliography

- [11] Giorgia Adani et al. “Environmental Risk Factors for Early-Onset Alzheimer’s Dementia and Frontotemporal Dementia: A Case-Control Study in Northern Italy”. In: *International Journal of Environmental Research and Public Health* 17.21 (2020), p. 7941. DOI: [10.3390/ijerph17217941](https://doi.org/10.3390/ijerph17217941).
- [12] Julius Adebayo et al. “Sanity Checks for Saliency Maps”. In: *Annual Conference on Neural Information Processing Systems 2018*. 2018, pp. 9525–9536.
- [13] Sitara Afzal et al. “A Segmentation-Less Efficient Alzheimer Detection Approach Using Hybrid Image Features”. In: *Handbook of Multimedia Information Security: Techniques and Applications*. Ed. by A. Singh and A. Mohan. Cham: Springer, 2019. DOI: [10.1007/978-3-030-15887-3_20](https://doi.org/10.1007/978-3-030-15887-3_20).
- [14] Marilyn S. Albert et al. “The diagnosis of mild cognitive impairment due to Alzheimer’s disease: recommendations from the National Institute on Aging Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 7 (2011), pp. 270–279. DOI: [10.1016/j.jalz.2011.03.008](https://doi.org/10.1016/j.jalz.2011.03.008).
- [15] Karim Armanious et al. “Age-Net: An MRI-Based Iterative Framework for Brain: A Journal of Neurology Biological Age Estimation”. In: *IEEE Transactions on Medical Imaging* 40 (2021), pp. 1778–1791. DOI: [10.1109/TMI.2021.3066857](https://doi.org/10.1109/TMI.2021.3066857).
- [16] Brian B. Avants et al. “A reproducible evaluation of ANTs similarity metric performance in brain image registration”. In: *NeuroImage* 54.3 (2011), pp. 2033–2044. DOI: [10.1016/j.neuroimage.2010.09.025](https://doi.org/10.1016/j.neuroimage.2010.09.025).
- [17] Vicki Babulas, Pam Factor-Litvak, Raymond Goetz, Catherine A. Schaefer, and Alan S. Brown. “Prenatal exposure to maternal genital and reproductive infections and adult schizophrenia”. In: *The American Journal of Psychiatry* 163.5 (2006), pp. 927–929. DOI: [10.1176/ajp.2006.163.5.927](https://doi.org/10.1176/ajp.2006.163.5.927).
- [18] Karl Bäckström, Mahmood Nazari, Irene Yu-Hua Gu, and Asgeir Store Jakola. “An efficient 3D deep convolutional network for Alzheimer’s disease diagnosis using MR images”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 149–153. DOI: [10.1109/ISBI.2018.8363543](https://doi.org/10.1109/ISBI.2018.8363543).

- [19] Alejandro Barredo A. et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115. DOI: [10.1016/j.inffus.2019.12.012](https://doi.org/10.1016/j.inffus.2019.12.012).
- [20] Cher Bass et al. “Icam-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans”. In: *arXiv* (2021).
- [21] Cher Bass et al. “ICAM: Interpretable Classification via Disentangled Representations and Feature Attribution Mapping”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 7697–7709.
- [22] Randall J. Bateman et al. “Clinical and biomarker changes in dominantly inherited Alzheimer’s disease”. In: *New England Journal of Medicine* 367.9 (2012), pp. 795–804. DOI: [10.1056/NEJMoA1202753](https://doi.org/10.1056/NEJMoA1202753).
- [23] Bárbara C. Beber et al. “Logopenic aphasia or Alzheimer’s disease: Different phases of the same disease?” In: *Dementia & Neuropsychologia* 8 (2014), pp. 302–307. DOI: [10.1590/S1980-57642014DN83000016](https://doi.org/10.1590/S1980-57642014DN83000016).
- [24] Duane L. Beekly et al. “The National Alzheimer’s Coordinating Center (NACC) database: the Uniform Data Set”. In: *Alzheimer Disease and Associated Disorders* 21.3 (2007), pp. 249–258. DOI: [10.1097/WAD.0b013e318142774e](https://doi.org/10.1097/WAD.0b013e318142774e).
- [25] Camilo Bermudez et al. “Anatomical context improves deep learning on the brain age estimation task”. In: *Magnetic Resonance Imaging* 62 (2019), pp. 70–77. DOI: [10.1016/j.mri.2019.06.018](https://doi.org/10.1016/j.mri.2019.06.018).
- [26] Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. “Graph Neural Networks in Network Neuroscience”. In: *arXiv preprint* (2021). eprint: [arXiv:2106.03535](https://arxiv.org/abs/2106.03535).
- [27] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. “Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers”. In: *Artificial Neural Networks and Machine Learning – ICANN*. 2016. DOI: [10.1007/978-3-319-44781-0_8](https://doi.org/10.1007/978-3-319-44781-0_8).
- [28] Kyriaki-Margarita Bintsi, Vasileios Baltatzis, Arinbjörn Kolbeinsson, Alexander Hammers, and Daniel Rueckert. “Patch-based brain age estimation from mr images”. In: *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*. 2020, pp. 98–107. DOI: [10.1007/978-3-030-66843-3_10](https://doi.org/10.1007/978-3-030-66843-3_10).
- [29] Bradley F. Boeve, Adam L. Boxer, Fiona Kumfor, Yolande Pijnenburg, and Jonathan D. Rohrer. “Advances and controversies in frontotemporal dementia: diagnosis, biomarkers, and therapeutic considerations”. In: *The Lancet Neurology* 21 (2022), pp. 258–272. DOI: [10.1016/S1474-4422\(21\)00341-0](https://doi.org/10.1016/S1474-4422(21)00341-0).

- [30] Simona M Brambati et al. “A tensor based morphometry study of longitudinal gray matter contraction in FTD”. In: *Neuroimage* 35.3 (2007), pp. 998–1003. DOI: [10.1016/j.neuroimage.2007.01.028](https://doi.org/10.1016/j.neuroimage.2007.01.028).
- [31] Xavier Bresson and Thomas Laurent. “Residual Gated Graph ConvNets”. In: *arXiv preprint* (2017). eprint: [arXiv:1711.07553](https://arxiv.org/abs/1711.07553).
- [32] Esther E. Bron et al. “Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge”. In: *NeuroImage* 111 (2015), pp. 562–579. DOI: [10.1016/j.neuroimage.2015.01.048](https://doi.org/10.1016/j.neuroimage.2015.01.048).
- [33] Esther E. Bron et al. “Multiparametric computer-aided differential diagnosis of Alzheimer’s disease and frontotemporal dementia using structural and advanced MRI”. In: *European Radiology* 27 (2017), pp. 3372–3382. DOI: [10.1007/s00330-016-4691-x](https://doi.org/10.1007/s00330-016-4691-x).
- [34] Esther E. Bron et al. “Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer’s disease”. In: *NeuroImage: Clinical* 31 (2021), p. 102712. DOI: [10.1016/j.nicl.2021.102712](https://doi.org/10.1016/j.nicl.2021.102712).
- [35] Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H. Michael Arrighi. “Forecasting the global burden of Alzheimer’s disease”. In: *Alzheimer’s Dementia* 3 (2007), pp. 186–191. DOI: [10.1016/j.jalz.2007.04.381](https://doi.org/10.1016/j.jalz.2007.04.381).
- [36] Juan R. Bustillo et al. “Glutamatergic and neuronal dysfunction in gray and white matter: a spectroscopic imaging study in a large schizophrenia sample”. In: *Schizophrenia Bulletin* 43.3 (2017), pp. 611–619. DOI: [10.1093/schbul/sbw122](https://doi.org/10.1093/schbul/sbw122).
- [37] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209. DOI: [10.1038/s41586-018-0579-z](https://doi.org/10.1038/s41586-018-0579-z).
- [38] Chiao-Hsiang Chang, Chin-Sheng Lin, Yu-Sheng Luo, Yung-Tsai Lee, and Chin Lin. “Electrocardiogram-Based Heart Age Estimation by a Deep Learning Model Provides More Information on the Incidence of Cardiovascular Disorders”. In: *Frontiers in Cardiovascular Medicine* 9 (2022), p. 754909. DOI: [10.3389/fcvm.2022.754909](https://doi.org/10.3389/fcvm.2022.754909).
- [39] Brian H. Chen et al. “DNA methylation-based measures of biological age: meta-analysis predicting time to death”. In: *Aging* 8 (2016), pp. 1844–1865. DOI: [10.18632/aging.101020](https://doi.org/10.18632/aging.101020).
- [40] Danni Cheng, Manhua Liu, Jianliang Fu, and Yaping Wang. “Classification of MR brain images by combination of multi-CNNs for AD diagnosis”. In: *International Conference on Digital Image Processing (ICDIP 2017)*. Vol. 10420. 2017, pp. 875–879. DOI: [10.1117/12.2281808](https://doi.org/10.1117/12.2281808).

- [41] Jian Cheng et al. “Brain Age Estimation From MRI Using Cascade Networks With Ranking Loss”. In: *IEEE Transactions on Medical Imaging* 40 (2021), pp. 3400–3412. DOI: [10.1109/TMI.2021.3085948](https://doi.org/10.1109/TMI.2021.3085948).
- [42] Tiffany W. Chow et al. “Overlap in Frontotemporal Atrophy Between Normal Aging and Patients With Frontotemporal Dementias”. In: *Alzheimer Disease & Associated Disorders* 22 (2008), pp. 327–335. DOI: [10.1097/WAD.0b013e31818026c4](https://doi.org/10.1097/WAD.0b013e31818026c4).
- [43] James H. Cole et al. “Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker”. In: *NeuroImage* 163 (2017), pp. 115–124. DOI: [10.1016/j.neuroimage.2017.07.059](https://doi.org/10.1016/j.neuroimage.2017.07.059).
- [44] James H. Cole et al. “Brain age predicts mortality”. In: *Molecular Psychiatry* 23 (2018), pp. 1385–1392. DOI: [10.1038/mp.2017.62](https://doi.org/10.1038/mp.2017.62).
- [45] James H. Cole et al. “Longitudinal Assessment of Multiple Sclerosis with the Brain: A Journal of Neurology-Age Paradigm”. In: *Annals of Neurology* 88 (2020), pp. 93–105. DOI: [10.1002/ana.25746](https://doi.org/10.1002/ana.25746).
- [46] James H. Cole and Katja Franke. “Predicting Age Using Neuroimaging: Innovative Brain: A Journal of Neurology aging Biomarkers”. In: *Trends in Neurosciences* 40 (2017), pp. 681–690. DOI: [10.1016/j.tins.2017.10.001](https://doi.org/10.1016/j.tins.2017.10.001).
- [47] Rosa Cortese, Sara Collorone, Olga Ciccarelli, and Ahmed T. Toosy. “Advances in brain imaging in multiple sclerosis”. In: *Therapeutic Advances in Neurological Disorders* 12 (2019), p. 175628641985972. DOI: [10.1177/1756286419859722](https://doi.org/10.1177/1756286419859722).
- [48] Pierrick Coupé et al. “Scoring by nonlocal image patch estimator for early detection of Alzheimer’s disease”. In: *NeuroImage Clinical* 1 (2012), pp. 141–152. DOI: [10.1016/j.nicl.2012.10.002](https://doi.org/10.1016/j.nicl.2012.10.002).
- [49] Pierrick Coupé et al. “Detection of Alzheimer’s disease signature in MR images seven years before conversion to dementia: Toward an early individual prognosis”. In: *Human Brain Mapping* 36 (2015), pp. 4758–4770. DOI: [10.1002/hbm.22926](https://doi.org/10.1002/hbm.22926).
- [50] Pierrick Coupé et al. “AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation”. In: *NeuroImage* 219 (2020), p. 117026. DOI: [10.1016/j.neuroimage.2020.117026](https://doi.org/10.1016/j.neuroimage.2020.117026).
- [51] Pierrick Coupé, Simon F. Eskildsen, José V. Manjón, Vladimir S. Fonov, and D. Louis Collins. “Simultaneous segmentation and grading of anatomical structures for patient’s classification: application to Alzheimer’s disease”. In: *NeuroImage* 59 (2012), pp. 3736–3747. DOI: [10.1016/j.neuroimage.2011.10.080](https://doi.org/10.1016/j.neuroimage.2011.10.080).

- [52] Pierrick Coupé, José Vicente Manjón, Enrique Lanuza, and Gwenaelle Catheline. “Lifespan Changes of the Human Brain: A Journal of Neurology In Alzheimer’s Disease”. In: *Scientific Report* 9 (2019), p. 3998. DOI: [10.1038/s41598-019-39809-8](https://doi.org/10.1038/s41598-019-39809-8).
- [53] Pierrick Coupé et al. “An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images”. In: *IEEE Transactions on Medical Imaging* 27.4 (2008), pp. 425–441. DOI: [10.1109/TMI.2007.906087](https://doi.org/10.1109/TMI.2007.906087).
- [54] Pierrick Coupé et al. “Hippocampal-amygdalo-ventricular atrophy score: Alzheimer disease detection using normative and pathological lifespan models”. In: *Human Brain Mapping* 43 (2022), pp. 3270–3282. DOI: [10.1002/hbm.25850](https://doi.org/10.1002/hbm.25850).
- [55] Pierrick Coupé et al. “Lifespan Neurodegeneration Of The Human Brain In Multiple Sclerosis”. In: *bioRxiv : the preprint server for biology* (2023), p. 2023.03.14.532535. DOI: [10.1101/2023.03.14.532535](https://doi.org/10.1101/2023.03.14.532535).
- [56] Marta Crous-Bou, Carolina Minguillón, Nina Gramunt, and José L. Molinuevo. “Alzheimer’s disease prevention: From risk factors to early intervention”. In: *Alzheimer’s Research & Therapy* 9.71 (2017). DOI: [10.1186/s13195-017-0297-z](https://doi.org/10.1186/s13195-017-0297-z).
- [57] Ruoxuan Cui and Manhua Liu. “Hippocampus Analysis by Combination of 3-D DenseNet and Shapes for Alzheimer’s Disease Diagnosis”. In: *IEEE journal of Biomedical and Health Informatics* 23 (2019), pp. 2099–2107. DOI: [10.1109/JBHI.2018.2882392](https://doi.org/10.1109/JBHI.2018.2882392).
- [58] Christos Davatzikos, Susan M. Resnick, X. Wu, P. Parmpi, and Christopher M. Clark. “Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI”. In: *NeuroImage* 41 (2008), pp. 1220–1227. DOI: [10.1016/j.neuroimage.2008.03.050](https://doi.org/10.1016/j.neuroimage.2008.03.050).
- [59] Lynn E. DeLisi et al. “Understanding structural brain changes in schizophrenia”. In: *Dialogues in Clinical Neuroscience* 8 (2006), pp. 71–78. DOI: [10.31887/DCNS.2006.8.1/ldelisi](https://doi.org/10.31887/DCNS.2006.8.1/ldelisi).
- [60] Janez Demšar. “Statistical Comparisons of Classifiers over Multiple Data Sets”. In: *Journal of Machine Learning Research* 7 (2006), pp. 1–30.
- [61] Adriana Di Martino et al. “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism”. In: *Molecular Psychiatry* 19.6 (2014), pp. 659–667. DOI: [10.1038/mp.2013.78](https://doi.org/10.1038/mp.2013.78).
- [62] Nicola K. Dinsdale, Mark Jenkinson, and Ana I.L. Namburete. “Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal”. In: *NeuroImage* 228 (2021), p. 117689. DOI: <https://doi.org/10.1016/j.neuroimage.2020.117689>.

- [63] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *International Conference on Learning Representations, ICLR 2021*. 2021.
- [64] Jian Du, Shanghang Zhang, Guanhang Wu, Jose M. F. Moura, and Soumya Kar. “Topology Adaptive Graph Convolutional Networks”. In: *arXiv preprint* (2017). eprint: [arXiv:1710.10370](https://arxiv.org/abs/1710.10370).
- [65] An-Tao Du et al. “Different regional patterns of cortical thinning in Alzheimer’s disease and frontotemporal dementia”. In: *Brain: A Journal of Neurology* 130 (2006), pp. 1159–1166. DOI: [10.1093/brain/awm016](https://doi.org/10.1093/brain/awm016).
- [66] Ranjan Duara et al. “Frontotemporal Dementia and Alzheimer’s Disease:Differential Diagnosis”. In: *Dementia and Geriatric Cognitive Disorders* 10 (1999), pp. 37–42. DOI: [10.1159/000051210](https://doi.org/10.1159/000051210).
- [67] Brittany N. Dugger and Dennis W. Dickson. “Pathology of Neurodegenerative Diseases”. In: *Cold Spring Harbor Perspectives in Biology* 9.7 (2017), a028035. DOI: [10.1101/cshperspect.a028035](https://doi.org/10.1101/cshperspect.a028035).
- [68] Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. “Deep learning to detect Alzheimer’s disease from neuroimaging: A systematic literature review”. In: *Computer Methods and Programs in Biomedicine* 187 (2020), p. 105242. DOI: [10.1016/j.cmpb.2019.105242](https://doi.org/10.1016/j.cmpb.2019.105242).
- [69] Kathryn A. Ellis et al. “The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer’s disease”. In: *International Psychogeriatrics* 21.4 (2009), pp. 672–687. DOI: [10.1017/S1041610209009405](https://doi.org/10.1017/S1041610209009405).
- [70] Roberto Erro and Maria Stamelou. “The Motor Syndrome of Parkinson’s Disease”. In: *International Review of Neurobiology* 132 (2017), pp. 25–32.
- [71] Julian M. Fearnley and Andrew J. Lees. “Ageing and Parkinson’s Disease: Substantia Nigra Regional Selectivity”. In: *Brain: A Journal of Neurology* 114.5 (1991), pp. 2283–2301. DOI: [10.1093/brain/114.5.2283](https://doi.org/10.1093/brain/114.5.2283).
- [72] Roberto Ferrari, Dimitrios Kapogiannis, Edward D. Huey, and Parastoo Momeni. “FTD and ALS: A Tale of Two Diseases”. In: *Current Alzheimer Research* 8.3 (2011). DOI: [10.2174/156720511795563700](https://doi.org/10.2174/156720511795563700).
- [73] Stanislao Fichelle et al. “MRI of helium-3 gas in healthy lungs: posture related variations of alveolar size”. In: *Journal of Magnetic Resonance Imaging* 20.2 (2004), pp. 331–335. DOI: [10.1002/jmri.20104](https://doi.org/10.1002/jmri.20104).
- [74] Leslie J. Findley. “The economic impact of Parkinson’s disease”. In: *Parkinsonism & Related Disorders* 13 (2007), S8–S12.

- [75] Elizabeth C. Finger. “Frontotemporal dementias”. In: *Continuum* 22.2 (2016), pp. 464–489. DOI: [10.1212/CON.0000000000000300](https://doi.org/10.1212/CON.0000000000000300).
- [76] A. L. Foundas et al. “Atrophy of the hippocampus, parietal cortex, and insula in Alzheimer’s disease: a volumetric magnetic resonance imaging study”. In: *Neuropsychiatry, neuropsychology, and behavioral neurology* 10 (1997), pp. 81–89.
- [77] Katja Franke et al. “Dementia classification based on brain age estimation”. In: *Proc MICCAI workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*. 2014, pp. 48–54.
- [78] Katja Franke and Christian Gaser. “Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain: A Journal of Neurology Aging: What Insights Have We Gained?” In: *Frontiers in Neurology* 10 (2019), p. 789. DOI: [10.3389/fneur.2019.00789](https://doi.org/10.3389/fneur.2019.00789).
- [79] Katja Franke, Eileen Luders, Arne May, Marko Wilke, and Christian Gaser. “Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI”. In: *NeuroImage* 63 (2012), pp. 1305–1312. DOI: [10.1016/j.neuroimage.2012.08.001](https://doi.org/10.1016/j.neuroimage.2012.08.001).
- [80] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. “Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters”. In: *NeuroImage* 50 (2010), pp. 883–892. DOI: [10.1016/j.neuroimage.2010.01.005](https://doi.org/10.1016/j.neuroimage.2010.01.005).
- [81] Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. “The clinical use of structural MRI in Alzheimer disease”. In: *Nature Reviews Neurology* 6 (2010), pp. 67–77. DOI: [10.1038/nrneurol.2009.215](https://doi.org/10.1038/nrneurol.2009.215).
- [82] Roba Gamal, Hoda Barka, and Mayada Hadhoud. “GAU U-Net for multiple sclerosis segmentation”. In: *Alexandria Engineering Journal* 73 (2023), pp. 625–634.
- [83] Christian Gaser, Katja Franke, Stefan Klöppel, Nikolaos Koutsouleris, and Heinrich Sauer. “BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer’s Disease”. In: *PloS One* 8 (2013), p. 67346. DOI: [10.1371/journal.pone.0067346](https://doi.org/10.1371/journal.pone.0067346).
- [84] Maryjo M. George, S. Kalaivani, and M. S. Sudhakar. “A non-iterative multi-scale approach for intensity inhomogeneity correction in MRI”. In: *Magnetic Resonance Imaging* 42 (2017), pp. 43–59. DOI: [10.1016/j.mri.2017.05.005](https://doi.org/10.1016/j.mri.2017.05.005).
- [85] Emilie Gerardin et al. “Multidimensional classification of hippocampal shape features discriminates Alzheimer’s disease and mild cognitive impairment from normal aging”. In: *NeuroImage* 47.4 (2009), pp. 1476–1486. DOI: [10.1016/j.neuroimage.2009.05.036](https://doi.org/10.1016/j.neuroimage.2009.05.036).

- [86] Sarah J. Getz and Bonnie Levin. “Cognitive and neuropsychiatric features of early Parkinson’s disease”. In: *Archives of Clinical Neuropsychology* 32.7 (2017), pp. 769–785. DOI: [10.1093/arclin/acx091](https://doi.org/10.1093/arclin/acx091).
- [87] Bernardino Ghetti et al. “Frontotemporal dementia caused by microtubule-associated protein tau gene (MAPT) mutations: a chameleon for neuropathology and neuroimaging”. In: *Neuropathology and Applied Neurobiology* 41.1 (2015), pp. 24–46. DOI: [10.1111/nan.12213](https://doi.org/10.1111/nan.12213).
- [88] Brian A. Gordon et al. “Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer’s disease: a longitudinal study”. In: *Lancet. Neurology* 17 (2018), pp. 241–250. DOI: [10.1016/S1474-4422\(18\)30028-0](https://doi.org/10.1016/S1474-4422(18)30028-0).
- [89] Cynthia Graber. “Diagnostics: Getting a clear picture”. In: *Nature* 484 (2012), S7–S7. DOI: [10.1038/nature11101](https://doi.org/10.1038/nature11101).
- [90] Hao Guan, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. “Attention-Guided Deep Domain Adaptation for Brain: A Journal of Neurology Dementia Identification with Multi-site Neuroimaging Data”. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Vol. 12444. 2020, pp. 31–40. DOI: [10.1007/978-3-030-60548-3_4](https://doi.org/10.1007/978-3-030-60548-3_4).
- [91] Yane Guo et al. “Grey-matter volume as a potential feature for the classification of Alzheimer’s disease and mild cognitive impairment: an exploratory study”. In: *Neuroscience Bulletin* 30 (2014), pp. 477–489. DOI: [10.1007/s12264-013-1432-x](https://doi.org/10.1007/s12264-013-1432-x).
- [92] Renske Hamel et al. “The trajectory of cognitive decline in the pre-dementia phase in memory clinic visitors: findings from the 4C-MCI study”. In: *Psychological Medicine* 45 (2015), pp. 1509–1519. DOI: [10.1017/S0033291714002645](https://doi.org/10.1017/S0033291714002645).
- [93] William L. Hamilton, Rex Ying, and Jure Leskovec. “Inductive Representation Learning on Large Graphs”. In: *arXiv preprint* (2017). eprint: [arXiv:1706.02216](https://arxiv.org/abs/1706.02216).
- [94] Dan Hendrycks and Kevin Gimpel. “Gaussian Error Linear Units (GELUs)”. In: *arXiv* (2016). eprint: [arXiv:1606.08415](https://arxiv.org/abs/1606.08415).
- [95] Maya L. Henry and Maria L. Gorno-Tempini. “The logopenic variant of primary progressive aphasia”. In: *Current Opinion in Neurology* 23 (2010), pp. 633–637. DOI: [10.1097/WCO.0b013e32833fb93e](https://doi.org/10.1097/WCO.0b013e32833fb93e).
- [96] Adelina Comas Herrera et al. *World Alzheimer Report 2016: Improving healthcare for people with dementia. Coverage, quality and costs now and in the future*. Tech. rep. 2016. DOI: [10.13140/RG.2.2.22580.04483](https://doi.org/10.13140/RG.2.2.22580.04483).

- [97] Kilian Hett, Vinh-Thong Ta, José V. Manjón, and Pierrick Coupé. “Adaptive fusion of texture-based grading for Alzheimer’s disease classification”. In: *Computerized Medical Imaging and Graphics* 70 (2018), pp. 8–16. DOI: [10.1016/j.compmedimag.2018.08.002](https://doi.org/10.1016/j.compmedimag.2018.08.002).
- [98] Kilian Hett, Vinh-Thong Ta, José V. Manjón, and Pierrick Coupé. “Graph of brain structures grading for early detection of Alzheimer’s disease”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. DOI: [10.1007/978-3-030-00931-1_49](https://doi.org/10.1007/978-3-030-00931-1_49).
- [99] Kilian Hett, Vinh-Thong Ta, Ipek Oguz, José V. Manjón, and Pierrick Coupé. “Multi-scale graph-based grading for Alzheimer’s disease prediction”. In: *Medical Image Analysis* 67 (2021), p. 101850. DOI: [10.1016/j.media.2020.101850](https://doi.org/10.1016/j.media.2020.101850).
- [100] David M. Holtzman, John C. Morris, and Alison M. Goate. “Alzheimer’s disease: the challenge of the second century”. In: *Science Translational Medicine* 3.77 (2011), 77sr1. DOI: [10.1126/scitranslmed.3002369](https://doi.org/10.1126/scitranslmed.3002369).
- [101] Marcia Hon and Naimul Mefraz Khan. “Towards Alzheimer’s disease classification through transfer learning”. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017, pp. 1166–1169. DOI: [10.1109/BIBM.2017.8217822](https://doi.org/10.1109/BIBM.2017.8217822).
- [102] Jean-François Horn et al. “Differential automatic diagnosis between Alzheimer’s disease and frontotemporal dementia based on perfusion SPECT images”. In: *Artificial Intelligence in Medicine* 47.2 (2009), pp. 147–158. DOI: [10.1016/j.artmed.2009.05.001](https://doi.org/10.1016/j.artmed.2009.05.001).
- [103] Sharpley Hsieh, Simon Schubert, Catherine Hoon, Eneida Mioshi, and John R. Hodges. “Validation of the Addenbrooke’s Cognitive Examination III in Frontotemporal dementia and Alzheimer’s disease”. In: *Dementia and geriatric cognitive disorders* 36.3-4 (2013), pp. 242–250. DOI: [10.1159/000351671](https://doi.org/10.1159/000351671).
- [104] Jingjing Hu et al. “Deep Learning-Based Classification and Voxel-Based Visualization of Frontotemporal Dementia and Alzheimer’s Disease”. In: *Frontiers in Neuroscience* 14 (2021), p. 626154. DOI: [10.3389/fnins.2020.626154](https://doi.org/10.3389/fnins.2020.626154).
- [105] Jiashuang Huang, Luping Zhou, Lei Wang, and Daoqiang Zhang. “Attention-Diffusion-Bilinear Neural Network for Brain: A Journal of Neurology Network Analysis”. In: *IEEE Transactions on Medical Imaging* 39 (2020), pp. 2541–2552. DOI: [10.1109/TMI.2020.2973650](https://doi.org/10.1109/TMI.2020.2973650).
- [106] Yadong Huang. “Roles of apolipoprotein E4 (ApoE4) in the pathogenesis of Alzheimer’s disease: lessons from ApoE mouse models”. In: *Biochemical Society Transactions* 39.4 (2011), pp. 924–932. DOI: [10.1042/BST0390924](https://doi.org/10.1042/BST0390924).

- [107] Yechong Huang, Jiahang Xu, Yuncheng Zhou, Tong Tong, and Xiahai Zhuang. “Diagnosis of Alzheimer’s Disease via Multi-Modality 3D Convolutional Neural Network”. In: *Frontiers in Neuroscience* 13 (2019). DOI: [10.3389/fnins.2019.00509](https://doi.org/10.3389/fnins.2019.00509).
- [108] Wyke Huizinga et al. “A spatio-temporal reference model of the aging brain”. In: *NeuroImage* 169 (2018), pp. 11–22. DOI: [10.1016/j.neuroimage.2017.10.040](https://doi.org/10.1016/j.neuroimage.2017.10.040).
- [109] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. “Differential data augmentation techniques for medical imaging classification tasks”. In: *AMIA Annual Symposium Proceedings*. Vol. 2017. 2017, p. 979.
- [110] Amanda Hutchinson and Jane L. Mathias. “Neuropsychological deficits in frontotemporal dementia and Alzheimer’s disease: a meta-analytic review”. In: *Journal of Neurology, Neurosurgery and Psychiatry* 78 (2007), pp. 917–928. DOI: [10.1136/jnnp.2006.100669](https://doi.org/10.1136/jnnp.2006.100669).
- [111] Bradley T. Hyman, Gary W. Van Hoesen, Antonio R. Damasio, and Clifford L. Barnes. “Alzheimer’s Disease: Cell-Specific Pathology Isolates the Hippocampal Formation”. In: *Science* 225 (1984), pp. 1168–1170. DOI: [10.1126/science.6474172](https://doi.org/10.1126/science.6474172).
- [112] Vera Ignjatovic et al. “Age-Related Differences in Plasma Proteins: How Plasma Proteins Change from Neonates to Adults”. In: *PloS One* 6 (2011), p. 17213. DOI: [10.1371/journal.pone.0017213](https://doi.org/10.1371/journal.pone.0017213).
- [113] Parkinson Progression Marker Initiative. “The Parkinson Progression Marker Initiative (PPMI)”. In: *Progress in Neurobiology* 95.4 (2011), pp. 629–635. DOI: [10.1016/j.pneurobio.2011.09.005](https://doi.org/10.1016/j.pneurobio.2011.09.005).
- [114] Clifford R. Jack et al. “Medial temporal atrophy on MRI in normal aging and very mild Alzheimer’s disease”. In: *Neurology* 49 (1997), pp. 786–794. DOI: [10.1212/wnl.49.3.786](https://doi.org/10.1212/wnl.49.3.786).
- [115] Clifford R. Jack et al. “The Alzheimer’s Disease Neuroimaging Initiative (ADNI): MRI methods”. In: *Journal of magnetic resonance imaging* 27.4 (2008), pp. 685–691. DOI: [10.1002/jmri.21049](https://doi.org/10.1002/jmri.21049).
- [116] Clifford R. Jack et al. “Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade”. In: *Lancet. Neurology* 9 (2010), pp. 119–128. DOI: [10.1016/S1474-4422\(09\)70299-6](https://doi.org/10.1016/S1474-4422(09)70299-6).
- [117] Rachna Jain, Nikita Jain, Akshay Aggarwal, and D. Jude Hemanth. “Convolutional neural network based Alzheimer’s disease classification from magnetic resonance brain images”. In: *Cognitive Systems Research* 57 (2019), pp. 147–159. DOI: [10.1016/j.cogsys.2018.12.015](https://doi.org/10.1016/j.cogsys.2018.12.015).

- [118] Jinseong Jang and Dosik Hwang. “M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20718–20729. DOI: [10.1109/CVPR52688.2022.02006](https://doi.org/10.1109/CVPR52688.2022.02006).
- [119] X. Jia et al. “Longitudinal Study of Gray Matter Changes in Parkinson Disease”. In: *American Journal of Neuroradiology* 36 (2015), pp. 2219–2226. DOI: [10.3174/ajnr.A4447](https://doi.org/10.3174/ajnr.A4447).
- [120] Kunlin Jin et al. “Increased hippocampal neurogenesis in Alzheimer’s disease”. In: *Proceedings of the National Academy of Sciences* 101 (2004), pp. 343–347. DOI: [10.1073/pnas.2634794100](https://doi.org/10.1073/pnas.2634794100).
- [121] Taeho Jo, Kwangsik Nho, and Andrew J. Saykin. “Deep Learning in Alzheimer’s Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data”. In: *Frontiers in Aging Neuroscience* 11 (2019), p. 220. DOI: [10.3389/fnagi.2019.00220](https://doi.org/10.3389/fnagi.2019.00220).
- [122] Andreas Johnen and Maxime Bertoux. “Psychological and Cognitive Markers of Behavioral Variant Frontotemporal Dementia-A Clinical Neuropsychologist’s View on Diagnostic Criteria and Beyond”. In: *Frontiers in Neurology* 10 (2019), p. 594. DOI: [10.3389/fneur.2019.00594](https://doi.org/10.3389/fneur.2019.00594).
- [123] Paul Johns. “Dementia”. In: *Clinical Neuroscience*. 2014, pp. 145–162.
- [124] L. W. de Jong et al. “Strongly reduced volumes of putamen and thalamus in Alzheimer’s disease: an MRI study”. In: *Brain: A Journal of Neurology* 131 (2008), pp. 3277–3285. DOI: [10.1093/brain/awn278](https://doi.org/10.1093/brain/awn278).
- [125] Benedikt A. Jonsson et al. “Brain age prediction using deep learning uncovers associated sequence variants”. In: *Nature Communications* 10 (2019), p. 5409. DOI: [10.1038/s41467-019-13163-9](https://doi.org/10.1038/s41467-019-13163-9).
- [126] Wonsik Jung, Eunji Jun, and Heung-Il Suk. “Deep recurrent model for individualized prediction of Alzheimer’s disease progression”. In: *NeuroImage* 237 (2021), p. 118143. DOI: [10.1016/j.neuroimage.2021.118143](https://doi.org/10.1016/j.neuroimage.2021.118143).
- [127] Reda Abdellah Kamraoui et al. “DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation”. In: *Medical Image Analysis* 76 (2022), p. 102312. DOI: [10.1016/j.media.2021.102312](https://doi.org/10.1016/j.media.2021.102312).
- [128] Katherine H. Karlsgodt, Daqiang Sun, and Tyrone D. Cannon. “Structural and Functional Brain: A Journal of Neurology Abnormalities in Schizophrenia”. In: *Current Directions in Psychological Science* 19 (2010), pp. 226–231. DOI: [10.1177/0963721410377601](https://doi.org/10.1177/0963721410377601).

- [129] J. Patrick Kesslak, Orhan Nalcioglu, and Carl W. Cotman. “Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer’s disease”. In: *Neurology* 41 (1991), pp. 51–54. DOI: [10.1212/wnl.41.1.51](#).
- [130] Ali S. Khashan et al. “Higher risk of offspring schizophrenia following antenatal maternal exposure to severe adverse life events”. In: *Archives of General Psychiatry* 65.2 (2008), pp. 146–152. DOI: [10.1001/archgenpsychiatry.2007.20](#).
- [131] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. “BiaSwap: Removing dataset bias with bias-tailored swapping augmentation”. In: *CoRR* abs/2108.10008 (2021). arXiv: [2108.10008](#).
- [132] Jun Pyo Kim et al. “Machine learning based hierarchical classification of frontotemporal dementia and Alzheimer’s disease”. In: *NeuroImage: Clinical* 23 (2019), p. 101811. DOI: [10.1016/j.nicl.2019.101811](#).
- [133] Thomas N. Kipf and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *5th International Conference on Learning Representations, ICLR 2017*. 2017.
- [134] Christine Klein and Ana Westenberger. “Genetics of Parkinson’s disease”. In: *Cold Spring Harbor Perspectives in Medicine* 2.1 (2012), a008888. DOI: [10.1101/cshperspect.a008888](#).
- [135] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. “Residual and plain convolutional neural networks for 3D brain MRI classification”. In: *IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017, pp. 835–838. DOI: [10.1109/ISBI.2017.7950647](#).
- [136] Vijay Kotu et al. “Data Mining Process”. In: 2015, pp. 17–36.
- [137] Lambrini Kourkouta, Areti Tsaloglidou, Christos Iliadis, Alexandros Monios, and Konstantinos Koukourikos. “Effects of Multiple Sclerosis”. In: *International Journal of Innovative Medicine and Health Science* 5 (2015), pp. 4–7.
- [138] Nikolaos Koutsouleris et al. “Accelerated Brain: A Journal of Neurology Aging in Schizophrenia and Beyond: A Neuroanatomical Marker of Psychiatric Disorders”. In: *Schizophrenia Bulletin* 40 (2014), pp. 1140–1153. DOI: [10.1093/schbul/sbt142](#).
- [139] Pamela J. LaMontagne et al. “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults”. In: *medRxiv preprint* (2019). eprint: [medRxiv:2019.12.13.19014902](#).

- [140] Hans Lassmann and Jack van Horssen. “The molecular basis of neurodegeneration in multiple sclerosis”. In: *Federation of European Biochemical Societies letters* 585.23 (2011), pp. 3715–3723. DOI: [10.1016/j.febslet.2011.08.004](https://doi.org/10.1016/j.febslet.2011.08.004).
- [141] Aleksandra K. Lebedeva et al. “MRI-Based Classification Models in Prediction of Mild Cognitive Impairment and Dementia in Late-Life Depression”. In: *Frontiers in Aging Neuroscience* 9 (2017), p. 13. DOI: [10.3389/fnagi.2017.00013](https://doi.org/10.3389/fnagi.2017.00013).
- [142] Christian Ledig, Schuh Andreas, Guerrero Ricardo, Heckemann Rolf A., and Rueckert Daniel. “Structural brain imaging in Alzheimer’s disease and mild cognitive impairment: biomarker analysis and shared morphometry database”. In: *Scientific Report* 8 (2018), p. 11258. DOI: [10.1038/s41598-018-29295-9](https://doi.org/10.1038/s41598-018-29295-9).
- [143] Jeyeon Lee et al. “Deep learning-based brain age prediction in normal aging and dementia”. In: *Nature Aging* 2 (2022), pp. 412–424. DOI: [10.1038/s43587-022-00219-7](https://doi.org/10.1038/s43587-022-00219-7).
- [144] Esten H. Leonardsen et al. “Deep neural networks learn general and clinically relevant representations of the aging brain”. In: *NeuroImage* 256 (2022), p. 119210. DOI: [10.1016/j.neuroimage.2022.119210](https://doi.org/10.1016/j.neuroimage.2022.119210).
- [145] Maxime Leroy, Maxime Bertoux, Emilie Skrobala, et al. “Characteristics and progression of patients with frontotemporal dementia in a regional memory clinic network”. In: *Alzheimer’s Research & Therapy* 13 (2021), p. 19. DOI: [10.1186/s13195-020-00753-9](https://doi.org/10.1186/s13195-020-00753-9).
- [146] Muriel D. Lezak, Diane B. Howieson, Erin D. Bigler, and Daniel Tranel. *Neuropsychological assessment*. 2012.
- [147] Chao Li et al. “Trans-ResNet: Integrating Transformers and CNNs for Alzheimer’s disease classification”. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022, pp. 1–5. DOI: [10.1109/ISBI52829.2022.9761549](https://doi.org/10.1109/ISBI52829.2022.9761549).
- [148] Fan Li, Danni Cheng, and Manhwa Liu. “Alzheimer’s disease classification based on combination of multi-model convolutional networks”. In: *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*. 2017, pp. 1–5. DOI: [10.1109/IST.2017.8261566](https://doi.org/10.1109/IST.2017.8261566).
- [149] Fan Li and Manhwa Liu. “Alzheimer’s disease diagnosis based on multiple cluster dense convolutional networks”. In: *Computerized Medical Imaging and Graphics* 70 (2018), pp. 101–110. DOI: [10.1016/j.compmedimag.2018.09.009](https://doi.org/10.1016/j.compmedimag.2018.09.009).
- [150] Xiaoxiao Li et al. “BrainGNN: Interpretable Brain: A Journal of Neurology Graph Neural Network for fMRI Analysis”. In: *Medical Image Analysis* (2020). DOI: [10.1016/j.media.2021.102233](https://doi.org/10.1016/j.media.2021.102233).

- [151] Yi Li and Nuno Vasconcelos. “REPAIR: Removing Representation Bias by Dataset Resampling”. In: *CoRR* abs/1904.07911 (2019). arXiv: [1904.07911](#).
- [152] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. “Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer’s Disease Diagnosis Using Structural MRI”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020), pp. 880–893. DOI: [10.1109/TPAMI.2018.2889096](#).
- [153] Franziskus Liem et al. “Predicting brain-age from multimodal imaging data captures cognitive impairment”. In: *NeuroImage* 148 (2017), pp. 179–188. DOI: [10.1016/j.neuroimage.2016.11.005](#).
- [154] Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. “Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy”. In: *Nature Reviews. Neurology* 9.2 (2013), pp. 106–118. DOI: [10.1038/nrneurol.2012.263](#).
- [155] Jiang Liu et al. “One Model to Synthesize Them All: Multi-contrast Multi-scale Transformer for Missing Data Imputation”. In: *IEEE Transactions on Medical Imaging* (2023), pp. 1–1. DOI: [10.1109/TMI.2023.3261707](#).
- [156] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. “Deep Multi-Task Multi-Channel Learning for Joint Classification and Regression of Brain Status”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Vol. 10435. 2017, pp. 3–11. DOI: [10.1007/978-3-319-66179-7_1](#).
- [157] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. “Landmark-based deep multi-instance learning for brain disease diagnosis”. In: *Medical Image Analysis* 43 (2018), pp. 157–168. DOI: [10.1016/j.media.2017.10.005](#).
- [158] Mingxia Liu, Jun Zhang, Dong Nie, Pew-Thian Yap, and Dinggang Shen. “Anatomical Landmark Based Deep Feature Representation for MR Images in Brain: A Journal of Neurology Disease Diagnosis”. In: *IEEE journal of biomedical and health informatics* 22 (2018), pp. 1476–1485. DOI: [10.1109/JBHI.2018.2791863](#).
- [159] Yawu Liu et al. “Education increases reserve against Alzheimer’s disease—evidence from structural MRI analysis”. In: *Neuroradiology* 54 (2012), pp. 929–938. DOI: [10.1007/s00234-012-1005-0](#).
- [160] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022. DOI: [10.1109/ICCV48922.2021.00986](#).
- [161] Gill Livingston et al. “Dementia prevention, intervention, and care”. In: *The Lancet* 390.10113 (2017), pp. 2673–2734. DOI: [10.1016/S0140-6736\(17\)31363-6](#).

- [162] Miguel López et al. “Neurological image classification for the Alzheimer’s Disease diagnosis using Kernel PCA and Support Vector Machines”. In: *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*. 2009, pp. 2486–2489. DOI: [10.1109/NSSMIC.2009.5402069](https://doi.org/10.1109/NSSMIC.2009.5402069).
- [163] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint* (2017). eprint: [arXiv:1711.05101](https://arxiv.org/abs/1711.05101).
- [164] Po H. Lu et al. “Patterns of Brain: A Journal of Neurology Atrophy in Clinical Variants of Frontotemporal Lobar Degeneration”. In: *Dementia and Geriatric Cognitive Disorders* 35 (2013), pp. 34–50. DOI: [10.1159/000345523](https://doi.org/10.1159/000345523).
- [165] Hansen Lui et al. “Progranulin Deficiency Promotes Circuit-Specific Synaptic Pruning by Microglia via Complement Activation”. In: *Cell* 165.4 (2016), pp. 921–935. DOI: [10.1016/j.cell.2016.04.001](https://doi.org/10.1016/j.cell.2016.04.001).
- [166] Yanjun Lyu et al. “Classification of Alzheimer’s Disease via Vision Transformer: Classification of Alzheimer’s Disease via Vision Transformer”. In: *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. 2022, pp. 463–468. DOI: [10.1145/3529190.3534754](https://doi.org/10.1145/3529190.3534754).
- [167] Da Ma, Donghuan Lu, Karteek Popuri, Lei Wang, and Mirza Faisal Beg. “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 (2020), p. 853. DOI: [10.3389/fnins.2020.00853](https://doi.org/10.3389/fnins.2020.00853).
- [168] Ines Mahjoub, Mohamed Ali Mahjoub, and Islem Rekik. “Brain multiplexes reveal morphological connectional biomarkers fingerprinting late brain dementia states”. In: *Scientific Reports* 8 (2018), p. 4103. DOI: [10.1038/s41598-018-21568-7](https://doi.org/10.1038/s41598-018-21568-7).
- [169] Ian B. Malone et al. “MIRIAD—Public release of a multiple time point Alzheimer’s MR imaging dataset”. In: *NeuroImage* 70 (2013), pp. 33–36. DOI: [10.1016/j.neuroimage.2012.12.044](https://doi.org/10.1016/j.neuroimage.2012.12.044).
- [170] Pravat K. Mandal, Rashima Mahajan, and Ivo D. Dinov. “Structural brain atlases: design, rationale, and applications in normal and pathological cohorts”. In: *Journal of Alzheimer’s Disease* 31.3 (2012), S169–S188. DOI: [10.3233/JAD-2012-120412](https://doi.org/10.3233/JAD-2012-120412).
- [171] José V. Manjón et al. “Robust MRI brain tissue parameter estimation by multistage outlier rejection”. In: *Magnetic Resonance in Medicine* 59.4 (2008), pp. 866–873. DOI: [10.1002/mrm.21521](https://doi.org/10.1002/mrm.21521).
- [172] José V. Manjón, Pierrick Coupé, Luis Martí-Bonmatí, D. Louis Collins, and Montserrat Robles. “Adaptive non-local means denoising of MR images with spatially varying noise levels”. In: *Journal of Magnetic Resonance Imaging : JMRI* 31.1 (2010), pp. 192–203. DOI: [10.1002/jmri.22003](https://doi.org/10.1002/jmri.22003).

- [173] José V. Manjón et al. “Diffusion weighted image denoising using overcomplete local PCA”. In: *PloS One* 8.9 (2013), e73021. DOI: [10.1371/journal.pone.0073021](https://doi.org/10.1371/journal.pone.0073021).
- [174] José V. Manjón et al. “Nonlocal intracranial cavity extraction”. In: *International Journal of Biomedical Imaging* 2014 (2014), p. 820205. DOI: [10.1155/2014/820205](https://doi.org/10.1155/2014/820205).
- [175] William R. Markesbery. “Neuropathologic alterations in mild cognitive impairment: a review”. In: *Journal of Alzheimer’s Disease* 19 (2010), pp. 221–228. DOI: [10.3233/JAD-2010-1220](https://doi.org/10.3233/JAD-2010-1220).
- [176] Thais Massetti et al. “Virtual reality in multiple sclerosis - A systematic review”. In: *Multiple Sclerosis and Related Disorders* 8 (2016), pp. 107–112. DOI: [10.1016/j.msard.2016.05.014](https://doi.org/10.1016/j.msard.2016.05.014).
- [177] John Mazziotta et al. “A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)”. In: *Philosophical Transactions of the Royal Society of London* 356.1412 (2001), pp. 1293–1322. DOI: [10.1098/rstb.2001.0915](https://doi.org/10.1098/rstb.2001.0915).
- [178] Robert A. McCutcheon, John H. Krystal, and Oliver D. Howes. “Dopamine and glutamate in schizophrenia: biology, symptoms and treatment”. In: *World psychiatry : official journal of the World Psychiatric Association (WPA)* 19.1 (2020), pp. 15–33. DOI: [10.1002/wps.20693](https://doi.org/10.1002/wps.20693).
- [179] Leland McInnes, Healy John, and Melville James. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: *arXiv* (2020). eprint: [arXiv:1802.03426](https://arxiv.org/abs/1802.03426).
- [180] Guy M. McKhann et al. “The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 7 (2011), pp. 263–269. DOI: [10.1016/j.jalz.2011.03.005](https://doi.org/10.1016/j.jalz.2011.03.005).
- [181] Jie Mei et al. “Machine Learning for the Diagnosis of Parkinson’s Disease: A Review of Literature”. In: *Frontiers in Aging Neuroscience* 13 (2021), p. 633752. DOI: [10.3389/fnagi.2021.633752](https://doi.org/10.3389/fnagi.2021.633752).
- [182] Sidney Mintz and Michael Alpert. “Imagery Vividness, Reality Testing, and Schizophrenic Hallucinations”. In: *Journal of Abnormal Psychology* 79.3 (1972), pp. 310–316. DOI: [10.1037/h0033209](https://doi.org/10.1037/h0033209).
- [183] Shiwangi Mishra, Iman Beheshti, and Pritee Khanna. “A Review of Neuroimaging-driven Brain: A Journal of Neurology Age Estimation for identification of Brain: A Journal of Neurology Disorders and Health Conditions”. In: *IEEE Reviews in Biomedical Engineering* (2021), pp. 1–1. DOI: [10.1109/RBME.2021.3107372](https://doi.org/10.1109/RBME.2021.3107372).

- [184] Christiane Möller et al. “Alzheimer Disease and Behavioral Variant Frontotemporal Dementia: Automatic Classification Based on Cortical Atrophy for Single-Subject Diagnosis”. In: *Radiology* 279 (2016), pp. 838–848. DOI: [10.1148/radiol.2015150220](https://doi.org/10.1148/radiol.2015150220).
- [185] Tohid Mortezaazadeh et al. “Imaging modalities in differential diagnosis of Parkinson’s disease: opportunities and challenges”. In: *Egyptian Journal of Radiology and Nuclear Medicine* 52 (2021), p. 79. DOI: [10.1186/s43055-021-00454-9](https://doi.org/10.1186/s43055-021-00454-9).
- [186] Yangling Mu and Fred H. Gage. “Adult hippocampal neurogenesis and its role in Alzheimer’s disease”. In: *Molecular Neurodegeneration* 6 (2011), p. 85. DOI: [10.1186/1750-1326-6-85](https://doi.org/10.1186/1750-1326-6-85).
- [187] Susanne G. Mueller et al. “The Alzheimer’s disease neuroimaging initiative”. In: *Neuroimaging Clinics of North America* 15.4 (2005), pp. 869–xii. DOI: [10.1016/j.nic.2005.09.008](https://doi.org/10.1016/j.nic.2005.09.008).
- [188] Eitaro Nakamura and Kenji Miyao. “A Method for Identifying Biomarkers of Aging and Constructing an Index of Biological Age in Humans”. In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 62 (2007), pp. 1096–1105. DOI: [10.1093/gerona/62.10.1096](https://doi.org/10.1093/gerona/62.10.1096).
- [189] Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. “Data augmentation for brain-tumor segmentation: a review”. In: *Frontiers in Computational Neuroscience* 13 (2019), p. 83. DOI: [10.3389/fncom.2019.00083](https://doi.org/10.3389/fncom.2019.00083).
- [190] Cullen O’Gorman, Robyn Lucas, and Bruce Taylor. “Environmental Risk Factors for Multiple Sclerosis: A Review with a Focus on Molecular Mechanisms”. In: *International Journal of Molecular Sciences* 13.9 (2012), pp. 11718–11752. DOI: [10.3390/ijms130911718](https://doi.org/10.3390/ijms130911718).
- [191] Kanghan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. “Classification and Visualization of Alzheimer’s Disease using Volumetric Convolutional Neural Network and Transfer Learning”. In: *Scientific Report* 9 (2019), p. 18150. DOI: [10.1038/s41598-019-54548-6](https://doi.org/10.1038/s41598-019-54548-6).
- [192] Dan Pan et al. “Early Detection of Alzheimer’s Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning”. In: *Frontiers in Neuroscience* 14 (2020). DOI: [10.3389/fnins.2020.00259](https://doi.org/10.3389/fnins.2020.00259).
- [193] Nalin Payakachat, J. Mick Tilford, and Wendy J. Ungar. “National database for autism research (NDAR): Big data opportunities for health services research and health technology assessment”. In: *PharmacoEconomics* 34.2 (2016), pp. 127–138. DOI: [10.1007/s40273-015-0331-6](https://doi.org/10.1007/s40273-015-0331-6).

- [194] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [195] Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. “Accurate brain age prediction with lightweight deep neural networks”. In: *Medical Image Analysis* 68 (2021), p. 101871. DOI: [10.1016/j.media.2020.101871](https://doi.org/10.1016/j.media.2020.101871).
- [196] Agnès Pérez-Millan et al. “Classifying Alzheimer’s disease and frontotemporal dementia using machine learning with cross-sectional and longitudinal magnetic resonance imaging data”. In: *Human Brain Mapping* 44 (2023), pp. 2234–2244. DOI: [10.1002/hbm.26205](https://doi.org/10.1002/hbm.26205).
- [197] Ruth Peters. “Aging and the brain”. In: *Postgraduate Medical Journal* 82 (2006), pp. 84–88. DOI: [10.1136/pgmj.2005.036665](https://doi.org/10.1136/pgmj.2005.036665).
- [198] Vincent Planche et al. “Structural progression of Alzheimer’s disease over decades: the MRI staging scheme”. In: *Brain Communications* 4.3 (2022), fcac109. DOI: [10.1093/braincomms/fcac109](https://doi.org/10.1093/braincomms/fcac109).
- [199] Vincent Planche et al. “Anatomical MRI staging of frontotemporal dementia variants”. In: *Alzheimer’s & dementia : the journal of the Alzheimer’s Association* (2023). DOI: [10.1002/alz.12975](https://doi.org/10.1002/alz.12975).
- [200] Katherine L. Possin, Victor R. Laluz, Oscar Z. Alcantar, Bruce L. Miller, and Joel H. Kramer. “Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer’s disease and behavioral variant frontotemporal dementia”. In: *Neuropsychologia* 49.1 (2011), pp. 43–48. DOI: [10.1016/j.neuropsychologia.2010.10.026](https://doi.org/10.1016/j.neuropsychologia.2010.10.026).
- [201] Shangran Qiu et al. “Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment”. In: *Alzheimer’s & Dementia* 10 (2018), pp. 737–749. DOI: [10.1016/j.dadm.2018.08.013](https://doi.org/10.1016/j.dadm.2018.08.013).
- [202] Gil Rabinovici et al. “Distinct MRI Atrophy Patterns in Autopsy-Proven Alzheimer’s Disease and Frontotemporal Lobar Degeneration”. In: *American Journal of Alzheimer’s Disease & Other Dementias* 22 (2008), pp. 474–488. DOI: [10.1177/1533317507308779](https://doi.org/10.1177/1533317507308779).
- [203] Vineet K. Raghu, Jakob Weiss, Udo Hoffmann, Hugo J.W.L. Aerts, and Michael T. Lu. “Deep Learning to Estimate Biological Age From Chest Radiographs”. In: *Cardiovascular Imaging* 14 (2021), pp. 2226–2236. DOI: [10.1016/j.jcmg.2021.01.008](https://doi.org/10.1016/j.jcmg.2021.01.008).

- [204] Tarek Rajji, Zahinoor Ismail, and Benoit Mulsant. “Age at onset and cognition in schizophrenia: Meta-analysis”. In: *British Journal of Psychiatry* 195.4 (2009), pp. 286–293. DOI: [10.1192/bjp.bp.108.060723](https://doi.org/10.1192/bjp.bp.108.060723).
- [205] Anil Rao, Ying Lee, Achim Gass, and Andreas Monsch. “Classification of Alzheimer’s Disease from structural MRI using sparse logistic regression with optional spatial regularization”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE Engineering in Medicine and Biology Society. Annual International Conference. 2011, pp. 4499–4502. DOI: [10.1109/IEMBS.2011.6091115](https://doi.org/10.1109/IEMBS.2011.6091115).
- [206] Katya Rascovsky et al. “Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia”. In: *Brain: A Journal of Neurology* 134 (2011), pp. 2456–2477. DOI: [10.1093/brain/awr179](https://doi.org/10.1093/brain/awr179).
- [207] Hanne Rasmussen, Eystein Stordal, and Tor A. Rosness. “Risk factors for frontotemporal dementia.” In: *Tidsskrift for den Norske lægeforening : tidsskrift for praktisk medicin, ny raekke* 138.14 (2018). DOI: [10.4045/tidsskr.17.0763](https://doi.org/10.4045/tidsskr.17.0763).
- [208] Nicholas Robinson and Sarah E. Bergen. “Environmental Risk Factors for Schizophrenia and Bipolar Disorder and Their Relationship to Genetic Risk: Current Knowledge and Future Directions”. In: *Frontiers in Genetics* 12 (2021), p. 686666. DOI: [10.3389/fgene.2021.686666](https://doi.org/10.3389/fgene.2021.686666).
- [209] Walter A. Rocca et al. “Trends in the incidence and prevalence of Alzheimer’s disease, dementia, and cognitive impairment in the United States”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 7.1 (2011), pp. 80–93. DOI: [10.1016/j.jalz.2010.11.002](https://doi.org/10.1016/j.jalz.2010.11.002).
- [210] Jonathan D. Rohrer. “Structural brain imaging in frontotemporal dementia”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1822.3 (2012), pp. 325–332. DOI: [10.1016/j.bbadis.2011.07.014](https://doi.org/10.1016/j.bbadis.2011.07.014).
- [211] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [212] Martino Ruggieri, Paola Iannetti, Angela Polizzi, and et al. “Multiple sclerosis in children under 10 years of age”. In: *Neurological Sciences* 25.Suppl 4 (2004), s326–s335. DOI: [10.1007/s10072-004-0335-z](https://doi.org/10.1007/s10072-004-0335-z).
- [213] Benoît Sauty and Stanley Durrleman. “Progression Models for Imaging Data with Longitudinal Variational Auto Encoders”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Vol. 13431. 2022. DOI: [10.1007/978-3-031-16431-6_1](https://doi.org/10.1007/978-3-031-16431-6_1).

- [214] Daniel Schmitter et al. “An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer’s disease”. In: *NeuroImage Clinical* 7 (2015), pp. 7–17. DOI: [10.1016/j.nicl.2014.11.001](https://doi.org/10.1016/j.nicl.2014.11.001).
- [215] Norbert Schuff et al. “MRI of hippocampal volume loss in early Alzheimer’s disease in relation to ApoE genotype and biomarkers”. In: *Brain: A Journal of Neurology* 132 (2009), pp. 1067–1077. DOI: [10.1093/brain/awp007](https://doi.org/10.1093/brain/awp007).
- [216] Flávio L. Seixas, Bianca Zadrozny, Jerson Laks, Aura Conci, and Débora C. Muchaluat Saade. “A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer’s disease and mild cognitive impairment”. In: *Computers in Biology and Medicine* 51 (2014), pp. 140–158. DOI: [10.1016/j.combiomed.2014.04.010](https://doi.org/10.1016/j.combiomed.2014.04.010).
- [217] Mohammad S.E. Sendi et al. “Brain age acceleration as biomarker of Alzheimer’s disease progression: Functional network connectivity analysis”. In: *Alzheimer’s & Dementia* 17 (2021). DOI: [10.1002/alz.050562](https://doi.org/10.1002/alz.050562).
- [218] Uday S. Shanthamallu, Andreas Spanias, Cihan Tepedelenlioglu, and Mike Stanley. “A brief survey of machine learning methods and their sensor and IoT applications”. In: *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*. 2017, pp. 1–8. DOI: [10.1109/IISA.2017.8316459](https://doi.org/10.1109/IISA.2017.8316459).
- [219] Yunsheng Shi et al. *Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification*. 2020. eprint: [arXiv:2009.03509](https://arxiv.org/abs/2009.03509).
- [220] Suhad Al-Shoukry, Taha H. Rassem, and Nasrin M. Makbol. “Alzheimer’s Diseases Detection by Using Deep Learning Algorithms: A Mini-Review”. In: *IEEE Access* 8 (2020), pp. 77131–77141. DOI: [10.1109/ACCESS.2020.2989396](https://doi.org/10.1109/ACCESS.2020.2989396).
- [221] Amar Shukla, Rajeev Tiwari, and Shamik Tiwari. “Review on Alzheimer Disease Detection Methods: Automatic Pipelines and Machine Learning Techniques”. In: *Sci* 5 (2023), p. 13. DOI: [10.3390/sci5010013](https://doi.org/10.3390/sci5010013).
- [222] David Silhan et al. “The parietal atrophy score on brain magnetic resonance imaging is a reliable visual scale”. In: *Current Alzheimer Research* 17.6 (2020), pp. 534–539. DOI: [10.2174/1567205017666200807193957](https://doi.org/10.2174/1567205017666200807193957).
- [223] Stephen M. Smith et al. “Estimation of brain age delta from brain imaging”. In: *NeuroImage* 200 (2019), pp. 528–539. DOI: [10.1016/j.neuroimage.2019.06.017](https://doi.org/10.1016/j.neuroimage.2019.06.017).
- [224] Kateřina Storey et al. “FTLD-TDP and progressive supranuclear palsy in comorbidity—a report of two cases with different clinical presentations”. In: *Neurocase* 23.1 (2017), pp. 5–11.

- [225] Andrew J. Swift et al. “Emphysematous changes and normal variation in smokers and COPD patients using diffusion 3He MRI”. In: *European Journal of Radiology* 54.3 (2005), pp. 352–358. DOI: [10.1016/j.ejrad.2004.08.002](https://doi.org/10.1016/j.ejrad.2004.08.002).
- [226] Saori C. Tanaka et al. “A multi-site, multi-disorder resting-state magnetic resonance image database”. In: *Scientific Data* 8.1 (2021), pp. 1–11. DOI: [10.1038/s41597-021-01004-8](https://doi.org/10.1038/s41597-021-01004-8).
- [227] Jason R. Taylor et al. “The Cambridge Centre for aging and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample”. In: *NeuroImage* 144 (2017), pp. 262–269.
- [228] Alessandro Termine, Carlo Fabrizio, Carlo Caltagirone, Laura Petrosini, and on behalf of the Frontotemporal Lobar Degeneration Neuroimaging Initiative. “A Reproducible Deep-Learning-Based Computer-Aided Diagnosis Tool for Frontotemporal Dementia Using MONAI and Clinica Frameworks”. In: *Life* 12 (2022), p. 947. DOI: [10.3390/life12070947](https://doi.org/10.3390/life12070947).
- [229] Andrea Termine, Carlo Fabrizio, Carlo Caltagirone, and Laura Petrosini. “A Reproducible Deep-Learning-Based Computer-Aided Diagnosis Tool for Frontotemporal Dementia Using MONAI and Clinica Frameworks”. In: *Life* 12 (2022), p. 947. DOI: [10.3390/life12070947](https://doi.org/10.3390/life12070947).
- [230] Elina Thibeau-Sutre, Olivier Colliot, Didier Dormont, and Ninon Burgos. “Visualization approach to assess the robustness of neural networks for medical image classification”. In: *Medical Imaging 2020: Image Processing*. 2020, p. 54. DOI: [10.1117/12.2548952](https://doi.org/10.1117/12.2548952).
- [231] Elina Thibeau-Sutre, Baptiste Couvy-Duchesne, Didier Dormont, Olivier Colliot, and Ninon Burgos. “MRI field strength predicts Alzheimer’s disease: a case example of bias in the ADNI data set”. In: *ISBI 2022 - International Symposium on Biomedical Imaging*. Vol. Proc. ISBI 2022 - International Symposium on Biomedical Imaging. 2022. DOI: [10.1109/ISBI52829.2022.9761504](https://doi.org/10.1109/ISBI52829.2022.9761504).
- [232] Danielle J. Tisserand et al. “A Voxel-based Morphometric Study to Determine Individual Differences in Gray Matter Density Associated with Age and Cognitive Change Over Time”. In: *Cerebral Cortex* 14 (2004), pp. 966–973. DOI: [10.1093/ercor/bhh057](https://doi.org/10.1093/ercor/bhh057).
- [233] Max Toepper. “Dissociating Normal Aging from Alzheimer’s Disease: A View from Cognitive Neuroscience”. In: *Journal of Alzheimer’s Disease* 57 (2017), pp. 331–352. DOI: [10.3233/JAD-161099](https://doi.org/10.3233/JAD-161099).

- [234] Tong Tong et al. “A Novel Grading Biomarker for the Prediction of Conversion From Mild Cognitive Impairment to Alzheimer’s Disease”. In: *Transactions on Biomedical Engineering* 64 (2017), pp. 155–165. DOI: [10.1109/TBME.2016.2549363](https://doi.org/10.1109/TBME.2016.2549363).
- [235] Hugo Touvron et al. “Three things everyone should know about vision transformers”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. 2022, pp. 497–515.
- [236] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. “Fixing the train-test resolution discrepancy”. In: *Advances in neural information processing systems* 32 (2019).
- [237] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International conference on machine learning*. 2021, pp. 10347–10357.
- [238] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. “Are Convolutional Neural Networks or Transformers more like human vision?” In: *CoRR* (2021). arXiv: [2105.07197](https://arxiv.org/abs/2105.07197). URL: <https://arxiv.org/abs/2105.07197>.
- [239] Nick J. Tustison et al. “N4ITK: Improved N3 Bias Correction”. In: *IEEE Transactions on Medical Imaging* 29.6 (2010), pp. 1310–1320. DOI: [10.1109/TMI.2010.2046908](https://doi.org/10.1109/TMI.2010.2046908).
- [240] Masaru Ueda et al. “An Age Estimation Method Using 3D-CNN From Brain: A Journal of Neurology MRI Images”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 380–383. DOI: [10.1109/ISBI.2019.8759392](https://doi.org/10.1109/ISBI.2019.8759392).
- [241] Aly Valliani and Ameet Soni. “Deep Residual Nets for Improved Alzheimer’s Diagnosis”. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2017, pp. 615–615. DOI: [10.1145/3107411.3108224](https://doi.org/10.1145/3107411.3108224).
- [242] Tinu Varghese, Kuman R. Sheela, P. S. Mathuranath, and Albert Singh. “Evaluation of different stages of Alzheimer’s disease using unsupervised clustering techniques and voxel based morphometry”. In: *2012 World Congress on Information and Communication Technologies*. Trivandrum, India, 2012, pp. 953–958. DOI: [10.1109/WICT.2012.6409212](https://doi.org/10.1109/WICT.2012.6409212).
- [243] Ali Varzandian, Miguel Angel Sanchez Razo, Michael Richard Sanders, Akhila Atmakuru, and Giuseppe Di Fatta. “Classification-Biased Apparent Brain: A Journal of Neurology Age for the Prediction of Alzheimer’s Disease”. In: *Frontiers in Neuroscience* 15 (2021), p. 673120. DOI: [10.3389/fnins.2021.673120](https://doi.org/10.3389/fnins.2021.673120).

- [244] Petar Veličković et al. “Graph Attention Networks”. In: *arXiv preprint* (2017). eprint: [arXiv:1710.10903](https://arxiv.org/abs/1710.10903).
- [245] Antonio Vita, Luca De Peri, Giuseppe Deste, and Emilio Sacchetti. “Progressive loss of cortical gray matter in schizophrenia: a meta-analysis and meta-regression of longitudinal MRI studies”. In: *Translational Psychiatry* 2.11 (2012), e190. DOI: [10.1038/tp.2012.116](https://doi.org/10.1038/tp.2012.116).
- [246] Sandra Vukusic et al. “Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France”. In: *Multiple Sclerosis Journal* 26.1 (2020), pp. 118–122. DOI: [10.1177/1352458518815602](https://doi.org/10.1177/1352458518815602).
- [247] Christian Wachinger, David H. Salat, Michael Weiner, and Martin Reuter. “Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala”. In: *Brain: A Journal of Neurology* 139 (2016), pp. 3253–3266. DOI: [10.1093/brain/aww243](https://doi.org/10.1093/brain/aww243).
- [248] Guotai Wang et al. *Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation*. 2018.
- [249] Jason D. Warren, Jonathan D. Rohrer, and Martin N. Rossor. “Clinical review. Frontotemporal dementia”. In: *British Medical Journal* 347 (2013), f4827. DOI: [10.1136/bmj.f4827](https://doi.org/10.1136/bmj.f4827).
- [250] Junhao Wen et al. “Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation”. In: *Medical Image Analysis* 63 (2020), p. 101694. DOI: [10.1016/j.media.2020.101694](https://doi.org/10.1016/j.media.2020.101694).
- [251] Jennifer L. Whitwell et al. “Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: a cluster analysis study”. In: *Brain: A Journal of Neurology* 132 (2009), pp. 2932–2946. DOI: [10.1093/brain/awp232](https://doi.org/10.1093/brain/awp232).
- [252] Anders Wimo et al. “The worldwide economic impact of dementia 2010”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 9.1 (2013), 1–11.e3. DOI: [10.1016/j.jalz.2012.11.006](https://doi.org/10.1016/j.jalz.2012.11.006).
- [253] Anders Wimo et al. “The worldwide costs of dementia 2015 and comparisons with 2010”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 13.1 (2017), pp. 1–7. DOI: [10.1016/j.jalz.2016.07.150](https://doi.org/10.1016/j.jalz.2016.07.150).
- [254] Tom N. Wolf, Sebastian Pölsterl, and Christian Wachinger. “DAFT: A universal module to interweave tabular data and 3D images in CNNs”. In: *NeuroImage* 260 (2022), p. 119505. DOI: [10.1016/j.neuroimage.2022.119505](https://doi.org/10.1016/j.neuroimage.2022.119505).
- [255] World Health Organization. *Dementia*. <https://www.who.int/news-room/fact-sheets/detail/dementia>. 2021.

- [256] Zhuofan Xia, Xuran Pan, Shiji Song, Erran Li Li, and Gao Huang. “Vision transformer with deformable attention”. In: *Conference on computer vision and pattern recognition*. 2022, pp. 4794–4803. DOI: [10.1109/CVPR52688.2022.00475](https://doi.org/10.1109/CVPR52688.2022.00475).
- [257] Evangeline Yee, Da Ma, Karteek Popuri, Leib Wang, and Mirza Faisal Beg. “Construction of MRI-Based Alzheimer’s Disease Score Based on Efficient 3D Convolutional Neural Network: Comprehensive Validation on 7,902 Images from a Multi-Center Dataset”. In: *Journal of Alzheimer’s disease* 79 (2021), pp. 47–58. DOI: [10.3233/JAD-200830](https://doi.org/10.3233/JAD-200830).
- [258] Belindaa Yew, Suvarnad Alladi, Mekalad Shailaja, John R. Hodges, and Michael Hornberger. “Lost and forgotten? Orientation versus memory in Alzheimer’s disease and Frontotemporal dementia”. In: *Journal of Alzheimer’s disease* 33 (2013), pp. 473–481. DOI: [10.3233/JAD-2012-120769](https://doi.org/10.3233/JAD-2012-120769).
- [259] Peter N. E. Young et al. “Imaging biomarkers in neurodegeneration: current and future practices”. In: *Alzheimer’s Research & Therapy* 12.1 (2020), pp. 1–13. DOI: [10.1186/s13195-020-00612-7](https://doi.org/10.1186/s13195-020-00612-7).
- [260] Mojtaba Zarei et al. “Cortical thinning is associated with disease stages and dementia in Parkinson’s disease”. In: *Journal of Neurology, Neurosurgery and Psychiatry* 84.8 (2013), pp. 875–881. DOI: [10.1136/jnnp-2012-304126](https://doi.org/10.1136/jnnp-2012-304126).
- [261] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. “mixup: Beyond Empirical Risk Minimization”. In: *6th International Conference on Learning Representations, ICLR 2018*. 2018.
- [262] Shengjie Zhang et al. “3D Global Fourier Network for Alzheimer’s Disease Diagnosis Using Structural MRI”. In: *Medical Image Computing and Computer Assisted Intervention*. 2022, pp. 34–43. DOI: [10.1007/978-3-031-16431-6_4](https://doi.org/10.1007/978-3-031-16431-6_4).
- [263] Yu Zhang et al. “MRI signatures of brain macrostructural atrophy and microstructural degradation in frontotemporal lobar degeneration subtypes”. In: *Journal of Alzheimer’s Disease* 33.2 (2013), pp. 431–444. DOI: [10.3233/JAD-2012-121156](https://doi.org/10.3233/JAD-2012-121156).
- [264] Yudong Zhang, Shuihua Wang, and Zhengchao Dong. “Classification of Alzheimer Disease Based on Structural Magnetic Resonance Imaging by Kernel Support Vector Machine Decision Tree”. In: *Progress In Electromagnetics Research* 144 (2014), pp. 185–191. DOI: [10.2528/PIER13121310](https://doi.org/10.2528/PIER13121310).
- [265] Mingbo Zhao, Rosa H. Chan, Peng Tang, Tommy W. Chow, and Savio W. Wong. “Trace Ratio Linear Discriminant Analysis for Medical Diagnosis: A Case Study of Dementia”. In: *IEEE Signal Processing Letters* 20.5 (2013), pp. 431–434. DOI: [10.1109/LSP.2013.2250281](https://doi.org/10.1109/LSP.2013.2250281).

- [266] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. “Learning Deep Features for Discriminative Localization”. In: (2015), pp. 2921–2929.